Three-Dimensional Parametrization for Parsing Morphologically Rich Languages

Reut Tsarfaty and Khalil Sima'an
Institute for Logic, Language and Computation
University of Amsterdam
Plantage Muidergracht 24, 1018TV Amsterdam, The Netherlands
{rtsarfat,simaan}@science.uva.nl

Abstract

Current parameters of accurate unlexicalized parsers based on Probabilistic Context-Free Grammars (PCFGs) form a two-dimensional grid in which rewrite events are conditioned on both horizontal (head-outward) and vertical (parental) histories. In Semitic languages, where arguments may move around rather freely and phrase-structures are often shallow, there are additional morphological factors that govern the generation process. Here we propose that agreement features percolated up the parse-tree form a third dimension of parametrization that is orthogonal to the previous two. This dimension differs from mere "state-splits" as it applies to a whole set of categories rather than to individual ones and encodes linguistically motivated co-occurrences between them. This paper presents extensive experiments with extensions of unlexicalized PCFGs for parsing Modern Hebrew in which tuning the parameters in three dimensions gradually leads to improved performance. Our best result introduces a new, stronger, lower bound on the performance of treebank grammars for parsing Modern Hebrew, and is on a par with current results for parsing Modern Standard Arabic obtained by a fully lexicalized parser trained on a much larger treebank.

1 Dimensions of Unlexicalized Parsing

Probabilistic Context Free Grammars (PCFGs) are the formal backbone of most high-accuracy statistical parsers for English, and a variety of techniques was developed to enhance their performance relative to the naïve treebank implementation — from unlexicalized extensions exploiting simple category splits (Johnson, 1998; Klein and Manning, 2003) to fully lexicalized parsers that condition events below a constituent upon the head and additional lexical content (Collins, 2003; Charniak, 1997). While it is clear that conditioning on lexical content improves the grammar's disambiguation capabilities, Klein and Manning (2003) demonstrate that a well-crafted unlexicalized PCFG can close the gap, to a large extent, with current state-of-the-art lexicalized parsers for English.

The factor that sets apart vanilla PCFGs (Charniak, 1996) from their unlexicalized extensions proposed by, e.g., (Johnson, 1998; Klein and Manning, 2003), is the choice for statistical parametrization that weakens the independence assumptions implicit in the treebank grammar. Studies on accurate unlexicalized parsing models outline two dimensions of parametrization. The first, proposed by (Johnson, 1998), is the annotation of parental history, and the second encodes a head-outward generation process (Collins, 2003). Johnson (1998) augments node labels with the label of their parent, thus incorporating a dependency on the node's grandparent. Collins (2003) proposes to generate the head of a phrase first and then generate its sisters using Markovian processes, thereby exploiting head/sister-dependencies.

Proceedings of the 10th Conference on Parsing Technologies, pages 156–167,
Prague, Czech Republic, June 2007.

Klein and Manning (2003) systematize the distinction between these two forms of parametrization by drawing them on a horizontal-vertical grid: parent encoding is vertical (external to the rule) whereas head-outward generation is horizontal (internal to the rule). By varying the value of the parameters along the grid, Klein and Manning (2003) tune their treebank grammar to achieve improved performance. This two-dimensional parametrization has been instrumental in devising parsing models that improve disambiguation capabilities for English as well as other languages, such as German (Dubey and Keller, 2003) Czech (Collins et al., 1999) and Chinese (Bikel and Chiang, 2000). However, accuracy results for parsing languages other than English still lag behind.[1]

We propose that for various languages including the Semitic family, e.g. Modern Hebrew (MH) and Modern Standard Arabic (MSA), a third dimension of parametrization is necessary 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 the vertical and horizontal conditioning. We propose a third dimension of parameterizations that encodes morphological features such as those realizing syntactic agreement. These features are percolated from surface forms in a bottom-up fashion and express information that is complementary to the horizontal and vertical generation histories proposed before. Such morphological information refines syntactic categories based on their morpho-syntactic role, and captures linguistically motivated co-occurrences and dependencies manifested via, e.g., morpho-syntactic agreement.

This work aims at parsing MH and explores the empirical contribution of the three dimensions of parameters specified above. We present extensive experiments that gradually lead to improved performance as we extend the degree to which the three dimensions are exploited. Our best model uses all three dimensions of parametrization, and our best result is on a par with those achieved for MSA using a fully lexicalized parser and a much larger treebank. The remainder of this document is organized as follows. In section 2 we review characteristic aspects of MH (and other Semitic languages) and illustrate the special role of morphology and dependencies displayed by morpho-syntactic processes using the case of syntactic definiteness in MH. In section 3 we define our three-dimensional parametrization space. In section 4 we spell out the method and procedure for the empirical evaluation of one, two and three parametrization dimensions, and in section 5 we report and analyze results for different parametrization choices. Finally, section 6 discusses related work and in section 7 we summarize and conclude.

2 Dimensions of Modern Hebrew Syntax

Parsing MH is in its infancy. Although a syntactically annotated corpus has been available for quite some time (Sima'an et al., 2001), we know of only two studies attempting to parse MH using statistical methods (see section 6). One reason for the sparseness in this field is that the adaptation of existing models to parsing MH is technically involved yet does not guarantee to yield comparable results as the processes that license grammatical structures of phrases and sentences in MH differ from those assumed for English. This section outlines differences between English and MH and discusses their reflection in the MH treebank annotation scheme. We argue that on top of syntactic processes exploited by current parsers there is an orthogonal morpho-syntactic dimension which is invaluable for syntactic disambiguation, and it can be effectively learned using simple treebank grammars.

2.1 Modern Hebrew Structure

Phrases and sentences in MH, as well as in 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 exclusively syntactic and have to do with rhetorical and pragmatic factors as well.[2]

[1]The learning curves over increasing training data (e.g., for German (Dubey and Keller, 2003)) show that treebank size cannot be the sole factor to account for the inferior performance.

[2]See, for instance, (Melnik, 2002) for an Information Structure-syntactic account of verb initial sentences.

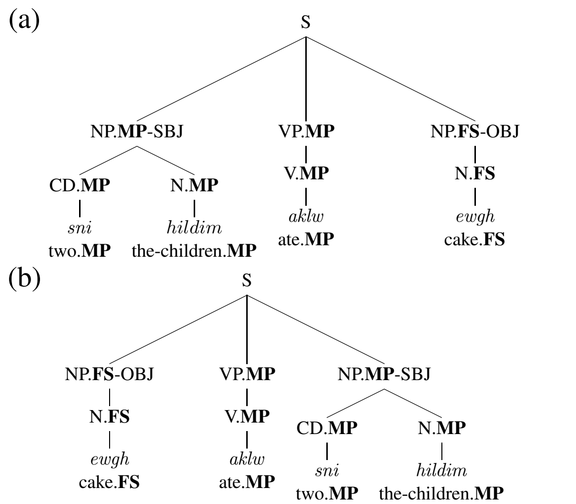


Figure 1: **Word Order and Agreement Features in MH Phrases:** Agreement on **MP** features reveals the subject-predicate dependency between surface forms and their dominating constituents in a variable phrase-structure (marking **M**(asculine), **F**(eminine), **S**(ingular), **P**(lural).)

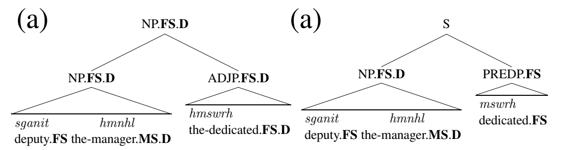


Figure 2: **Definiteness in MH as a Phrase-Level Agreement Feature:** Agreement on definiteness helps to determine the internal structure of a higher level NP (a), and the absence thereof helps to determine the attachment to a predicate in a verb-less sentence (b) (marking **D**(efiniteness))

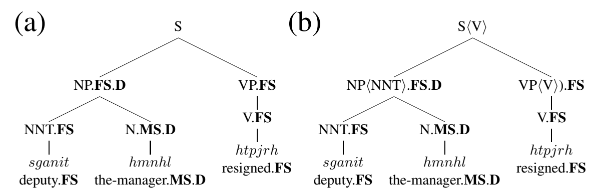


Figure 3: **Phrase-Level Agreement Features and Head-Dependencies in MH:** The direction of percolating definiteness in MH is distinct of that of the head (marking ⟨head-tag⟩)

It would be too strong a claim, however, to classify MH (and similar languages) as a free-word-order language in the canonical sense. The level of freedom in the order and number of internal constituents varies between syntactic categories. Within a verb phrase or a sentential clause, for instance, the order of constituents obeys less strict rules than within, e.g., a noun phrase.[3] Figure 1 illustrates two syntactic structures that express the same grammatical relations yet vary in their internal order of constituents. Within the noun phrase constituents, however, determiners always precede nouns.

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 between surface forms. Figure 1 also shows that agreement features help to reveal such relations between higher levels of constituents as well.

Determining the child constituents that contribute each of the features is not a trivial matter either. To illustrate the extent and the complexity of that matter let us consider *definiteness* in MH, which is morphologically marked (as an h prefix to the stem, glossed here explicitly as "the-") and behaves as a syntactic property (Danon, 2001). Definite noun-phrases exhibit agreement with other modifying phrases, and such agreement helps to determine the internal structure, labels, and the correct level of attachment as illustrated in figure 2. The agreement on definiteness helps to determine the internal structure of noun phrases 2(a), and the absence thereof helps in determining the attachment to predicates in verb-less sentences, as in 2(b). Finally, definiteness may be percolated from a different form than the one determining the gender and number of a phrase. In figure 3(a), for instance, the definiteness feature (marked as D) percolates from '$hmnhl$' (the-manager.MS.D) while the gender and number are percolated from '$sganit$' (deputy.FS). The direction of percolation of definiteness may be distinct of that of percolating head information, as can be seem in figure 3(b). (The direction of head-dependencies in MH typically coincides with that of percolating gender.)

To summarize, agreement features are helpful in analyzing and disambiguating syntactic structures in MH, not only at the lexical level, but also at higher levels of constituency. In MH, features percolated from different surface forms jointly determine the features of higher-level constituents, and such features manifest multiple dependencies, which in turn cannot be collapsed onto a single head.

[3]See (Wintner, 2000) and (Goldberg et al., 2006) for formal and statistical accounts (respectively) of noun phrases in MH.

2.2 The Modern Hebrew Treebank Scheme

The annotation scheme of version 2.0 of the MH treebank (Sima'an et al., 2001)[4] aims to capture the morphological and syntactic properties of MH just described. This results in several aspects that distinguish the MH treebank from, e.g., the WSJ Penn treebank annotation scheme (Marcus et al., 1994).

The MH treebank is built over word segments. This means that 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 segment corresponds to a single POS tag. (This in turn means that prefixes marking determiners, relativizers, prepositions and definite articles are segmented away and appear as leaves in a syntactic parse tree.) 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 pre-defined syntactic dependencies (Krymolowski et al., 2007). Since agreement features of non-terminal constituents may be contributed by more than one child, the annotation scheme defines multiple dependency labels that guide the percolation of the different features higher up the tree. Definiteness in the MH treebank is treated as a segment at the POS tags level and as a feature at the level of non-terminals. As any other feature, it is percolated higher up the tree according to marked dependency labels. Table 1 lists the features and values annotated on top of syntactic categories and table 2 describes the dependencies according to which these features are percolated from child constituents to their parents.

In order to comply with the flexible phrase structure in MH, clausal categories (S, SBAR and FRAG and their corresponding interrogatives SQ, SQBAR 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, meaning that all inflected verb forms are represented as unary productions under an inflected VP. NP and PP are annotated

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

Table 1: **Features and Values in the MH Treebank**

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

Table 2: **Dependency Labels in the MH Treebank**

as nested structures capturing the recursive structure of construct-state nouns, numerical expressions and possession. An additional category, PREDP, is added in the treebank scheme to account for sentences in MH that lack a copular element, and it may also be decorated with inflectional features agreeing with the subject. The MH treebank scheme also features null elements that mark traces and additional labels that mark functional features (e.g., SBJ,OBJ) which we strip off and ignore throughout this study.

Morphological features percolated up the tree manifest dependencies that are marked locally yet have a global effect. We propose to learn treebank grammars in which the syntactic categories are augmented with morphological features at all levels of the hierarchy. This allows to learn finer-grained categories with subtle differences in their syntactic behavior and to capture non-independence between certain parts of the syntactic parse-tree.

3 Refining the Parameter Space

(Klein and Manning, 2003) argue that parent encoding on top of syntactic categories and RHS markovization of CFG productions are two instances of the same idea, namely that of encoding the generation history of a node to a varying degree. They subsequently describe two dimensions that define their parameters' space. The *vertical* dimension (v), capturing the history of the node's ancestors in a top-

[4] Version 2.0 of the MH treebank is publicly available at `http://mila.cs.technion.ac.il/english/index.html` along with a complete overview of the MH annotation scheme and illustrative examples (Krymolowski et al., 2007).

down generation process (e.g., its parent and grandparent), and the *horizontal* dimension (h), capturing the previously generated horizontal ancestors of a node (effectively, its sisters) in a head-outward generation process. By varying the value of h and v along this two-dimensional grid they improve performance of their induced treebank grammar.

Formally, the probability of a parse tree π is calculated as the probability of its derivation, the sequential application of rewrite rules. This in turn is calculated as the product of rules' probabilities, approximated by assuming independence between them $P(\pi) = \prod_i P(r_i|r_1 \circ ... \circ r_{i-1}) \approx \prod_i P(r_i)$. The vertical dimension v can be thought of as a function Ψ_0 selecting features from the generation history of the constituent thus restoring selected dependencies:

$$P(r_i) = P(r_i|\Psi_0(r_1 \circ .. \circ r_{i-1}))$$

The horizontal dimension h can be thought of as two functions Ψ_1, Ψ_2 over decomposed rules, where Ψ_1 selects hidden internal features of the parent, and Ψ_2 selects previously generated sisters in a head-outward Markovian process (we retain here the assumption that the head child H always matters).

$$P(r_i) = P_h(H|\Psi_1(LHS(r_i))) \\ \times \prod_{C \in RHS(r_i)-H} P_C(C|\Psi_2(RHS(r_i)), H)$$

The fact that the default notion of a treebank grammar takes $v = 1$ (i.e., $\Psi_0(r_1 \circ .. \circ r_{i-1}) = \emptyset$) and $h = \infty$ (RHS cannot decompose) is, according to Klein and Manning (2003), a historical accident.

We claim that languages with freeer word order and richer morphology call for an additional dimension of parametrization. The additional parameter shows to what extent morphological features encoded in a specialized structure back up the derivation of the tree. This dimension can be thought of as a function Ψ_3 selecting aspects of morphological orthogonal analysis of the rules, where MA denotes morphological analysis of the syntactic categories in both LHS and RHS of the rule.

$$P(r_i) = P(r_i|\Psi_3(MA(r_i)))$$

The fact that in current parsers $\Phi_3(MA(r_i)) = \emptyset$ is, we claim, another historical accident. Parsing English is quite remarkable in that it can be done with impoverished morphological treatment, but for languages in which morphological processes are more pertinent, we argue, bi-dimensional parametrization shall not suffice.

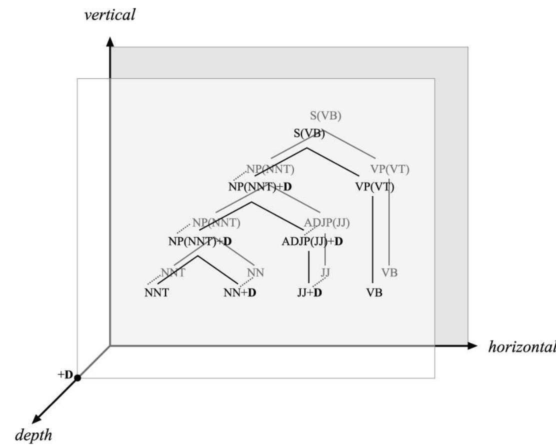
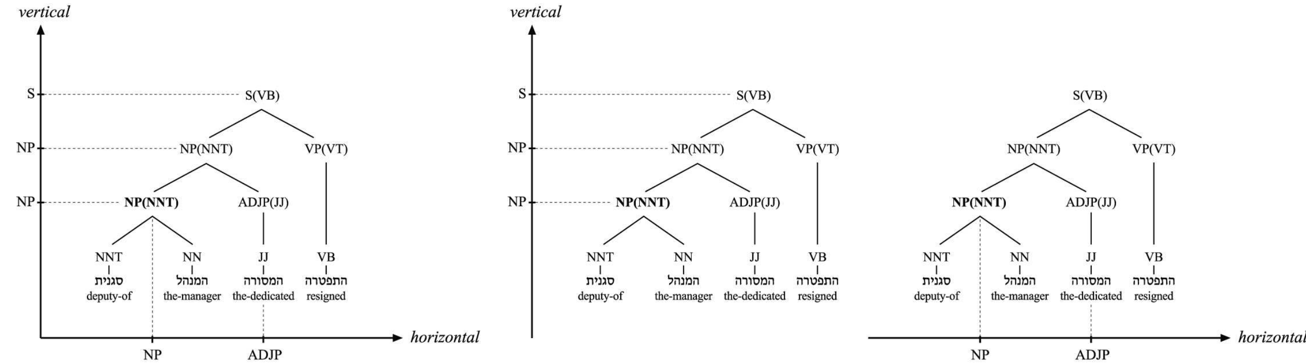


Figure 4: **The Three-Dimensional Parametrization Space**

The emerging picture is as follows. Bare-category skeletons reside in a bi-dimensional parametrization space (figure 3(a)) in which the vertical (figure 3(b)) and horizontal (figure 3(c)) parameter instantiations elaborate the generation history of a non-terminal node. Specialized structures enriched with (an increasing amount of) morphological features reside deeper along a third dimension we refer to as *depth* (d). Figure 4 illustrates an instantiation of $d = 1$ with a single definiteness feature. Higher d values would imply adding more (accumulating) features.

Klein and Manning (2003) view the *vertical* and *horizontal* parametrization dimensions as implementing *external* and *internal* annotation strategies respectively. External parameters indicate features of the external environment that influence the node's expansion possibilities, and internal parameters mark aspects of hidden internal content which influence constituents' external distribution. We view the third dimension of parametrization as implementing a *relational* strategy of annotation encoding the way different constituents may combine to form phrases and sentences. In a bottom up process this annotation strategy imposes soft constraints on a the top-down head-outward generation process. Figure 6(a) focuses on a selected NP node highlighted in figure 4 and shows its expansion possibilities in three dimensions. Figure 6(b) illustrates how the depth expansion interacts with both parent anno-



(a) **The** *horizontal/vertical* **Grid** (b) **The** *vertical* **dimension** (c) **The** *horizontal* **dimension**

Figure 5: **The Two-Dimensional Space:** The horizontal and vertical dimensions outlined by (Klein and Manning, 2003)

tation and neighbor dependencies thereby affecting both distributions.

3.1 A Note on State-Splits

Recent studies (Klein and Manning, 2003; Matsuzaki et al., 2005; Prescher, 2005; Petrov et al., 2006) suggest that category-splits help in enhancing the performance of treebank grammars, and a previous study on MH (Tsarfaty, 2006) outlines specific POS-tags splits that improve MH parsing accuracy. Yet, there is a major difference between category-splits, whether manually or automatically acquired, and the kind of state-splits that arise from agreement features that refine phrasal categories. While category-splits aim at each category in isolation, agreement features apply to a whole set of categories all at once, thereby capturing refinement of the categories as well as linguistically motivated co-occurrences between them. Individual category-splits are viewed as taking place in a two-dimensional space and it is hard to analyze and empirically evaluate their interaction with other annotation strategies. Here we propose a principled way to statistically model the interaction between different linguistic processes that license grammatical structures and empirically contrast their contribution.

3.2 A Note on Stochastic AV grammars

The practice of having morphological features orthogonal to a constituency structure is not a new one and is familiar from formal theories of syntax such as HPSG (Sag et al., 2003) and LFG (Kaplan and Bresnan, 1982). Here we propose to reframe systematic morphological decoration of syntactic categories at all levels of the hierarchy as

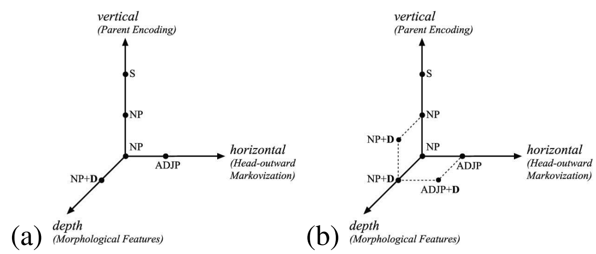


Figure 6: **The Expansion Possibilities of a Non-Terminal Node:** Expanding the NP from figure 4 in a three-dimensional parameterization Space

an additional dimension of statistical estimation for learning unlexicalized treebank PCFGs. Our proposal deviates from various stochastic extensions of such constraints-based grammatical formalisms (cf. (Abney, 1997)) and has the advantage of elegantly bypassing the issue of loosing probability mass to failed derivations due to unification failures. To the best of our knowledge, this proposal has not been empirically explored before.

4 Experimental Setup

Our goal is to determine the optimal strategy for learning treebank grammars for MH and to contrast it with bi-dimensional strategies explored for English. The methodology we use is adopted from (Klein and Manning, 2003) and our procedure is identical to the one described in (Johnson, 1998). We define transformations over the treebank that accept as input specific points in the (h, v, d) space depicted in figure 7. We use the transformed training sets for learning different treebank PCFGs which we then used to parse unseen sentences, and detransform the parses for the purpose of evaluation.[5]

[5]Previous studied on MH used different portions of the treebank and its annotation scheme due to its gradual development

Data We use version 2.0 of the MH treebank which consists of 6501 sentences from the daily newspaper 'Ha'aretz'. We 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 all used for training. After removing empty sentences, sentences with uneven bracketing and sentences that do not match the annotation scheme[6] 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). 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 the development set for parameter-tuning and error analysis and use the test set only for confirming our best results.

Models The models we implement use one-, two- or three-dimensional parametrization and different instantiation of values thereof. (Due to the small size of our data set we only use the values $\{0, 1\}$ as possible instantiations.)

The v dimension is implemented using a transform as in (Johnson, 1998) where $v = 0$ corresponds to bare syntactic categories and $v = 1$ augments node labels with the label of their parent node.

The h dimension is peculiar in that it distinguishes PCFGs ($h = \infty$), where RHS cannot decompose, from their head-driven unlexicalized variety. To implement $h \neq \infty$ we use a PCFG transformation emulating (Collins, 2003)'s first model, in which sisters are generated conditioned on the head tag and a simple 'distance' function (Hageloh, 2007).[7] The instantiated value of h then selects the number of previously generated (non-head) sisters to be taken into account when generating the next sister in a Markovian process (Ψ_2 in our formal exposition).

The d dimension we proposed is implemented using a transformation that augments syntactic categories with morphological features percolated up the tree. We use $d = 0$ to select bare syntactic categories and instantiate $d = 1$ with the definiteness feature. The decision to select definiteness (rather than, e.g., gender or number) is rather pragmatic as its direction of percolation may be distinct of head information and the question remains whether the combination of such non-overlapping dependencies is instrumental for parsing MH.

Our baseline model is a vanilla treebank PCFG as described in (Charniak, 1996) which we locate on the $(\infty, 0, 0)$ point of our coordinates-system. In a first set of experiments we implement simple PCFG extensions of the treebank trees based on selected points on the (∞, v, d) plain. In a second set of experiments we use an unlexicalized head-driven baseline à la (Collins, 2003) located on the $(0, 0, 0)$ coordinate. We transform the treebank trees in correspondence with different points in the three-dimensional space defined by (h, v, d). The models we implement are marked in the coordinate-system depicted in figure 7. The implementation details of the transformations we use are spelled out in tables 3–4.

Procedure We implement different models that correspond to different instantiations of h, v and d. For each instantiation we transform the training set and learn a PCFG using Maximum Likelihood estimates, and we use BitPar (Schmidt, 2004), an efficient general-purpose parser, to parse unseen sentences. The input to the parser is a sequence of word segments where each segment corresponds to a single POS tag, possibly decorated with morphological features. This setup assumes partial morphological disambiguation (namely, segmentation) but crucially we do *not* disambiguate their respective POS categories. This setup is more appropriate for using general-purpose parsing tools and it makes our results comparable to studies in other languages.[8]

process. As the MH treebank is approaching maturity we feel that the time is ripe to standardize its use for MH statistical parsing. The software we implemented will be made available for non-commercial use upon request to the author(s) and the feature percolation software by (Krymolowski et al., 2007) is publicly available through the Knowledge Center for Processing Hebrew. By this we hope to increase the interest in MH within the parsing community and to facilitate the application of more sophisticated models by cutting down on setup time.

[6] Marked as "NO_MATCH" in the treebank.

[7] A formal overview of the transformation and its correspondence to (Collins, 2003)'s models is available at (Hageloh, 2007). We use the distance function defined therein, marking the direction and whether it is the first node to be generated.

[8] Our working assumption is that better performance of a parsing model in our setup will improve performance also

Transliterate The lexical items (leaves) in the MH treebank are written left-to-write and are encoded in utf8. A transliteration software is used to convert the utf encoding into Latin characters and to reverse their order, essentially allowing for standard left-to-right processing.

Correct The manual annotation resulted in unavoidable errors in the annotation scheme, such as typos (e.g., SQBQR instead of SQBAR) wrong delimiters (e.g., “-” instead of “_”) or wrong feature order (e.g., number-gender instead of gender-number). We used an automatic script to detect these error, we manually determine their correction. Then we created an automatic script to apply all fixes (57 errors in 1% sentences).

Re-attach VB elements are attached by convention to a VP which inherits its morphological features. 9 VB instances in the treebank are mistakenly attached to an S parent without an intermediate VP level. Our software re-attaches those VB elements to a VP parent and percolates its morphological features.

Disjoint Due to recursive processes of generating noun phrases and numerical expression (*smixut)* in MH the sets of POS and syntactic categories are not disjoint. This is a major concern for PCFG parsers that assume disjoint sets of pre- and non-terminals. The overlap between the sets also introduces additional infinite derivations to which we loose probability mass. Our software takes care to decorate POS categories used as non-terminal with an additional “P”, creating a new set of categories encoding partial derivations.

Lexicalize A pre-condition for applying horizontal parameterizations à la Collins is the annotation of heads of syntactic phrases. The treebank provided by the knowledge center does not define unique heads, but rather, mark multiple dependencies for some categories and none for others. Our software uses rules for choosing the syntactic head according to specified dependencies and a head table when none are specified.

Linearize In order to implement the head-outward constituents' generation process we use software made available to us by (Hageloh, 2007) which converts PCFG production such as the generation of a head is followed by left and right markovized derivation processes. We used two versions of Markovization, one which conditions only on the head and a distance function, and another which conditions also on immediately neighboring sister(s).

Decorate Our software implements an additional general transform which selects the features that are to be annotated on top of syntactic categories to implement various parametrization decisions. This transform can be used for, e.g., displaying parent information, selecting morphological features, etc.

Table 3: **Transforms over the MH Treebank:** We clean and correct the treebank using **Transliterate, Correct, Re-attach** and **Disjoint**, and transform the training set according to certain parametrization decisions using **Lexicalize, Linearize** and **Decorate**.

Smoothing pre-terminal rules is done explicitly by collecting statistics on “rare word” occurrences and providing the parser with possible open class categories and their corresponding frequency counts. The frequency threshold defining “rare words” was tuned empirically and set to 1. The resulting test parses are detransformed and to skeletal constituent structures, and are compared against the gold parses to evaluate parsing accuracy.

Evaluation We evaluate our models using EVALB in accordance with standard PARSEVAL evaluation metrics. The evaluation of all models focuses on Labeled Precision and Recall considering bare syntactic categories (stripping off all morphological or parental features and removing intermediate nodes for linearization). We report the average F-measure for sentences of length up to 40 and for all sentences ($F_{\leq 40}$ and F_{All} respectively). We report the results for two evaluation options, once including punctuation marks (WP) and once excluding them (WOP).

within an integrated model for morphological and syntactic disambiguation in the spirit of (Tsarfaty, 2006). We conjecture that the kind of models developed here which takes into account morphological information is more appropriate for the morphological disambiguation task defined therein.

5 Results

Our baseline for the first set of experiments is a vanilla PCFG as described in (Charniak, 1996) (without a preceding POS tagging phase and without right branching corrections). We transform the treebank trees based on various points in the (∞, v, d) two-dimensional space to evaluate the performance of the resulting PCFG extensions.

Table 5 reports the accuracy results for all models on section 0 (*devset*) of the treebank. The accuracy results for the vanilla PCFG are approximately 10% lower than reported by (Charniak, 1996) for English demonstrating that parsing MH using the currently available treebank is a harder task. For all unlexicalized extensions learned from the transfromed treebanks, the resulting grammars show enhanced disambiguation capabilities and improved parsing accuracy. We observe that the vertical dimension contributes the most from both one-dimensional mod-

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 feature percolation	**Decorate(definiteness)**

Table 4: **Implementing Different Parametrization Options using Transforms**

Implementation	(h, v, d)	F_{ALL} WP	$F_{\leq 40}$ WP	F_{ALL} WOP	$F_{\leq 40}$ WOP
PCFG	$(\infty, 0, 0)$	65.17	66.63	66.17	67.7
PA	$(\infty, 0, 1)$	70.6	71.96	70.96	72.18
DEF	$(\infty, 1, 0)$	67.53	68.78	68.82	70.06
PA+DEF	$(\infty, 1, 1)$	72.63	73.89	73.01	**74.11**

Table 5: **PCFG Two-Dimensional Extensions:** Accuracy results for parsing the *devest* (section 0)

els. A qualitative error analysis reveals that parent annotation strategy distinguishes effectively various kinds of distributions clustered together under a single category. For example, S categories that appear under TOP tend to be more flat than S categories appearing under SBAR (SBAR clauses typically generate a non-finite VP node under which additional PP modifiers can be attached).

Orthogonal morphological marking provide additional information that is indicative of the kind of dependencies that exist between a category and its various child constituents, and we see that the d dimension instantiated with *definiteness* not only contribute more than 2% to the overall parsing accuracy of a vanilla PCFG, but also contributes as much to the improvement obtained from a treebank already annotated with the vertical dimension. The contributions are thus additive providing preliminary empirical support to our claim that these two dimensions provide information that is complementary.

In our next set of experiments we evaluate the contribution of the depth dimension to extensions of the head-driven unlexicalized variety à la (Collins, 2003). We set our baseline at the $(0, 0, 0)$ coordinate and evaluate models that combine one, two and three dimensions of parametrization. Table 6 shows the accuracy results for parsing section 0 using the resulting models.

The first outcome of these experiments is that our new baseline improves on the accuracy results of a simple treebank PCFG. This result indicates that head-dependencies which play a role in determining grammatical structures in English are also instrumental for parsing MH. However, the marginal contribution of the head-driven variation is surprisingly low. Next we observe that for one-dimensional models the vertical dimension still contributes the most to parsing accuracy. However, morphological information represented by the depth dimension contributes more to parsing accuracy than information concerning immediately preceding sisters on the horizontal dimension. This outcome is consistent with our observation that the grammar of MH puts less significance on the position of constituents relative to one others and that morphological information is more indicative of the kind of syntactic relations that appear between them. For two-dimensional models, incorporating the depth dimension (orthogonal morphological marking) is better than not doing so, and relying solely on horizontal/vertical parameters performs slightly worse than the vertical/depth combination. The best performing model for two-dimensional head-driven extensions is the one combining vertical history and morphological depth. This is again consistent with the properties of MH highlighted in section 2 — parental information gives cues about the possible expansion on the current node, and morphological information indicates possible interrelation between child constituents that may be generated in a flexible order.

Our second set of experiments shows that a three-dimensional annotation strategy strikes the best balance between bias and variance and achieves the best accuracy results among all models. Different dimensions provide different sorts of information which are complementary, resulting in a model that is capable of generalizing better. The total error reduction from a plain PCFG is more than 20%, and our best result is on a par with those achieved for other languages (e.g., 75% for MSA).

Implementation	Params (h, v, d)	F_{ALL} WP	$F_{\leq 40}$ WP	F_{ALL} WOP	$F_{\leq 40}$ WOP
DIST	$(0, 0, 0)$	66.56	68.20	67.59	**69.24**
MRK	$(1, 0, 0)$	66.69	68.14	67.93	69.37
PA	$(0, 1, 0)$	68.87	70.48	69.64	70.91
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	$(1, 0, 1)$	69.46	70.79	71.05	72.37
PA+DEF	$(0, 1, 1)$	71.15	72.34	71.98	**72.91**
PA+MRK+DEF	$(1, 1, 1)$	72.34	73.63	73.27	**74.41**

Table 6: **Head-Driven Three-Dimensional Extensions:** Accuracy results for parsing the *devest* (section 0)

Implementation	Params (h, v, d)	F_{ALL} WP	$F_{\leq 40}$ WP	F_{ALL} WOP	$F_{\leq 40}$ WOP
PCFG	$(\infty, 0, 0)$	65.08	67.31	65.82	68.22
PCFG+PA+DEF	$(\infty, 1, 1)$	72.26	74.46	72.42	**74.52**
DIST	$(0, 0, 0)$	66.33	68.79	67.06	69.47
PA+MRK+DEF	$(1, 1, 1)$	72.64	74.64	73.21	**75.25**

Table 7: **PCFG and Head-Driven Unlexicalized Models:** Accuracy Results for parsing the *testst* (section 12)

Figure 8 shows the $F_{All}(WOP)$ results for all models we implemented. In general, we see that for parsing MH higher dimensionality is better. Moreover, we see that for all points on the $(v, h, 0)$ plain the corresponding models on the $(v, h, 1)$ plain always perform better. We further see that the contribution of the depth dimension to a parent annotated PCFG can compensate, to a large extent on the lack of head-dependency information. These accumulative results, then, provide empirical evidence to the importance of morphological and morpho-syntactic processes such as definiteness for syntactic analysis and disambiguation as argued for in section 2.

We confirm our results on the *testset* and report in table 7 our results on section 12 of the treebank. The performance has slightly increased and we obtain better results for our best strategy. We retain the high error-reduction rate and propose our best result, 75.25% for sentences of length ≤ 40, as an empirically established string baseline on the performance of treebank grammars for MH.

6 Related Work

The MH treebank (Sima'an et al., 2001), a morphologically and syntactically annotated corpus, has been successfully used for various NLP tasks such as morphological disambiguation, POS tagging (Bar-Haim et al., 2007) and NP chunking (Goldberg et al., 2006). However its use for statistical parsing has been more scarce and less successful. The only previous studies attempting to parse MH we know of are (Sima'an et al., 2001), applying a variation of the DOP tree-gram model to 500 sentences, and (Tsarfaty, 2006), using a treebank PCFG in an integrated system for morphological and syntactic disambiguation.[9] The adaptation of state-of-the-art parsing models to MH is not immediate as the flat variable structures of phrases are hard to parse and a plentiful of morphological features that would facilitate disambiguation are not exploited by currently available parsers. Also, the MH treebank is much smaller than the ones for, e.g., English (Marcus et al., 1994) and Arabic (Maamouri and Bies, 2004), making it hard to apply data-intensive methods such as the all-subtrees approach (Bod, 1992) or full lexicalization (Collins, 2003). Our best performing model incorporates three dimensions of parametrization and our best result (75.25%) is similar to the one obtained by the parser of (Bikel, 2004) for Modern Standard Arabic (75%) using a fully lexicalized model and a training corpus about three times as large as our newest MH treebank.

This work has shown that devising an adequate baseline for parsing MH requires more than simple category-splits and sophisticated head-driven extensions, and our results provide preliminary evidence for the variation in performance of different parametrization strategies relative to the properties and structure of a given language. The comparison with parsing accuracy for MSA suggests that parametrizing an orthogonal depth dimension may be able to compensate, to some extent, on the lack of sister-dependencies, lexical information, and perhaps even the lack of annotated data, but establishing empirically its contribution to parsing MSA is a matter for further research. In the future we intend to further investigate the significance of the depth dimension by extending our models to include more morphological features, more variation in the pa-

[9]Both studies acheived between 60%–70% accuracy, however the results are not comparable to our study because of the use of different training sets, different annotation conventions, and different evaluation schemes.

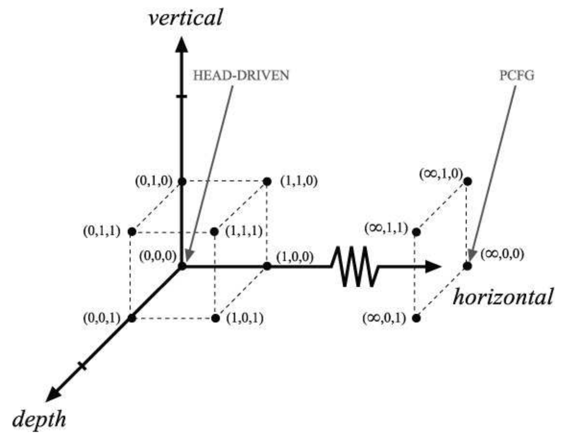


Figure 7: **All Models:** Locating Unlexicalized Parsing Models in a Three-Dimensional Parametrization Space

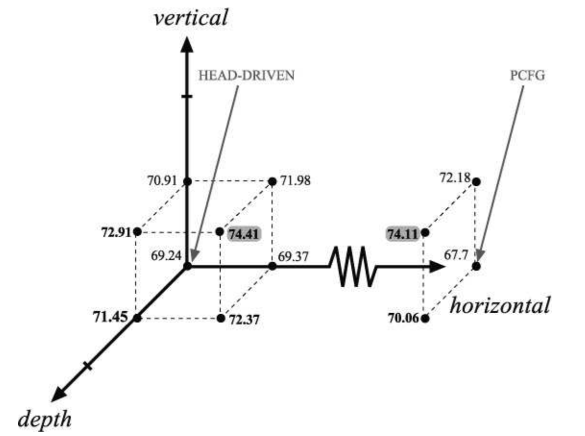


Figure 8: **All Results:** Parsing Results for Unlexicalized Models in a Three-Dimensional Parametrization Space

rameter space, and applications to more languages.

7 Conclusion

Morphologically rich languages introduce a new dimension into the expansion possibilities of a non-terminal node in a syntactic parse tree. This dimension is orthogonal to the vertical (Collins, 2003) and horizontal (Johnson, 1998) dimensions previously outlined by Klein and Manning (2003), and it cannot be collapsed into any one of the previous two. These additional dependencies exist alongside the syntactic head dependency and are attested using morphosyntactic phenomena such as long distance agreement. We demonstrate using syntactic definiteness in MH that incorporating morphologically marked features as a third, orthogonal dimension for annotating syntactic categories is invaluable for weakening the independence assumptions implicit in a treebank PCFG and increasing the model's disambiguation capabilities. Using a three-dimensional model we establish a new, stronger, lower bound on the performance of unlexicalized parsing models for Modern Hebrew, comparable to those achieved for other languages (Czech, Chinese, German and Arabic) with much larger corpora.

Tuning the dimensions and value of the parameters for learning treebank grammars is largely an empirical matter, and we do not wish to claim here that a three-dimensional annotation strategy is the best for any given language. Rather, we argue that for different languages different optimal parametrization strategies may apply. MH is not a free-word-order language in the canonical sense, and our qualitative analysis shows that all dimensions contribute to the models' disambiguation capabilities. Orthogonal dimensions provide complementary information that is invaluable for the parsing process to the extent that the relevant linguistic phenomena license grammatical structures in the language. Our results point out a principled way to quantitatively characterizing differences between languages, thus guiding the selection of parameters for the development of annotated resources, custom parsers and cross-linguistic robust parsing engines.

Acknowledgments We thank the Knowledge Center for Processing Hebrew and Dalia Bojan for providing us with the newest version of the MH treebank. We are particularly grateful to the development team of version 2.0, Adi Mile'a and Yuval Krymolowsky, supervised by Yoad Winter for continued collaboration and technical support. We further thank Felix Hageloh for allowing us to use the software resulting from his M.Sc. thesis work. We also like to thank Remko Scha, Jelle Zuidema, Yoav Seginer and three anonymous reviewers for helpful comments on the text, and Noa Tsarfaty for technical help in the graphical display. The work of the first author is funded by the Netherlands Organization for Scientific Research (NWO), grant number 017.001.271, for which we are grateful.

References

S. Abney. 1997. Stochastic Attribute-Value Grammars. *Computational Linguistics*, 23 (4):597–618.

R. Bar-Haim, K. Sima'an, and Y. Winter. 2007. Part-of-Speech Tagging of Modern Hebrew Text. *Journal of Natural Language Engineering*.

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

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

R. Bod. 1992. Data Oriented Parsing. In *Proceedings of COLING*.

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

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

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

M. Collins. 2003. Head-Driven Statistical Models for Natural Language Parsing. *Computational Linguistics*, 29(4).

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

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

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

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

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

R. Kaplan and J. Bresnan. 1982. Lexical-Functional Grammar: A formal system for grammatical representation. In J. Bresnan, editor, *The Mental Representation of Grammatical Relations*, Cambridge, MA. The MIT Press.

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

Y. Krymolowski, Y. Adiel, N. Guthmann, S. Kenan, A. Milea, N. Nativ, R. Tenzman, and P. Veisberg. 2007. Treebank Annotation Guide. MILA, Knowledge Center for Hebrew Processing.

M. Maamouri and A. Bies. 2004. Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools. In *Proceedings of COLING*.

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.

T. Matsuzaki, Y. Miyao, and J. Tsujii. 2005. Probabilistic CFG with Latent Annotations. In *Proceedings of ACL'05*.

N. Melnik. 2002. *Verb-Initial Constructions in Modern Hebrew*. Ph.D. thesis, Berkeley University of California.

S. Petrov, L. Barrett, R. Thibaux, and D. Klein. 2006. Learning Accurate, Compact, and Interpretable Tree Annotation. In *Proceedings of ACL-COLING*, pages 433–440, Sydney, Australia, July.

D. Prescher. 2005. Head-Driven PCFGs with Latent-Head Statistics. In *In Proceedings of the International Workshop on Parsing Technologies*.

I. A. Sag, T. Wasow, and E. M. Bender. 2003. *Syntactic Theory: A Formal Introduction*. CSLI Publications, address, second edition.

H. Schmidt. 2004. Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors. In *Proceedings of COLING*, Geneva, Switzerland.

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

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

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