# An Alternative to Head-Driven Approaches for Parsing a (Relatively) Free Word-Order Language

**Reut Tsarfaty** **Khalil Sima'an** **Remko Scha**
Institute for Logic Language and Computation
University of Amsterdam
{r.tsarfaty,k.simaan,r.scha}@uva.nl

## Abstract

Applying statistical parsers developed for English to languages with freer word-order has turned out to be harder than expected. This paper investigates the adequacy of different statistical parsing models for dealing with a (relatively) free word-order language. We show that the recently proposed *Relational-Realizational (RR)* model consistently outperforms state-of-the-art *Head-Driven (HD)* models on the Hebrew Treebank. Our analysis reveals a weakness of HD models: their intrinsic focus on configurational information. We conclude that the form-function separation ingrained in RR models makes them better suited for parsing nonconfigurational phenomena.

## 1 Introduction

Parsing technology has come a long way since Charniak (1996) demonstrated that a simple treebank PCFG performs better than any other parser (with $F_1 75$ accuracy) on parsing the WSJ Penn treebank (Marcus et al., 1993). Treebank Grammars (Scha, 1990; Charniak, 1996) trained on large corpora nowadays present the best available means to parse natural language text.

The performance curve for parsing the WSJ was a steep one at first, as the incorporation of notions such as *head*, *distance*, *subcategorization* (Charniak, 1997; Collins, 1999) brought about a dramatic increase in parsing accuracy to the level of $F_1 88$. Discriminative approaches, Data-Oriented Parsing ('all-subtrees') approaches, and self-training techniques brought further improvements, and recent results are starting to level off at around $F_1 92.1$ (McClosky et al., 2008).

As the interest of the NLP community grows to encompass more languages, we observe efforts towards adapting an English parser for parsing other languages (e.g., (Collins et al., 1999)), or towards designing a language-independent framework based on principles underlying the models for parsing English (Bikel, 2002). The performance curve for parsing other languages with these models looks rather different. A case in point is Modern Standard Arabic. Since the initial effort of (Bikel, 2002) to parse the Arabic treebank (Maamouri et al., 2004), which yielded $F_1 75$ accuracy, four years and successive revisions have led to no more than $F_1 79$ (Maamouri et al., 2008).

This pattern from Arabic is not peculiar. The level of state-of-the-art results for other languages still lags behind those for English, even after putting considerable effort into the adaptation.[1] Given that these languages are inherently different from English and from one another, it appears that we cannot avoid a question concerning the *adequacy* of the models used to parse them. That is, given the properties of a language, which modeling strategy would be appropriate for parsing it?

Until recently, there has been practically no computationally affordable alternative to the *Head-Driven (HD)* approach in the development of phrase-structure based statistical parsing models. Recently, we proposed the *Relational-Realizational (RR)* approach that rests upon different premises (Tsarfaty and Sima'an, 2008). The question of how the RR model fares against the HD models that have so far been predominantly used has never been tackled. Yet, it is precisely such a comparison that can shed new light on the question of adequacy we posed above.

Empirically quantifying the effects of different modeling choices has been addressed for English by, e.g., (Johnson, 1998; Klein and Manning, 2003), and for German by, e.g., (Dubey, 2004;

[1] Consider, e.g., "The PaGe shared task on parsing German" (Kubler, 2008), reporting $F_1 75, F_1 79, F_1 83$ for the participating parsers.

Rafferty and Manning, 2008). This paper provides an empirical systematic comparison of conceptually different modeling strategies with respect to parsing Hebrew. This comparison is intended to provide a first answer to the question of parser adequacy in the face of word-order freedom.

Our two empirical results are unequivocal. Firstly, RR models significantly outperform HD models (about 2 points absolute improvement in $F_1$) in parsing the Modern Hebrew treebank. In particular, RR models show better performance in identifying the constituents for which syntactic positions are relatively free. Secondly, we show a novel variation of the HD model, incorporating the *Relational* notions of the RR model, on the hypothesis that this might bridge the gap. The RR model remains superior.

Our post-experimental analysis shows that HD modeling is inherently problematic for parsing a language with freer word-order because of the hard-wiring of notions such as *left*, *right* and *distance from the head*. RR models, taking a principled approach towards capturing variable form-function correspondence patterns, are better suited for parsing *nonconfigurational* phenomena.

## 2 The Data

This section describes some properties of Modern Hebrew (henceforth, Hebrew) that make it significantly different from English. These properties affect the syntactic representations found in the Hebrew Treebank and the kind of syntactic phenomena a parser for Hebrew has to cope with.

Modern Hebrew is a Semitic language with a canonical SVO word-order pattern,[2] yet it allows considerable freedom in the placement of syntactic constituents in a clause. For example, linguistic elements of any kind may be fronted, triggering an inversion familiar from Germanic languages as in (1b) (*Triggered Inversion (TI)* in (Shlonsky, 1997)). Under some information structuring conditions, *Verb Initial (VI)* constructions are also allowed, as in (1c) (Melnik, 2002). All sentences in (1) thus mean "Dani gave the present to Dina", despite their different word-ordering.

(1) a. *dani natan et hamatana ledina*
Dani gave ACC the-present to-Dina

b. *et hamatana natan dani ledina*
ACC the-present gave Dani to-Dina

[2]SVO is an abbreviation for the Subject-Verb-Object type in the *basic word-order* typology of (Greenberg, 1963).

| Word Order | Frequency | Relative Frequency |
|---|---|---|
| SV | 1612 | 41% |
| VS | 1144 | 29% |
| No S | 624 | 16% |
| No V | 550 | 14% |

Table 1: **Modern Hebrew Predicative Clause-Types** in 3930 Predicative Matrix Clauses in the Training Set of the Modern Hebrew Treebank.

c. *natan dani et hamatana ledina*
gave Dani ACC the-present to-Dina

A corpus study we conducted on a fragment of the Modern Hebrew treebank reveals that although there is a significant number of subjects preceding verbs in simple (matrix) clauses (41%), there are also a fair number of sentences for which this order is reversed (29%), and there is evidence for other configurations, such as empty realization of subjects (16%) and non-verbal realization of predicates (14%).

In the face of such lack of consistency in its configurational position, the grammatical function *Object* in Hebrew is indicated by *Differential Object Marking (DOM)* (Aissen, 2003). NP objects in Hebrew are marked for *accusativity* (using the marker *et*) if they are also marked for *definiteness* (indicated by the prefix *ha*). So, in contrast with (2a)-(2b), the indefinite object renders (2c) ungrammatical, and the missing accusativity renders (2d) awkward. The fact that marking NP objects involves the joint contribution of multiple surface elements (*et*, *ha*) contributing features to the NP constituent is referred to as *extended exponence* (Matthews, 1993, p. 182).

(2) a. *dani natan matana ledina*
Dani gave present to-Dina
"Dani gave a present to Dina"

b. *dani natan et hamatana ledina*
Dani gave ACC the-present to-Dina
"Dani gave the present to Dina"

c. **dani natan et matana ledina*
Dani gave ACC present to-Dina

d. *??dani natan hamatana ledina*
Dani gave the-present to-Dina

These data pose a challenge to generative parsing models, as they would be required to generate alternative word-order patterns while maintaining a coherent pattern of object marking, encom-

passing the contribution of multiple surface exponents. The question this paper addresses is therefore what kind of modeling approach would be adequate for modeling the interplay between *syntax* and *morphology* in marking grammatical relations in Hebrew, as reflected by the sentence-pair (3). They both mean, roughly, "Dani gave the present to Dina yesterday; their word-order vary, but the pattern of object marking is retained.

(3) a. *dani natan etmol et hamatana ledina*
Dani gave yesterday ACC the-present to-Dina

b. *et hamatana natan etmol dani ledina*
ACC the-present gave yesterday dani to-dina

## 3 The Models

The different models we experiment with are all trained on syntactic structures annotated in the Modern Hebrew Treebank (Sima'an et al., 2001). The native representation of clause-level categories in the Treebank employs flat structures. This choice was made due to the lack of empirical evidence in Hebrew for grouping freely positioned syntactic elements to form a constituent.[3] In order to compensate for the ambiguity in the *interpretation* of flat structures, additional information such as morphological marking and grammatical function labels is added to the phrase-structure trees.

### 3.1 The *State-Splits* Approach

The simplest way to encode grammatical functions information on top of the phrase-structure representation in the treebank is by decorating non-terminal nodes with morphological or functional features, similarly to the rich representation format of syntactic categories in GPSG. This is the approach taken by the annotators of the Hebrew treebank in which information about morphological marking appears at multiple levels of constituency (Guthmann et al., 2009), and functional features (such as *subject*, *object*, etc.) decorate phrase-level constituent labels (Sima'an et al., 2001). The S-level representation of our example sentences (3a)–(3b) then would be as we depict in figure 1, which can be read off as feature-rich PCFG productions. We refer to this approach as the *State-Splits (SP)* approach, which serves as the baseline for the rest of our investigation.

### 3.2 The *Head-Driven* Approach

Following the linguistic wisdom that the internal organization of syntactic constituents revolves around their *heads*, *Head-Driven (HD)* models have been proposed by (Magerman, 1995; Charniak, 1997; Collins, 1999). In a generative HD model, the head daughter is generated first, conditioned on properties of the mother node. Then, sisters of the head daughter are generated conditioned on the head, typically by *left* and *right* generation processes. Overall, HD processes have the modeling advantage that they capture structurally-marked positions that characterize the *argument structure* of the sentence. The simplest possible process uses unigram probabilities, but (Klein and Manning, 2003) show that using *vertical* and *horizontal* Markovization improves parsing accuracy.[4]

An unlexicalized generative HD model will generate our two example sentences as we illustrate in figure 2. The generation of the context-free events in figure 1 is then broken down to seven different context-free parameters each, encoding head-parent and head-sister structural relationships — the latter mediated with a structurally-marked *delta* function ($\Delta_i$). The rich morphological representation of phrase-level NP objects (*+def/acc*), for instance, is conditioned on the *head* sister, its *direction*, and the *distance from the head* (check, e.g., nodes $\Delta_{L_1}, \Delta_{R_2}$).

### 3.3 The *Relational-Realizational* Approach

The *Relational-Realizational (RR)* parsing model of (Tsarfaty and Sima'an, 2008) similarly decomposes the generation of the context-free events in figure 1 into multiple independent parameters, but does so in a conceptually different way. Instead of decomposing a context-free event to *head* and *sisters*, the RR model is best viewed as a generative grammar that decomposes it to *form* and *function*.

The RR grammar first generates a set of grammatical functions depicting the *Relational Network (RN)* (Perlmutter, 1982) of the clause. This

[3] Such clauses are defined formally as *exocentric* in formal theories of syntax, and are used to describe syntactic structures in, e.g., Tagalog, Hungarian and Warlpiri (Bresnan, 2001, page 110). This flat representation format is characteristic of treebanks for other languages with relatively-free word-order as well, such as German (cf. (Kubler, 2008)).

[4] The success of Head-Driven models (Charniak, 1997; Collins, 2003) was initially attributed to the fact that they were fully lexicalized, but (Klein and Manning, 2003) show that an unlexicalized model combining Head-Driven Markovian processes with linguistically motivated state-splits can approach the performance of fully lexicalized models.

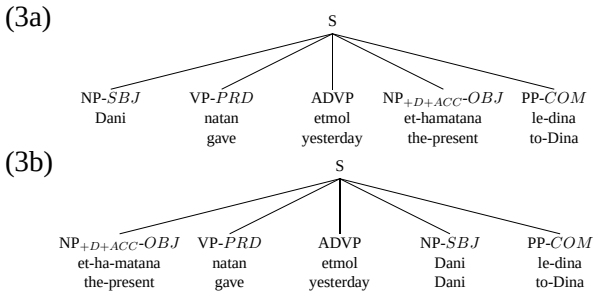


Figure 1: The *State-Splits* Approach for Ex. (3)

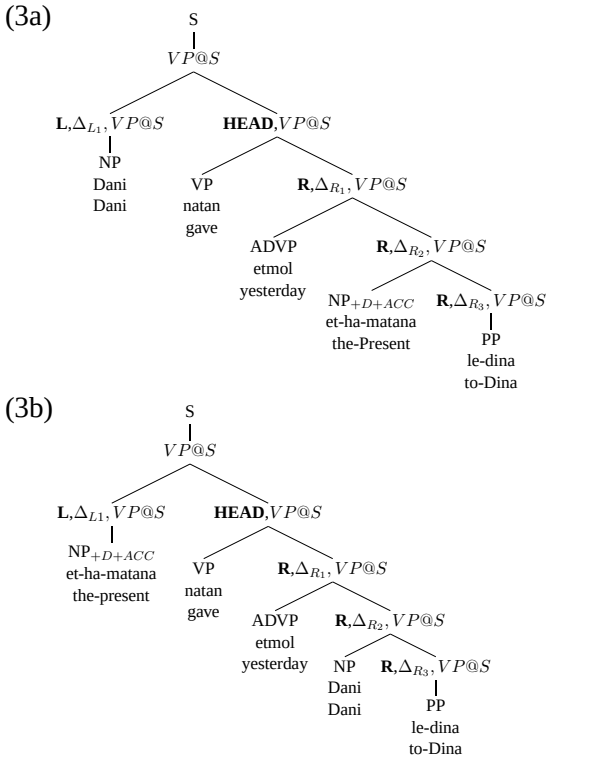


Figure 2: The *Head-Driven* Approach for Ex. (3)

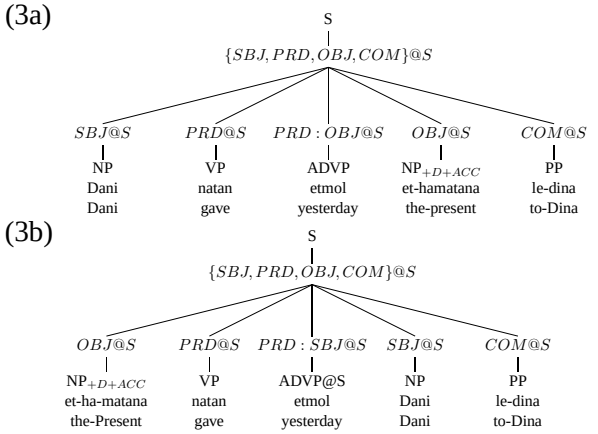


Figure 3: The *Relational-Realizational* Approach

RN provides an abstract set-theoretic representation of the *argument structure* of the clause.[5] This is called the *projection* phase. Then an ordering of the grammatical relations is generated, including reserved contextual slots for adjunction and/or punctuation marks. This is called the *configuration* phase. Finally, each of the grammatical function labels and adjunction slots gets realized as a morphosyntactic representation (a category label plus dominated morphological features) of the respective daughter constituent. This is called the *realization* phase.[6]

Figure 3 shows the generation of sentences (3a)–(3b) following the *projection, configuration* and *realization* phases corresponding to the top-down context-free layers of the tree. In both cases, the same relational network is generated, capturing the fact that they have the same argument structure. Then the different orderings of the grammatical elements are generated, reserving an adjunction slot for sentential modification (labeled by short context). Interestingly, the HD/RR models for our sentences are of comparable size (seven parameters) but the parameter types encode radically different notions. Illustrative of the difference is the realization of a morphologically marked NP object. In the RR model this is conditioned on a grammatical relation (check, for instance, node OBJ@S) and in the HD model it is conditioned on linear ordering or configurational notions such as *left, right* and *distance*.

## 4 Experiments

**Goal** We set out to compare the performance of the different modeling approaches for parsing Modern Hebrew. Considerable effort was devoted to making the models strictly comparable, in terms of preparing the data, defining statistical events, and unifying the rules determining cross-cutting linguistic notions (e.g., *heads* and *predicates, grammatical functions* and *subcat sets*). We spell out some of the setup considerations below.

**Data** We use the Modern Hebrew treebank (MHTB) (Sima'an et al., 2001) consisting of 6501 sentences from news-wire texts, morphologically analyzed and syntactically annotated as phrase-

[5] Unlike in HD models or dependency grammars, the *head* predicative element has no distinguished status here.

[6] Realization of adjunction slots (but not of function labels) may generate multiple sisters adjoining at a single position.

| *GF* | *Description* | *Applicable to…* |
|---|---|---|
| PRD | Predicative Elements | VP, PREDP |
| SBJ | Grammatical Subjects | NP, SBAR |
| OBJ | Direct Objects | NP |
| COM | Indirect Objects | NP, PP |
| | FInite Complements | SBAR |
| IC | Infinitival Complements | VP |
| CNJ | A Conjunct within | |
| | a Conjunction Structure | All |

Table 2: Grammatical Functions in the MHTB

| | | |
|---|---|---|
| **SP-PCFG** | Expansion | $P(C_{l_n}, \ldots, C_h, \ldots, C_{r_n}\|P)$ |
| **HD-PCFG** | Head | $P(C_h\|P)$ |
| | Left Branch? | $P(\mathbf{L}{:}\Delta_{l_1}, \mathbf{H}{:}\Delta_h\|C_h, P)$ |
| | Right Branch? | $P(C_h, \mathbf{R}{:}\Delta_{r_1}\|\Delta_h, C_h, P)$ |
| | Left Arg/Mod | $P(C_{l_i}, \Delta_{l_{i+1}}\|\ \mathbf{L}\ , \Delta_{l_i}, C_h, P)$ |
| | Right Arg/Mod | $P(C_{r_i}, \Delta_{r_{i+1}}\|\ \mathbf{R}\ , \Delta_{r_i}, C_h, P)$ |
| | Left Final? | $P(C_1\|\ \mathbf{L}\ , \Delta_{l_{n-1}}, C_h, P)$ |
| | Right Final? | $P(C_n\|\ \mathbf{R}\ , \Delta_{r_{n-1}}, C_h, P)$ |
| **RR-PCFG** | Projection | $P(\{gr_1, \ldots, gr_m\}\|P)$ |
| | Configuration | $P(\langle gr_1, \ldots, gr_m\rangle\|\{gr_1, \ldots, gr_m\}P)$ |
| | Realization | $P(C_j\|gr_j, P)$ |
| | Adjunction | $P(C_{j_1}, \ldots, C_{j_n}\|gr_j : gr_{j+1}, P)$ |

Table 3: PCFG Parameter Classes for All Models

structure trees. In our version of the MHTB, *definiteness* and *accusativity* features are percolated from the PoS-tags level to phrase-level categories, extending the procedure of (Guthmann et al., 2009). For all models, we applied non-terminal state-splits distinguishing finite from non-finite verb forms and possessive from non-possessive noun phrases. We head-annotated the treebank, and based on the 'subject', 'object', 'complement' and 'conjunction' labels in the MHTB we devised an automatic procedure to annotate all the grammatical functions indicated in table 2.[7]

**Procedure** For all models, we learn a PCFG by reading off the parameters described in table 3, in accordance with the trees depicted in figures 1–3.[8] For all models, we use relative frequency estimates. For lexical parameters, we use a simple smoothing procedure assigning probability to unknown words using the per-tag distribution of rare words ("rare" threshold set to $< 2$). The input to our parser consists of morphologically segmented surface forms, and the parser has to assign the syntactic as well as morphological analysis to the surface segments.[9] We use the standard development/training/test split as in (Tsarfaty and Sima'an, 2008). Since our goal is a detailed comparison and fine-grained analysis of the results we concentrate on the development set. We use a general-purpose CKY parser (Schmid, 2004) to exhaustively parse the sentences, and we strip off all model-specific information prior to evaluation.

**Evaluation** We use standard *Parseval* measures calculated for the original, flat, canonical representation of the parse trees.[10] We report *Precision/Recall* for the coarse-grained non-terminal categories. In addition to overall Parseval scores we report the accuracy results *Per Syntactic Category*. We further report model size in terms of the number of parameters. As is well known in Machine Learning, models with more parameters require more data to learn, and are more vulnerable to sparseness. In our evaluation we thus follow the rule of thumb that (all else being equal) for models of equal size the better performing model is preferred, and for models with equal performance, the smaller one is preferred.

## 5 Results and Analysis

### 5.1 Overall Results

Table 4 shows the parsing results for the **State-Split (SP) PCFG**, the **Head-Driven (HD) PCFG** and the **Relational-Realizational (RR) PCFG** models on parsing the Modern Hebrew Treebank, with *definiteness* and *accusativity* marked on PoS-tags as well as phrase-level categories. For all models, we experiment with grandparent encoding. For non-HD models, we also examine the utility of a head-category split.[11]

[7] The enhanced corpus will be available at `www.science.uva.nl/~rtsarfat/resources.htm`.

[8] Our training procedure is strictly equivalent to the transform-detransform methodology of (Johnson, 1998), but we implement a tree-traverse procedure as in (Bikel, 2002) collecting all parameters per event at once.

[9] This setup is more difficult than, e.g., the Arabic parsing setup of (Bikel, 2002), as they assume gold-standard pos-tags as input. Yet it is easier than the setup of (Tsarfaty, 2006; Goldberg and Tsarfaty, 2008) which uses unsegmented surface forms as input. The decision to use segmented and untagged forms was made to retain a realistic scenario. Morphological analysis is known to be ambiguous, and we do not assume that morphological features are known up front. Morphological segmentation is also ambiguous, but for our purposes it is unavoidable. When comparing different models on an individual sentence they may propose segmentation to sequences of different lengths, for which accuracy results cannot be faithfully compared. See (Tsarfaty, 2006) for discussion.

[10] The flat canonical representation also allows for a fair comparison that is not biased by the differing branching factors of the different models.

[11] In HD models, a head-tag is already assumed in the conditioning context for sister nodes (Klein and Manning, 2003).

| | | | | |
|---|---|---|---|---|
| **SP-PCFG** | | | | |
| Grand-Parent | − | − | + | + |
| Head-Tag | − | + | − | + |
| Prec/Rec | 70.05/72.40 | 71.14/72.03 | **74.66/74.35** | 71.99/72.17 |
| (#Params) | (4995) | (8366) | **(7385)** | (11633) |
| **HD-PCFG** | | | | |
| Grand-Parent | − | − | + | + |
| Markov | 0 | 1 | 0 | 1 |
| Prec/Rec | 66.87/71.64 | 70.40/74.35 | 73.04/71.94 | **73.52/74.84** |
| (#Params) | (6678) | (10015) | (19066) | **(21399)** |
| **RR-PCFG** | | | | |
| Grand-Parent | − | − | + | + |
| Head Tag | − | + | − | + |
| Prec/Rec | 69.90/73.96 | 72.96/75.73 | 74.19/75.03 | **76.32/76.51** |
| (#Params) | (3791) | (7546) | (7611) | **(13618)** |

Table 4: **The Performance of Different Models in Parsing Hebrew:** Parsing Results Prec/Recall for Sentences of Length $\leq 40$.

For all models, grandparent encoding is helpful. For HD models, a higher Markovian order improves performance. This shows that even in Hebrew there are linear-precedence tendencies that help steer the disambiguation in the right direction, which is in line with our observation that word-order patterns in Modern Hebrew are not completely free (cf. table 1).

The best SP model performs equally or better than all HD models. This might be due to the smaller size of SP grammars, resulting in more robust estimates. But it is remarkable that, given the feature-rich representation, such a simple treebank grammar provides better disambiguation capacity than linguistically articulated HD models. We attribute this to the fact that parent-daughter relations have a stronger association with grammatical functions than relations between neighbouring nodes. For Hebrew, such adjacency relations may be arbitrary due to word-order variability.

Overall, RR models show the best performance for the set of all models with parent encoding, and for the set of all models without. Our best RR model shows 6.6%/8.4% Prec/Rec error reduction from the best SP model. The Recall improvement shows that the RR model is much better in generalizing, recovering successfully more of the constituents found in the gold representation. The best RR model also outperforms HD models with 8.7%/6.7% Prec/Rec error reduction from the best HD model. The resulting precision improvement of the RR relative to HD is larger than the improvement relative to SP, and the Recall improvement pattern is reversed. So it seems that the HD model generalizes better than the SP model, but also gets generalizations wrong more often than the SP model.

The RR model combines the generalization advantage of breaking down context-free events while it maintains the coherence advantage of learning flat trees (cf. (Johnson, 1998)). The best RR model obtains the best performance among all models: $F_1 76.41$. To put this result in context, for the setting in which the Arabic parser of (Maamouri et al., 2008) obtains $F_1 78.1$, — i.e., with gold standard feature-rich tags — the best RR model obtains $F_1 83.3$ accuracy which is the best parsing result reported for a Semitic language so far. RR models also have the advantage of resulting in more compact grammars, which makes learning and parsing with them much more computationally efficient.

In our SP or RR models, head-information is used as yet another feature-value pair rather than an object with a distinguished status during generation.

| Model / Category | SP-PCFG | HD-PCFG | RR-PCFG |
|---|---|---|---|
| NP | 77.39 / 74.32 | 77.94 / 73.75 | **78.96 / 76.11** |
| PP | 71.78 / 71.14 | 71.83 / 69.24 | **74.4 / 72.02** |
| SBAR | 55.73 / 59.71 | 53.79 / 57.49 | **57.97 / 61.67** |
| ADVP | 71.37 / 77.01 | 72.52 / 73.56 | **73.57 / 77.59** |
| ADJP | **79.37 / 78.96** | 78.47 / 77.14 | 78.69 / 78.18 |
| S | **73.25 / 79.07** | 71.07 / 76.49 | 72.37 / 78.33 |
| SQ | 36.00 / **32.14** | 30.77 / 14.29 | **55.56** / 17.86 |
| PREDP | 36.31 / 39.63 | **44.74** / 39.63 | 44.51 / **46.95** |
| VP | 76.34 / 80.81 | 77.33 / **82.51** | **78.59** / 81.18 |

Table 5: **Per-Category Evaluation of Parsing Performance for Different Models:** Prec/Rec Per Category Calculated for All Sentences.

## 5.2 Per-Category Break-Down Analysis

To understand better the merits of the different models we conducted a break-down analysis of performance-per-category for the best performing models of each kind. The break-down results are shown in table 5. We divided the table into three sets of categories: those for which the RR model gave the best performance, those for which the SP model gave the best performance, and those for which there is no clear trend.

The most striking outcome is that the RR model identifies at higher accuracy precisely those syntactic elements that are freely positioned with re-

spect to the head: NPs, PPs, ADVPs and SBARs. Adjectives, in contrast, have clear ordering constraints — they always appear after the noun. S level elements, when embedded, always appear immediately after a conjunction or a relativizer. In particular, NPs and PPs realize arguments and adjuncts that may occupy different positions relative to the head. The RR model is better than the other models in identifying those elements partly because morphological information helps to disambiguate syntactically relevant chunks and make correct attachment decisions about them.

Remarkably, predicative (verb-less) phrases (PREDP), which are characteristic of Semitic languages, are hard to parse, but here too the RR does slightly better than the other two, as it allows for variability in the means to realize a (verbal or verb-less) predicate. Both RR and HD models outperform SP for VPs, which is due to the specific nature of VPs in the MHTB – they exist *only* for complement phrases with strict linear ordering.

## 6 Distances, Functions and Subcategorization Frames

Markovian processes to the *left* and to the *right* of the head provide a first approximation of the predicate's *argument structure*, as they capture trends in the co-occurrences of constituents reflected in their pattern of *positioning* and *adjacency*. But as our results so far show, such an approximation is empirically less rewarding for a language in which grammatical relations are not tightly correlated with structural notions.[12]

Collins (2003) attempted a more abstract formulation of argument-structure by articulating left and right *subcat-sets*. Each set represents those arguments that are expected to occur at each side of the head. Argument sisters ("complements") are generated if and only if they are required, and their generation 'cancels' the requirement in the set. Adjuncts ("modifiers") may be freely generated at any position.

At first glance, such a dissociation of configurational positions and subcategorization sets seems to be more adequate for parsing Hebrew, because it allows for some variability in the order of generation. But here too, since the model uses sets of

[12]Conditioning based on *adjacency* and *distance* is also common inside *dependency parsing* models, and we conjecture that this is one of the reasons for their difficulty in coping with freer word-order languages, a difficulty pointed out in (Nivre et al., 2007).

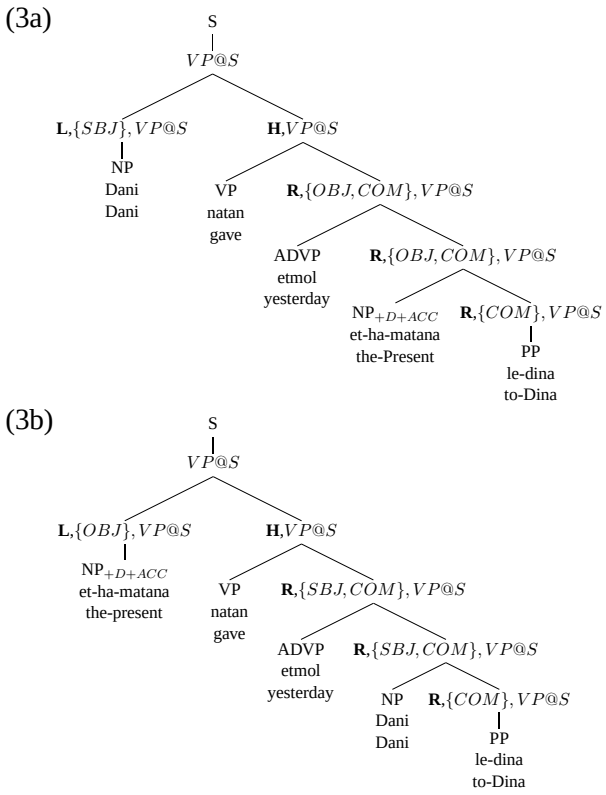


Figure 4: The *Relational Head-Driven* Approach

*constituent labels*, it disambiguates the grammatical functions of an NP solely based on the direction of the head, which is adequate for English but not for Hebrew. In order to relax this association further, we propose to replace constituent labels in the subcat-sets with grammatical relations identical to the functional elements in the relational network of the RR. This provides means to mediate the cancellation of constituents in the sets with their functions and correlate it with morphology.

To get an idea of the implications of such a modeling strategy, let us consider our example sentences in such a Relational-HD model as depicted in figure 4. Both representations share the event of generating the verbal head. Sisters are generated conditioned on the head and the functional elements remaining to be "cancelled". Each of the two trees consists of an event realizing an "object", one for an NP to the right of the head, and the other for an NP to its left. In both cases, an object constituent will be generated jointly with the morphological features associated with it. Evidently, when using sets of grammatical relations instead of constituent-labels, correlation of morphology and grammatical functions is more straight-forward to maintain.

| *Model* | **SP-PCFG** | **HD-PCFG** | **HD-PCFG** | **HD-PCFG** | **HD-PCFG** | **RR-PCFG** |
|---|---|---|---|---|---|---|
| *Type of Distance* $\Delta$ *or Subcategorization* | Phrase-Level State-Splits | Intervening Verb/Punc | Left and Right #Constituents | Left and Right Constituent Labels | Left and Right Function Labels | Subcat Sets Configuration |
| *Precision/Recall (#Params)* | 70.95/70.32 (13884) | 72.39 / 71.97 (11650) | 72.70 / 74.46 (18058) | 72.42 / 74.29 (16334) | 72.84/74.62 (16460) | 76.32/76.51 (13618) |

Table 6: **Incorporating Distance and Grammatical Functions into Head-Driven Parsing Models** Reporting Precison/Recall (#Parameters) for Sentences Length $< 40$.

### 6.1 Results and Analysis

Table 6 reports the results of experimenting with HD models with different instantiations of a *distance* function, starting from the standard notion of (Collins, 2003) and ending with our proposed, relational, function sets. For all HD models, we retain the *head*, *left* and *right* generation cycle and only change the conditioning context ($\Delta_i$) for sister generation.

As a baseline, we show the results of adding grammatical function information as state-splits on top of an SP-PCFG.[13] This SP model presents much lower performance than the RR model although they are almost of the same size and they start off with the same information. This result shows that sophisticated modeling can blunt the claws of the sparseness problem. One may obtain the same number of parameters for two different models, but correlate them with more profound linguistic notions in one model than in the other. In our case, there is more statistical evidence in the data for, e.g., case marking patterns, than for association of grammatical relations with structurally-marked positions.

For all HD variations, the RR model continues to outperform HD models. The function-set variation performs slightly (but not significantly) better than the category-set. What seems to be still standing in the way of getting useful disambiguation cues for HD models is the fact that the *left* and *right* direction of realization is hardwired in their representation. This breaks down a coherent distribution over morphosyntactic representations realizing grammatical relations to arbitrary position-dependent fragments, which results in larger grammars and inferior performance.[14]

[13]The startegy of adding grammatical functions as state-splits is used in, e.g., German (Rafferty and Manning, 2008).

[14]Due to the difference in the size of the grammars, one could argue that smoothing will bridge the gap between the HD and RR modeling strategies. However, the better size/accuracy trade-off shown here for RR models suggests that they provide a good bias/variance balancing point, especially for feature-rich models characterizing morphologically rich languages. A promising strategy then would be to smooth or split-and-merge (Petrov et al., 2006)) RR-based models rather than to add an elaborate smoothing component to configurationally-based HD models.

## 7 A Typological Detour

Hebrew, Arabic and other Semitic Languages are known to be substantially different from English in that English is strongly *configurational*. In configurational languages word-order is fixed, and information about the grammatical functions of constituents (e.g., *subject* or *object*) is often correlated with structurally-marked positions inside highly-nested constituency structures. *Nonconfigurational* languages (Hale, 1983), in contrast, allow for freedom in their word-ordering and information about grammatical relations between constituents is often marked by means of *morphology*.

*Configurationality* is hardly a clear-cut notion. The difference in the configurationality level of different languages is often conceived as depicted in figure 7. In *linguistic typology*, the branch of linguistics that studies the differences between languages (Song, 2001), the division of labor between linear ordering and morphological marking in the realization of grammatical relations is often viewed as a continuum. Common wisdom has it that the lower a language is on the configurationality scale, the more morphological marking we expect to be used (Bresnan, 2001, page 6).

For a statistical parser to cope with nonconfigurational phenomena as observed in, for instance, Hebrew or German, it should allow for flexibility in the *form* of realization of the grammatical *functions* within the phrase-structure representation of trees. Recent morphological theories employ *Form-Function* separation as a widely-accepted practice for enhancing the adequacy of models describing variability in the realization of grammatical *properties*. Our results suggest that the adequacy of syntactic processing models is related to such typological insights as well, and is enhanced by adopting a similar form-function separation for expressing grammatical *relations*.

*configurational* ——————— *nonconfigurational*
Chinese>English>{German,Hebrew}>Warlpiri

Figure 5: **The Configurationality Scale**

The HD assumptions take the function of a constituent to be transparently related to its formal position, which entails word-order rigidity. Such transparent relations between configurational positions and grammatical functions are assumed by other kinds of parsing frameworks such as the 'all-subtrees' approach of Data-Oriented Parsing, and the distinction between left and right application in CCG-based parsers.

The RR modeling strategy stipulates a strict separation between *form* — parametrizing explicitly basic word-order (Greenberg, 1963) and morphological realization (Greenberg, 1954) — and function — parametrizing relational networks borrowed from (Perlmutter, 1982) — which makes it possible to statistically learn complex form-function mapping reflected in the data. This is an adequate means to capture, e.g., morphosyntactic interactions, which characterize the *less-configurational* languages on the scale.

## 8 Conclusion

In our comparison of the HD and RR modeling approaches, the RR approach is shown to be empirically superior and typologically more adequate for parsing a language exhibiting word-order variation interleaved with extended morphology. HD models are less accurate and more vulnerable to sparseness as they assume transparent mappings between form and function, based on *left* and *right* decompositions hard-wired in the HD representation. RR models, in contrast, employ *form* and *function* separation which allows the statistical model to learn complex correspondance patterns reflected in the data. In the future we plan to investigate how the different models fare against one another in parsing different languages. In particular we wish to examine whether parsing different languages should be pursued by different models, or whether the RR strategy can effectively cope with different languages types. Finally, we wish to explore the implications of RR modeling for applications that consider the form of expression in multiple languages, for instance *Statistical Machine Translation (SMT)*.

## 9 Acknowledgements

We thank Jelle Zuidema, Inbal Tsarfati, David McCloskey and Yoav Golderg for excellent comments on earlier versions. We also thank Miles Osborne and Tikitu de Jager for comments on the camera-ready draft. All errors are our own. The work of the first author is funded by the Dutch Science Foundation (NWO) grant 017.001.271.

## References

J. Aissen. 2003. Differential Object Marking: Iconicity vs. Economy. *Natural Language and Linguistic Theory*, 21.

D. M. Bikel. 2002. Design of a Multi-lingual, Parallel-processing Statistical Parsing Engine. In *Proceedings of HLT*.

J. Bresnan. 2001. *Lexical-Functional Syntax*. Blackwell Textbooks in Linguistics. Blackwell.

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

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

M. Collins, J. Hajič, E. Brill, L. Ramshaw, and C. Tillmann. 1999. A Statistical Parser of Czech. In *Proceedings ACL*.

M. Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

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

A. Dubey. 2004. *Statistical Parsing for German: Modeling syntactic properties and annotation differences*. Ph.D. thesis, Saarland University, Germany.

Y. Goldberg and R. Tsarfaty. 2008. A Single Framework for Joint Morphological Segmentation and Syntactic Parsing. In *Proceedings of ACL*.

J.H. Greenberg. 1954. A Quantitative Approach to the Morphological Typology of Language. In R. F. Spencer, editor, *Method and Perspective in Anthropology*. University of Minessota Press.

J. H. Greenberg. 1963. Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. In Joseph H. Greenberg, editor, *Universals of Language*. MIT Press.

N. Guthmann, Y. Krymolowski, A. Milea, and Y. Winter. 2009. Automatic Annotation of Morpho-Syntactic Dependencies in a Modern Hebrew Treebank. In *Proceedings of TLT*.

K. L. Hale. 1983. Warlpiri and the Grammar of Non-Configurational Languages. *Natural Language and Linguistic Theory*, 1(1).

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

D. Klein and C. Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of ACL*.

S. Kubler. 2008. The PaGe Shared task on Parsing German. In *ACL Workshop on Parsing German*.

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

M. Maamouri, A. Bies, and S. Kulick. 2008. Enhanced Annotation and Parsing of the Arabic treebank. In *Proceedings of INFOS*.

D. M. Magerman. 1995. Statistical Decision-Tree Models for Parsing. In *Proceedings of ACL*.

M. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*.

P. H. Matthews. 1993. *Morphology*. Cambridge.

D. McClosky, E. Charniak, and M. Johnson. 2008. When is self-training effective for parsing? In *Proceedings of CoLing*.

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

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The Shared Task on Dependency Parsing. In *Proceedings of the CoNLL Shared Task*.

D. M. Perlmutter. 1982. Syntactic Representation, Syntactic Levels, and the Notion of a Subject. In Pauline Jacobson and Geoffrey Pullum, editors, *The Nature of Syntactic Representation*. Springer.

S. Petrov, L. Barrett, R. Thibaux, and D. Klein. 2006. Learning Accurate, Compact, and Interpretable Tree Annotation. In *Proceedings of ACL*.

A. Rafferty and C. D. Manning. 2008. Parsing Three German Treebanks: Lexicalized and Unlexicalized Baselines. In *ACL WorkShop on Parsing German*.

R. Scha. 1990. Language Theory and Language Technology; Competence and Performance. In Q. A. M. de Kort and G. L. J. Leerdam, editors, *Computertoepassingen in de Neerlandistiek*. Almere: LVVN.

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

U. Shlonsky. 1997. *Clause Structure and Word Order in Hebrew and Arabic*. Oxford University Press.

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

J. J. Song. 2001. *Linguistic Typology: Morphology and Syntax*. Pearson Education Limited, Edinbrugh.

R. Tsarfaty and K. Sima'an. 2008. Relational-Realizational Parsing. In *Proceedings of CoLing*.

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