Towards a linguistically motivated computational grammar for Hebrew

Shuly Wintner
Seminar für Sprachwissenschaft, Universität Tübingen
Kl. Wilhelmstr. 113, 72074 Tübingen, Germany
shuly@sfs.nphil.uni-tuebingen.de

Abstract

While the morphology of Modern Hebrew is well accounted for computationally, there are few computational grammars describing the syntax of the language. Existing grammars are scarcely based on solid linguistic grounds: they do not conform to any particular linguistic theory and do not provide a linguistically plausible analysis for the data they cover. This paper presents a first attempt towards the construction of a formal grammar for a fragment of Hebrew that is both linguistically motivated and computationally implementable. The grammar, concentrating on the structure of noun phrases, is designed in accordance with HPSG, a linguistic theory that lends itself most naturally to computational implementation. It is the first application of HPSG to any Semitic language. Several theoretical issues are addressed, including the status of the definite article, the application of the DP hypothesis to Hebrew, definiteness agreement in the noun phrase as well as definiteness inheritance in constructs. All the analyses presented in the paper were tested and their predictions were verified. This is a work in progress, and the results described herein are preliminary.

1 Introduction

Modern Hebrew (MH) poses some interesting problems for the grammar designer. The Hebrew script is highly ambiguous, a fact that results in many part-of-speech tags for almost every word (Ornan, 1994). Short prepositions, articles and conjunctions are usually attached to the words that immediately succeed them. In addition, Hebrew morphology is very rich: a noun base might have over fifteen different derivations, and a verb – over thirty. In spite of the difficulties, disambiguation of the script, as well as morphological analysis, were covered by a variety of works (Bentur et al., 1992; Choueka and Ne'eman, 1995; Ornan and Katz, 1995). From a practical point of view, Hebrew morphology is well accounted for.

The syntax of the language, however, remains an open problem. The first syntactic analyzer for Hebrew is described in (Cohen, 1984), but its grammar is implicit in a software system. Nirenburg and Ben-Asher (1984) describe a small-scale ATN for Hebrew, capable of recognizing very limited structures. Unification-based formalisms were used for developing Hebrew grammars only recently. A limited experiment using PATR-II is described in (Wintner, 1992); it is extended (Wintner and Ornan, 1996) to a reasonable subset of the language, on a different platform: Tomita's LR Parser/Compiler, which is based on LFG. The grammar recognizes sentences of wide variety and complexity, but the analyses it provides are not conveyed in the framework of any particular linguistic theory. A different work along the same lines is (Yizhar, 1993): using the same framework, it concentrates on the syntax of noun phrases, employing ideas from different linguistic theories.

Works related to the syntax of Hebrew, and in particular to noun phrases, are abundant in the theoretical linguistics literature (Borer, 1984; Ritter, 1991; Siloni, 1994). All of them are carried out in Chomskian frameworks; none can be directly implemented computationally, and their predictions cannot be verified on the basis of existing on-line corpora. The practical contribution of these works is thus limited.

This paper describes the first stages of an attempt to bridge the gap between linguistically theoretic analyses and computational implementations. Using HPSG (Pollard and Sag, 1994) as the linguistic theory in which analyses are conveyed, grammars can be directly implemented and their predictions verified. HPSG is used for formally describing the structure of a variety of languages, but this is the first time the theory is applied to any Semitic language. While some ideas of existing

Hebrew grammars, in particular (Wintner and Ornan, 1996) and (Yizhar, 1993), are incorporated into the work described here, the starting point is new: we present an account of several aspects of the Hebrew noun phrase, aligned with the general principles of HPSG. All the analyses described in the paper were computationally implemented using AMALIA (Wintner, 1997a) as the development framework. The phenomena we address include the status of the definite article, the application of the DP hypothesis to Hebrew, definiteness agreement in noun phrases as well as definiteness inheritance in constructs. This is a work in progress, and the results described here are preliminary. The grammar is not intended to have a broad coverage, but rather to provide explanatory structures to linguistically interesting phenomena. However, we hope to extend the coverage of the grammar in the future, maintaining its linguistic rigor.

2 The framework

HPSG is formulated as a set of constraints on *typed feature structures* (TFSs) that are used to model linguistic information in all levels: from the lexicon, through grammatical principles, to complete analyses. HPSG "rules" are organized as *principles* that set constraints on the properties of well-formed phrases, along with *ID schemata* that license certain phrase structures. The schemata are independent of the categories of the involved phrases; they state general conditions for the construction of larger phrases out of smaller ones, according to the function of the sub-phrases (e.g., *subject-head, head-complement, specifier-head* etc.) ID schemata only *license* certain phrase combinations. They do not specify *all* the constraints imposed on the involved sub-phrases, as these are articulated by the principles.

Like other current linguistic theories, HPSG is highly *lexical*: most of the information is encoded in highly articulated lexical entries associated with words. The constraints on the grammar are usually few and very general. An elaborate set of *lexical rules* relates lexical entries, either to account for morphology or to introduce changes in the TFSs associated with the basic entries.

3 The structure of noun phrases

3.1 The data

Hebrew has one definite article, ha-, which attaches to *words* (nouns, adjectives, numerals and demonstratives, henceforth *nominals*), not phrases. Many elements in the noun phrase are marked for, and must agree on, definiteness (1). MH provides two major ways of forming genitive relations: *free genitives* (FG), in which the genitive phrase is introduced by the preposition $ell 'of' (2); and *constructs* (CS), in which the head noun is morphologically marked (and is said to be in the *construct* state, *cs*) and the genitive phrase must immediately follow it, preceding any other modifiers (3). In FG the definiteness of the possessor is independent of that of the head, allowing for *four* different combinations of definiteness (both the head and the possessor can each be either definite or indefinite) (2); in CS, the definiteness of the phrase is inherited from the possessor, allowing only *two* combinations: either both are definite, or both are not (3). The definite article never combines with *cs*-nouns. A poorly studied yet closely related phenomenon is *cs*-adjectives, which exhibit the same definiteness behavior (4).

(1) ha- sepr ha- gadol ha- ze/lii
the book the big the this/third
'this big book / the third big book'

(2) (ha-) sparim $ell m$orer
(the) books of poet
'(the) books of a poet'

(ha-) sparim $ell ha- m$orer
(the) books of the poet
'(the) books of the poet'

(3) siprei m$orer xda$im
books-*cs* poet new
'new books of a poet'

siprei ha- m$orer ha- xda$im
books-*cs* the poet the new
'the new books of the poet'

(4) yruqqat (ha-) &einaym
green-*cs* (the) eyes
'a/(the) green eyed'

3.2 Are noun phrases NPs or DPs?

Following Abney (1987), analyses carried out in Chomskian frameworks view noun phrases as DPs, headed by the functional category D. The DP hypothesis (DPH) has been applied to a variety of languages and is incorporated into most existing accounts for Modern Hebrew. Originally motivated by the English '-ing' gerunds, that possess simultaneously properties of both sentences and noun phrases, the importance of the DPH is that it assigns parallel structures to clauses and noun phrases; in particular, both are headed by functional categories. In HPSG, however, functional categories are discouraged: English noun phrases are viewed as NPs, headed by the noun, and determiners – as subcategorized specifiers of nouns (Pollard and Sag, 1994, section 9.4). HPSG analyses for other languages, notably German, consider article-noun combinations to be DPs (Netter, 1994). Preferring either of the two analyses, in the context of HPSG, boils down to deciding whether it is the determiner or the noun that heads a nominal phrase. Applying the criteria of (Zwicky, 1985) we show that in Hebrew it is the noun that heads the noun phrases. Netter (1994) lists several considerations in favor of each of the alternatives. In German, all the morphosyntactic features that must be transferred to the maximal projection of a nominal phrase (for agreement or government purposes) are manifested equally well both on the article and on the noun. Determinerless noun phrases require, in German, disjunctive subcategorization frames for nouns under an NP analyses, and empty categories in a DP analysis. Finally, it is the declension phenomenon that causes Netter (1994) to favor a DP analysis. When applied to MH, these considerations yield a different result: information that is relevant for agreement, such as number and gender, is expressed on the noun only; determinerless phrases are always grammatical; and there are no declensions.

Nevertheless, most existing analyses of MH noun phrases apply the DPH, with the definite article as the D head (Ritter, 1988; Ritter, 1991; Siloni, 1991; Siloni, 1994). For lack of space we cannot survey the motivation for such analyses here; the argumentation relies on derived (deverbal) nouns, especially in CS noun phrases, including the following observations: the inability of *cs*-nouns to be rendered definite directly (i.e., the fact that ha- never attaches to them); the impossibility of direct modification of such nouns (i.e., the fact the any adjectives must follow the genitive complement in CS); and the inheritance of definiteness from the complement in CS. These, along with theory-internal considerations, yield an analysis by which noun phrases are DPs, headed by the functional, possibly phonologically null, category *D*, and necessitating a compulsory movement of the head noun. FG noun phrases require yet another functional (and empty) category. We show in (Wintner, 1998) that there is no theory-independent reason to apply the DPH to Hebrew; on the contrary, such accounts miss generalizations and yield wrong predictions. We show below that an NP analysis is not only possible but also plausible, accounting for a body of data, traditionally believed to require functional categories and compulsory head raising in noun phrases.

Many of the limitations of the analyses mentioned above are listed by Borer (1994), suggesting that definiteness is a feature of nouns, base generated on the N stem. An affixal view of the MH definite article is established in (Wintner, 1997b), and is the starting point for the analysis we propose here. We first account for the fact that *cs*-nouns must have an immediate complement. We then explain why the article does not combine with *cs*-nominals. We justify a treatment of possessives as complements, and finally present an analysis for both FG and CS noun phrases as NPs.

3.3 Prosodic dependency

Most subcategorized complements are optional in Hebrew: objectless VPs are grammatical in many contexts, as are subjectless clauses. But compulsory, immediate complementation is not unique to *cs*-nouns only; it is required in *cs*-adjectives and cardinals, as well as in prepositions and some quantifiers. In spite of the differences among these elements, there are some striking similarities: they can never occur without a complement, which cannot be extracted, or 'moved', but which can be replaced by a pronominal pronoun, which is always realized as a clitic (Borer, 1984, chapter 2). The data are summarized in (5).

(5) siprei ha- m$or_rim / sipreihem
books-*cs* the poets / books+3rd-pl-m
'the poets' books / their books'

lot ha- m$or_rim / lotam
three-*cs* the poets / three+3rd-pl-m
'the three poets / the three of them'

$ell ha- m$or_rim / $ellahem
of the poets / of+3rd-pl-m
'of the poets / of them'

'et ha- m$or_rim / 'otam
ACC the poets / ACC+3rd-pl-m
'the poets (ACC) / them (ACC)'

&al yad ha- m$or_rim / &al yadam
near the poets / near+3rd-pl-m
'near the poets / near them'

koll ha- m$or_rim / kullam
all the poets / all+3rd-pl-m
'all the poets / all of them'

The need for an immediate complement is a result of these elements being prosodically weak. We do not suggest a theory of prosody in HPSG; rather, taking advantage of the observation that the discussed constituents correlate well with phrases in MH, we account for them in the following way: we add a DEPendency feature to the lexical entries of *words*. The value of this feature can either be an empty list, or a list of one element, in which case the element must be reentrant with some element in some valence list of the word (in other words, DEP points to some element on the ARG_S value of the word). As the only relations between prosodically dependent words and their obligatory complements, in Hebrew, are those of head-complement or specifier-head, the obligatory complement is bound to be a member of the ARG_S of those words. In addition, we introduce the *prosodic dependency* principle, by which words that are specified as prosodically dependent must first combine with the obligatory complement they depend on; only then can the obtained phrases combine with other modifiers:

> In a headed phrase, in which one of the daughters is a word, either the DEP of this daughter is empty, or it is reentrant with (the SYNSEM value of) some other daughter.

3.4 The morphological nature of definiteness

Why doesn't the definite article combine with *cs*-nouns? Not only nouns have construct states: adjectives (4) and numerals do, too, and ha- does not combine with the other *cs*-nominals either. The rules that govern the combination of ha- with nominals are simple, when the article is viewed as an affix (Wintner, 1997b): (i) ha- attaches to words, not to phrases; (ii) it attaches only to nominals, and to all kinds of nominals; (iii) it only combines with indefinite words. An additional (boolean) feature, DEFiniteness, is required for encoding the value of definiteness in nominals. As definiteness agreement in Hebrew is not a semantic process, we add this feature to the CATegory of nominals (rather than to their CONTent). Since definiteness is a feature of phrases, inherited from the lexical head, DEF is a head feature, appropriate for all *nominal*s. Viewing definiteness as a lexical process, we introduce the *Definite Lexical Rule* (DLR, 6). It operates on all nominal words whose DEFiniteness feature is '−'. In all categories its effect on the phonology is determined by the same phonological rules, abstracted over by the function *definite*. The DLR changes the value of the path SYNSEM|LOC|CAT|HEAD|DEF from '−' to '+'. *Adjuncts* specify the heads they select as the value of the MOD feature in their lexical entries. Like any other nominal, they have a DEFiniteness feature, whose value is shared with the value of the path MOD|LOC|CAT|HEAD|DEF. When the DLR operates on adjuncts, it results in a specification of a '+' value for both paths. Thus it is guaranteed that definite adjectives, for example, are not only specified as definite but also select definite heads. As for *cs*-nominals, these are not indefinite; we show below that they are unspecified for definiteness, and hence the DLR cannot apply to them.

(6)

$$\begin{bmatrix} \textbf{word} \\ phon: \quad \boxed{1} \\ synsem: cat: \begin{bmatrix} \textbf{cat} \\ head: \begin{bmatrix} \textbf{nominal} \\ def: - \end{bmatrix} \end{bmatrix} \end{bmatrix} \mapsto$$

$$\begin{bmatrix} \textbf{word} \\ phon: \quad definite(\boxed{1}) \\ synsem: cat: \begin{bmatrix} \textbf{cat} \\ head: \begin{bmatrix} \textbf{nominal} \\ def: + \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

3.5 Possessives as complements

In standard HPSG (Pollard and Sag, 1994, section 9.4.5) possessives are *specifiers*: they combine with an N's to form complete NPs through the specifier-head schema, and they express the expectation of an N' as the value of the SPECified feature in their HEADs, just like other determiners do. As Pollard and Sag (1994, p. 375) note, this analysis is valid for German and English, but other languages might require different accounts. We advocate a position by which possessives of all kinds are *complements* in MH. First, possessives differ from other determiners in their distribution. While most determiners precede the noun, possessives follow it (7). Second, possessives can regularly co-occur with other determiners (8). Thus, if determiners occupy the specifier position in NPs, possessives cannot fill the same function. Third, MH exhibits also cases of clitic doubled constructions (Borer, 1984), where a genitive pronoun cliticizes onto the head noun and must agree with a doubled possessive on number, gender and person. Agreement is usually associated with complements (including subjects) and not with specifiers.

(7) koll sepr
every book
'every book'

koll / lot ha- sparim
all / three the books
'all books / the three books'

ha- sparim $selli / $ell dan
the books my / of Dan
'my/Dan's book'

(8) koll sepr $selli / $ell dan
every book my / of Dan
'each of my/Dan's books'

koll ha- sparim $selli / $ell dan
all the books my / of Dan
'all my/Dan's books'

lot ha- sparim $selli / $ell dan
three the books my / of Dan
'my/Dan's three books'

Other arguments for viewing possessives as complements, in two languages that show many similarities to Hebrew, namely Welsh and Arabic, are given in (Borsley, 1995). We therefore view possessors as (most oblique) complements of nouns. When the noun has additional arguments, they are listed in its valence feature preceding the possessor. Thus, in the lexical entry of sepr ('book'), the value of the COMPlement list has two members, an agent and an optional[1] possessor. When two possessives are present, the structure depicted in (9) is obtained.

3.6 The structure of CS

As *cs*-nominals are words, their lexical entries express an expectation for an immediate complement; that is, an indication (the SYNSEM value) of the compulsory complement of *cs*-nominals is present in the lexical entry of the nominal. It is thus possible to share, in the lexicon, the values of the definiteness feature in both the nominal and its complement. This results in only two possibilities of definiteness combinations for constructs, as opposed to the four possible combinations of free genitives. The construct form is generated from the absolute form by means of a morphological process, modelled by a lexical rule (10). Apart from modifying the phonology[2] of the nominal, this process has a double effect. First, the rule picks a genitive complement from the COMP list, replaces it by a nominative noun phrase and unifies the values of the DEF feature of the nominal and the complement it depends on. In addition, the rule sets the value of 'DEP' to this complement, to indicate that *cs*-nominals are prosodically dependent. When the nominal is combined with its complement, the resulting phrase inherits the definiteness from the latter. Notice that the results of this process, i.e., the lexical entries of *cs*-nouns, are not specified as 'DEF −' (in fact, they are not specified for definiteness at all), and hence the DLR cannot apply to them. The fact that *cs*-nominals cannot be rendered definite directly is naturally obtained.

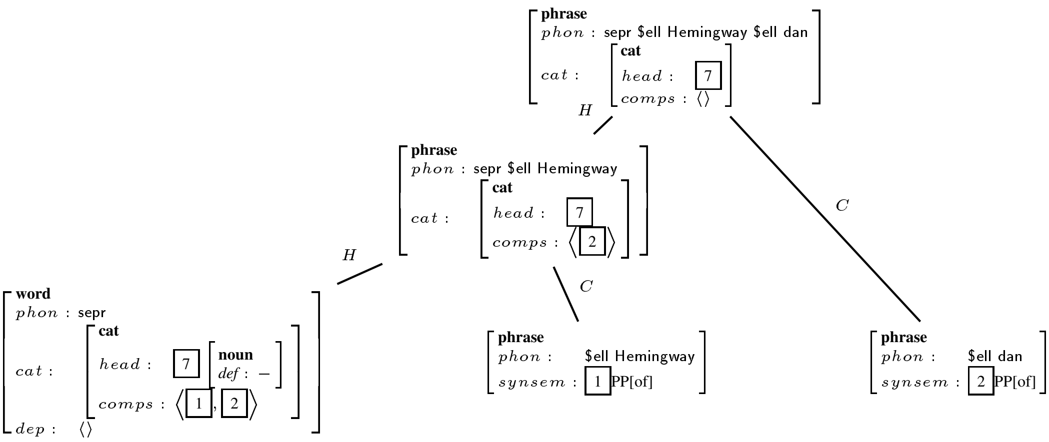
Noun–noun constructs are thus constructed by the head-complement schema. An independent *cs*-noun, with no immediate complement, cannot be promoted to the status of a phrase, as the dependency principle prohibits its combination with other phrases until its DEP requirements are discharged. Since the DEF value of the construct head and its complement are shared, and since DEF is a head feature, it is also shared by the mother; thus, the DEF feature of the phrase is inherited from the complement, as required. This process is depicted in (11); notice in particular how the definiteness of the phrase is inherited from the complement using a reentrancy in the head.

The similar properties noun–noun and adjective–noun constructs suggest that they are actually only two instances of one process: any analysis that would suggest two different mechanisms to account for both phenomena is bound to be redundant. We simply extend the analysis of noun–noun constructs to *cs*-adjectives: such adjectives are lexically specified to subcategorize for nouns. They cannot occur independently, with no immediate complement, and hence are marked as dependent; the phrase is constructed through the head-complement schema (12). We thus obtain a uniform, principled account for the

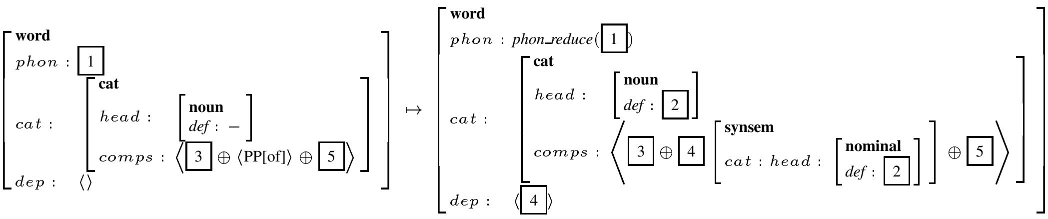
[1]Recall that most subcategorized elements are optional in Hebrew.

[2]The function *phon_reduce* computes the phonology of the construct noun.

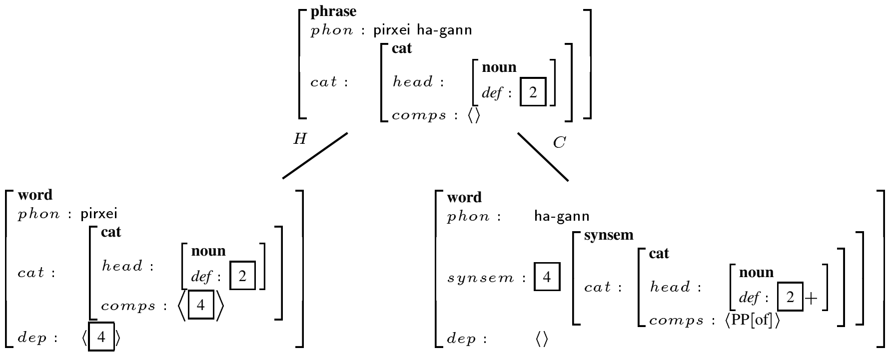
(9)



(10)



(11)



two phenomena, maintaining an NP view of noun phrases and requiring neither functional nor empty categories.

References

Steven Abney. 1987. *The English Noun Phrase in Its Sentential Aspect*. Ph.D. thesis, Massachusettes Institute of Technology, Cambridge.

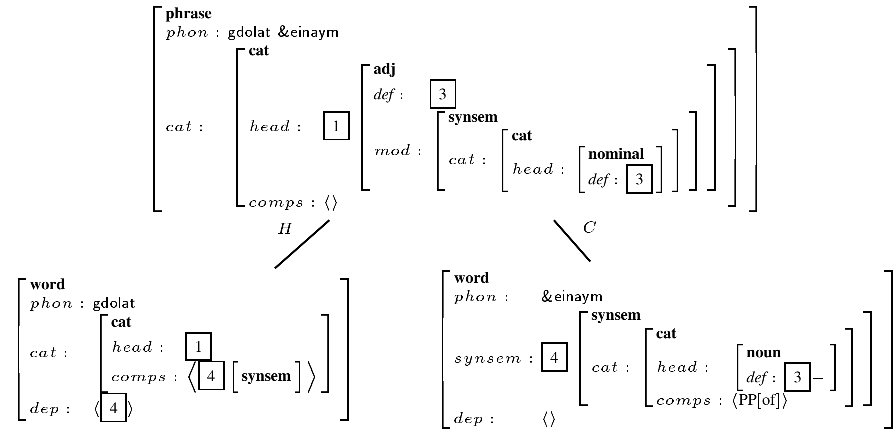
Esther Bentur, Aviella Angel, and Danit Segev. 1992. Computerized analysis of Hebrew words. *Hebrew Linguistics*, 36:33–38, December. (in Hebrew).

Hagit Borer. 1984. *Parametric Syntax - Case Studies in Semitic and Romance Languages*, volume 13 of *Studies in Generative Grammar*. Foris Publications, Dordrecht – Holland.

Hagit Borer. 1994. The construct in review. In Jacqueline Lecarme and Ur Shlonsky, editors, *Proceedings of the Second Conference on Afroasiatic Linguistics*, Sophia Antipolis, France, June. (to appear in *Studies in Afroasiatic Grammar*).

Robert D. Borsley. 1995. On some similarities and differences between Welsh and Syrian Arabic. *Linguis-*

(12)



tics, 33:99–122.

Yaacov Choueka and Yoni Ne'eman. 1995. "Nakdan-T", a text vocalizer for modern Hebrew. In *Proceedings of the Fourth Bar-Ilan Symposium on Foundations of Artificial Intelligence*, June.

Daniel Cohen. 1984. *Mechanical Syntactic Analysis of a Hebrew Sentence*. Ph.D. thesis, Hebrew University of Jerusalem. (In Hebrew).

Klaus Netter. 1994. Towards a theory of functional heads. In John Nerbonne, Klaus Netter, and Carl Pollard, editors, *German in Head-Driven Phrase Structure Grammar*, volume 46 of *CSLI Lecture Notes*, chapter 9, pages 297–340. CSLI, Stanford, CA.

Sergei Nirenburg and Yosef Ben-Asher. 1984. HUHU – the Hebrew University Hebrew understander. *Computer Languages*, 9(3/4).

Uzzi Ornan and Michael Katz. 1995. A new program for Hebrew index based on the Phonemic Script. Technical Report LCL 94-7, Laboratory for Computational Linguistics, Technion, Haifa, Israel, July.

Uzzi Ornan. 1994. Basic concepts in "romanization" of scripts. Technical Report LCL 94-5, Laboratory for Computational Linguistics, Technion, Haifa, Israel, March.

Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press and CSLI Publications.

Elizabeth Ritter. 1988. A head-movement approach to construct-state noun phrases. *Linguistics*, 26(6):909–929.

Elizabeth Ritter. 1991. Two functional categories in noun phrases: evidence from Modern Hebrew. In Susan D. Rothstein, editor, *Perspectives on Phrase Structure: Heads and Licensing*, volume 25 of *Syntax and Semantics*, pages 37–62. Academic Press.

Tali Siloni. 1991. Noun raising and the structure of noun phrases. In Jonathan David Bobaljik and Tony Bures, editors, *Papers from the third Student Conference in Linguistics*, volume 14 of *MIT Working Papers in Linguistics*, pages 255–270. Department of Linguistics and Philosophy, MIT, Cambridge, MA 02139.

Tali Siloni. 1994. *Noun Phrases and Nominalizations*. Ph.D. thesis, Départment de linguistique générale, Université De Genève.

Shuly Wintner and Uzzi Ornan. 1996. Syntactic analysis of Hebrew sentences. *Natural Language Engineering*, 1(3):261–288, September.

Shuly Wintner. 1992. Syntactic analysis of Hebrew sentences using PATR. In Uzzi Ornan, Gideon Ariely, and Edit Doron, editors, *Hebrew Computational Linguistics*, chapter 4, pages 105–115. Ministry of Science and Technology. (In Hebrew).

Shuly Wintner. 1997a. *An Abstract Machine for Unification Grammars*. Ph.D. thesis, Technion – Israel Institute of Technology, Haifa, Israel, January.

Shuly Wintner. 1997b. The affixal nature of the Hebrew definite article. Presented in *Computational Linguistics in the Netherlands (CLIN-97), University of Nijmegen*, December.

Shuly Wintner. 1998. Noun phrases as NPs – the case of Hebrew. In Gosse Bouma, Geert-Jan Kruijff, and Richard Oehrle, editors, *Proceedings of the 4th conference on Formal Grammar and the 5th conference on Head-driven Phrase Structure Grammar (FHCG'98)*, Saarbrücken, Germany, August.

Dana Yizhar. 1993. Computational grammar for Hebrew noun phrases. Master's thesis, Computer Science Department, Hebrew University, Jerusalem, Israel, June. (In Hebrew).

Arnold M. Zwicky. 1985. Heads. *Journal of Linguistics*, 21:1–29.