Noname manuscript No.
(will be inserted by the editor)

The Hebrew CHILDES Corpus

Transcription and Morphological Analysis

Aviad Albert · Brian MacWhinney ·
Bracha Nir · Shuly Wintner

Received: October 19, 2012/ Accepted: October 19, 2012

Abstract We present a corpus of transcribed spoken Hebrew that reflects spoken interactions between children and adults. The corpus is an integral part of the CHILDES database, which distributes similar corpora for over 25 languages. We introduce a dedicated transcription scheme for the spoken Hebrew data that is sensitive to both the phonology and the standard orthography of the language. We also introduce a morphological analyzer that was specifically developed for this corpus. The analyzer adequately covers the entire corpus, producing detailed correct analyses for all tokens. Evaluation on a new corpus reveals high coverage as well. Finally, we describe a morphological disambiguation module that selects the correct analysis of each token in context. The result is a high-quality morphologically-annotated CHILDES corpus of Hebrew, along with a set of tools that can be applied to new corpora.

Keywords CHILDES · Hebrew · Transcription of spoken language · Morphological analysis · Morphological disambiguation

1 Introduction

Recent years have witnessed the proliferation of computerized tools for processing natural languages with complex morphology. These tools serve language researchers by providing them with an interface that enables quick and

A. Albert
Department of Linguistics, Tel Aviv University. E-mail: aviad.albert@openu.ac.il

B. MacWhinney
Department of Psychology, Carnegie Mellon University. E-mail: macw@cmu.edu

B. Nir
Department of Communication Sciences, University of Haifa. E-mail: bnir@univ.haifa.ac.il

S. Wintner
Department of Computer Science, University of Haifa. E-mail: shuly@cs.haifa.ac.il

accurate analyses of large-scale corpora. This paper presents a corpus of transcribed spoken Hebrew that forms an integral part of a comprehensive data system that has been developed to suit the specific needs and interests of child language researchers: CHILDES (Child Language Data Exchange System; MacWhinney (2000)).

CHILDES is a system of programs and codes designed to facilitate the process of naturalistic speech analysis. It involves three integrated components:

1. A system for discourse notation and coding called CHAT (Codes for the Human Analysis of Transcripts), designed to accommodate different levels of linguistic analysis (e.g., phonological, morphological, or lexical), while maintaining a human-readable form of transcription;
2. A set of computer programs called CLAN (Computerized Language ANalysis), that provide researchers with pre-defined analyses specifically tailored for the study of child language acquisition; and
3. A large, internationally recognized database of language transcripts formatted in CHAT. These include child-caretaker interactions from normally-developing children, children with language disorders, adults with aphasia, second language learners, and bilinguals who have been exposed to more than one language in early childhood.

Researchers can directly test a vast range of empirical hypotheses against data from nearly one hundred major research projects in a wide variety of languages. Thus, although about half of the CHILDES corpus consists of English data, there is also a significant component of transcripts in over 25 other languages.

The CLAN software includes a language for expressing morphological grammars, implemented as a system, *MOR*, for the construction of morphological analyzers.[1] The main focus of the present paper is on the construction of a MOR grammar for Hebrew. Before examining this new system for Hebrew, however, it is important to understand why morphological analysis is so crucial for child language studies.

From its very beginning, the domain of language acquisition has put an emphasis on the development of grammatical competence. For example, the landmark analysis of the grammars of Adam, Eve, and Sarah (Brown, 1973) focused on the determinants of the order of acquisition of 14 grammatical morphemes. This seminal work led to parallel studies in dozens of other languages which taught us how the sequence of acquisition of grammatical morphemes and parts of speech was conditioned by a variety of interesting formal and functional factors. This research is summarized in the chapters of Slobin's 7-volume series on the cross-linguistic study of language acquisition (Slobin, 1985).

Eventually, various methods for morphological and part of speech analysis became codified in systems for assessment and diagnosis of language development, such as SALT (Miller and Chapman, 1983), DSS (Lee, 1974), LARSP (Crystal et al., 1976), and IPSyn (Scarborough, 1990). A major limitation

[1] The MOR program was initially developed by Roland Hausser and Mitzi Morris. It is described in detail in Hausser (1989).

of all of these systems has been that they require idiosyncratic hand-crafted tagging and analysis of each word in a transcript. As a result, these methods are both time-consuming and error-prone. To address this problem, the MOR program for automatic analysis and part of speech tagging was introduced into the CHILDES system. To date, MOR analysis programs have been constructed for Cantonese, Dutch, English, French, German, Italian, Japanese, Mandarin, and Spanish. Once a child language corpus has been automatically tagged by MOR, it is then possible to automate various systems for assessment and diagnosis. Recent examples include automated computation of the DSS score for English and Japanese (Miyata et al., 2009; Miyata and MacWhinney, 2011) and the IPSyn measure for English (Sagae et al., 2004).

Systems for further automatic analysis of child language corpora can also be grounded on processing of the morphological information. In particular, syntactic (Dependency Grammar) analyzers that rely on MOR tagging were developed for English and Spanish (Sagae et al., 2010). These morphological and syntactic analyses form a platform for further work on grammar induction from CHILDES corpora, as represented in various new learning algorithms (Bannard et al., 2009; Borensztajn et al., 2009; Freudenthal et al., 2010; Waterfall et al., 2010). Together, these methods for automatic analysis of morphosyntactic development promise to advance the study of language acquisition to new theoretical levels.

The current paper reports on the construction of a MOR system for Hebrew. Because of the richness of its allomorphic patterns and various orthographic complications, Hebrew poses a particularly interesting challenge to any system for automatic morphological analysis. We focus on the Hebrew section of the CHILDES database, and specifically on two major data-sets: the Berman longitudinal corpus, with data from four children between the ages of 1;06 and 3;05 (Berman and Weissenborn, 1991), and the Ravid longitudinal corpus, with data from two siblings between the ages of 0;09 to around 6 years of age. The corpora consist of 114,632 utterances comprising of 417,938 word-tokens (13,828 word-types).

This paper makes three main contributions: we devise a system for adequate transcription of the spoken utterances, and uniformly transcribe the two data-sets; we develop a highly accurate morphological grammar that is specifically designed for spoken Hebrew; and we present a disambiguation module that selects the correct analysis of ambiguous tokens in the context in which they occur. The main outcome of this work is therefore a morphologically-annotated corpus of spoken Hebrew, meticulously transcribed and annotated, which is already being used by several researchers of child language, language development, and language disorders. An additional outcome is a set of automated tools that can be used to accurately annotate new corpora with similar codes; such corpora are currently being transcribed, and we will apply the morphological analyzer and disambiguation module to more texts as they become available. This paper extends and revises several preliminary conference presentations (Nir et al., 2010; Albert et al., 2011, 2012).

We discuss the transcription scheme in the following section, and the scope of the annotation in Section 3. Section 4 details the morphological grammar; morphological disambiguation is the topic of Section 5. Section 6 provides a robust evaluation of the results. Finally, Section 7 discusses the ways in which the annotated database is already being used, and concludes with suggestions for future research.

2 Transcription of Spoken Hebrew Data

2.1 Desiderata

All data files in the CHILDES system are transcribed according to the CHAT format. This format allows researchers to decide whether they wish to apply orthographically-based transcription (as is, for example, the case for the English data-sets) or to rely on a phonemic representation of the speech data. The decisions involved in the transcription of data are not only technical but also theoretical (MacWhinney, 2000), and have a direct impact on later stages of the research process, when the data are subject to computerized analysis. The three major goals which the CHAT format is designed to achieve are systematicity and clarity, ease of data entry, and human and computerized readability (MacWhinney, 1996).

When dealing with Hebrew, relying on an orthographic representation of speech makes little sense. Hebrew is written from right to left; complex scripts are required for the full representation of vowels; and the representation of all phonemes is very different from the one used in Latin-based languages (Ravid, 2012). The Hebrew script can be used in two variants: *vocalized*, where diacritics mark the vowels; and *non-vocalized*, where much of the vocalic information is missing. The vast majority of modern texts use the latter.

The *non-vocalized* Hebrew orthography (the standard Hebrew script) lacks prosodic information and includes a very limited range of vocalic information. This state of affairs, where orthographic forms are, in fact, sequences of letters denoting mostly consonants, increases the number of homographs in conventional Hebrew script (Ornan and Katz, 1995; Wintner, 2004). Take, for example, the orthographic form **שירה**, which can be read in the following ways: *šayarā* "convoy"; *širā* "poetry / her poem"; *š̄ira* "Shira (proper name)"; *še# yarā* "that shot". Consequently, any computerized system that is to handle written Hebrew data would have to take into account the highly ambiguous nature of its orthography (Yona and Wintner, 2008).

The *vocalized* script solves the ambiguity problem but introduces a plethora of other problems. First, it uses diacritics, rather than alphabetic characters, to encode the vowels; this makes it difficult to specify morphological rules (Section 4) and to search the transcribed texts (search patterns can use variables to abstract over characters, but not over parts of characters). Second, the five-vowel system of Modern Hebrew is very different from the rich vowel system of biblical Hebrew, which is the one preserved in the orthography. Consequently, the standard diacritics encode redundant information and are therefore ambiguous. For example, indications of schwa are sometimes pronounced as a vowel (e.g., the בְּ in בְּרֹאשׁ *[beroš]*) but sometimes not (the same בְּ in בְּרוֹשׁ *[broš]*). Being able to make such distinctions is crucial for child-language research, where actual pronunciation of the elements in the utterance is the basis for studying all levels of linguistic development, from phonology and morphology to syntax, semantics, and pragmatics. Third, as a result of the discrepancy between the pronunciation of Modern Hebrew and the (vocalized) orthography, Hebrew speakers are unable to correctly and consistently produce the vocalized script, and hence transcription using this representation is likely to be costly and inaccurate.

2.2 A Hybrid Transcription Method

In light of the above, we chose to transcribe the Hebrew data in the CHILDES system in a Latin-based phonemic transcription. Existing Hebrew transcription approaches either use a phonemic transcription (Ornan, 1986, 1994) or employ one-to-one transliterations. The former reflect the inherent features of the language but are hard to learn and use; the latter miss much of the information, in particular the vowels. In contrast to previous transcription methods, the current transcription relies on phonemic, prosodic, and orthographic features. It also consistently separates out phrasal (e.g., prepositional) and clausal (e.g., subordinating) functional elements that in Hebrew are orthographically prefixed to the following word.

The Hebrew data in the CHILDES database were collected by different researchers and were meant to serve various research purposes, involving morphology, lexicon, and syntax. Consequently, the various data-sets were highly inconsistent in terms of the transcription methods that were originally used. The first step was thus standardizing the transcription of the data. All files were semi-automatically re-transcribed to conform to a newly devised set of CHAT-compatible transcription conventions (Nir et al., 2010). Since CHAT conventions do not permit the use of special ASCII characters (e.g., $, &, #) for representing consonants, we use a set of monoglyph Unicode characters (mostly in line with standard IPA conventions) that has already been applied for other complex scripts. Table 1 presents the complete set of transcription pairs.[2]

The advantages of our approach are summarized below:

- As illustrated above, phonemic transcriptions that include the five vowels of Modern Hebrew as well as prosodic information on primary stress location, yield fewer ambiguities.
- At the same time, the Hebrew orthography retains valuable phonetic and phonemic distinctions that no longer occur in Modern Hebrew speech. Such

[2] Diacritics are produced through addition of overprinting Unicode characters and are not single Unicode characters.

Consonants:

א	ב	ג	ד	ה	ו	ז	ח	ט	י	כ	ל	מ
ʔ	b/v	g	d	h	w	z	x	ṭ	y	k/ḳ	l	m
נ	ס	ע	פ	צ	ק	ר	ש	ת	ג'	ז'	צ'	
n	s	ʕ	p/f	c	q	r	š/ṣ	t	j̧	ž	ç	

Vowels (stressed and unstressed):

ā	ē	ī	ō	ū	a	e	i	o	u

Table 1 Transcription of Hebrew in CHAT format

orthographically distinct segments facilitate resolution of homophonic ambiguity that results from the loss of these phonemic distinctions in the first place. Consider, for example, the following pairs: כר *kar* "pillow" vs. קר *qar* "cold" (both are pronounced *[kar]*); or אח *ʔax* "brother" vs. אך *ʔaḳ* "however" (pronounced *[ax]*). Our transcription encodes such pairs differently.

- Recall that one of the possible readings for the Hebrew orthographic form שירה is *še# yarā* "that shot". The Hebrew letter ש, which denotes the consonant [š] in this case, can be interpreted as the subordinating conjunction *še#* "that". This is one of the clausal prefixes that orthographically combines with the stem. Marking these particles (with #) followed by a space in our transcripts allows us to consistently treat all syntactic functional elements and to recognize them as separate morphemes that never participate in homographs.

This final point reflects on one of the major issues involved in the transcription of spoken data, the question of what is a word. We discuss the various strategies we chose to handle this question in Section 2.3.

Figure 1 shows a brief example of an interaction between a Hebrew-speaking child and her caretakers, transcribed in CHAT. Child and adult utterances are listed one in a line, in what is called the *main tier*. Each line begins with a specification of the speaker (e.g., `CHI:` for the target child, `MOT:` for the mother). The CHAT format uses several special characters, such as `#` to indicate a prefix, or `[: ]` to indicate the correct form when the actual utterance is mispronounced.

It is important to note that Hebrew speakers find the transcription straightforward, and are able to read it with no training. Coding requires some minimal training, and in particular attention to details that are not present in the Hebrew orthography, such as stress; but lexicographers are able to transcribe new utterances reliably and fairly quickly. Coding errors can usually be detected due to the existence of a morphological analyzer: forms that cannot be analyzed are highlighted and can subsequently be inspected and corrected. Furthermore, conversion of our transcription to the standard (unvocalized) Hebrew script can be done automatically with high accuracy; we developed such a conversion program and use it for evaluation (see Section 6). Incidentally, the same issues were considered when the Japanese section of CHILDES

```
@Begin
@Languages: he
@Participants: CHI Hagar Child, MOT Inbal Mother, GRA Grandmother
GRA: Hagār, ʔat xolā .
GRA: ʔat yodāʕat še# ʔat xolā Hagāri ?
CHI: ava [: ʔavāl] [*] le# gag .
CHI: gag .
GRA: mi ze ?
CHI: ladow@c le# gag .
CHI: le# e@voc gag .
MOT: ʔīmaʔ loʔ holēḳet la# gag .
CHI: gag gag !
MOT: Hagāri .
GRA: ʔat rocā sipūr ?
GRA: bōʔi tavīʔi li sipūr we# ʔanī ʔasapēr laḳ .
CHI: le# gag !
MOT: loʔ meṣaxqīm ʕaḳšāyw ba# gag .
@End
```

Fig. 1 Example of the transcription

was transcribed; in the case of Japanese, practiced transcribers prefer Roman, but students and new transcribers prefer Kana-Kanji. Japanese, too, has an automatic conversion script between the two representations.

2.3 Lexical Strategies

As in any transcription method, including standard orthography, the question of "what is a word" is critical. Several issues emerge here, such as those involved in the characterization of phonological versus orthographic words (as in the cases described above of homophonic ambiguity, on the one hand, and of functional items that in Hebrew are written either in adjacency to the following content words or are separated by spaces).

A major issue that reflects not only on semantic but also on syntactic acquisition is that of complex expressions that are written as strings that contain more than one lexeme but that are considered by native speakers as constituting one lexical entry. Consider the following examples:

Multi Lexemic Expressions (MLEs) are sequences of words that together constitute one expression, or one complex syntactic entity (Sag et al., 2002). MLEs may combine words from different syntactic categories in various ways. In CHAT, MLEs are transcribed with an underscore replacing space between words (e.g., *šalāṭ_raxōq* "remote control", *lāyla_ṭov* "good night").

Compounds are special cases of MLEs (Clark and Berman, 1987) featuring initial lexicalized compounds at the early stages of acquisition and partially marked noun-noun combinations at around age 3 (Berman, 2009), as well as noun-noun combinations in *Construct-State* in speech directed to children (Borer, 1988, 1996). The transcription convention dictates that compounds be separated by a '+' sign instead of spaces (e.g., *bēged+yam* "swimsuit", *xadār+ʔōḳel* "dining room").

Merged forms are sequences involving proclitic particles (see Section 2.2) that are fully combined with the following word. Many Hebrew adverbs, for example, appear in such a construction, with the preposition *be#* "in" followed by an abstract noun (e.g., *beqōši* "barely" = *be# qōši* "in difficulty", Nir and Berman (2010)). Merged forms are transcribed with no separation between morphemes: *hayōm* "today" vs. *ha# yom* "the day".

Unfortunately, there is no straight-forward way to know whether complex expressions should be divided into sequences of isolated lexemes vs. one "frozen" lexical entity. While some cases may seem to be very clear in this respect, other cases corroborate claims that the degree of "lexicality" of complex expressions runs on a scale (Berman, 1979; Berman and Ravid, 1986). Various criteria, such as loss of compositional meaning, can prove helpful in making the decision but they do not appear to constitute lexicality (consider the merged forms examples above, which are mostly compositional yet considered fully merged). The current corpora were re-transcribed to meet the following criteria. An expression is considered lexicalized if its meaning is idiomatic, or if it is highly conventional in use by Hebrew speakers. In order to determine whether a particular expression is to be treated as lexicalized, we consolidated the judgments of five linguistics undergraduate and graduate students, all native Hebrew speakers who went over lists of the various entries recognized by transcribers as possible complex lexical entries. Those expressions that were perceived as one item by the majority of judges were introduced into the system as fixed expressions.

In addition to MLEs, spoken language in general, and child language in particular, includes an abundance of forms that are idiosyncratic and inconsistent with standardized forms, such as onomatopoeias and other cases of ad-hoc productions (including child forms, filler syllables, unique diminutive forms etc.). CHAT conventions mark such special forms with a code preceded by '@', which is suffixed to the word (e.g., *nad@c* is recognized as a child form due to the @c special-form marker). It is also possible to automatically recognize capitalized words as proper nouns, in line with English orthography (e.g., *Ron* is analyzed as a proper noun due to its initial upper-case letter). Note that since the Hebrew transcript includes some non-Latin characters which bear no case distinctions, the special-form marker @z:pn achieves the same goal by automatically recognizing a word as a proper noun (e.g., *ʔēli@z:pn* "Eli"). Our corpora mainly use the special-form markers listed in Table 2.

3 Scope of the Annotation

Corpus annotation has become a major effort in recent years, both for linguistic research and for natural language processing applications. Text corpora can be annotated in a variety of ways, reflecting various levels of linguistic infor-

@c	Child form
@co	Communicator
@si	Singing
@x	Unknown
@voc	Vocalizer
@z:pn	Proper noun
@z:oc	Completion
@z:dim	Diminutive noun
@z:dima	Diminutive adjective

Table 2 Special codes in the CHAT transcripts

mation. The Hebrew CHILDES data consist of transcribed spoken speech, and the choice of transcription adds yet another degree of freedom to the decision.

We opted for a transcription scheme that reflects morphological distinctions but does not provide narrow phonetic information; narrow phonetic transcription would have required much more time and effort. We did, however, retain phonetic information when it was transcribed, especially in the case of child utterances, using conventions available in CHAT. Note that the actual audio tapes on which the conversations were recorded are available, and we have very recently completed the process of digitizing them. Eventually, we would like to synchronize the recordings with the transcriptions, as was done for other languages.

The main focus of the present work is on morphological analysis and disambiguation. Having said that, we do intend to extend the corpora with syntactic information, in the form of grammatical (dependency) relations, in the near future; this work is currently underway. In their final form, then, the corpora will be of use to any linguist interested in the lexical, morphological, or syntactic processes involved in child language acquisition. Specifically, the annotated corpora will support three different lines of investigation:

1. Basic child language developmental research that examines the sequence of acquisition of grammatical morphemes and the various morphosyntactic processes of the language (e.g., Berman (1981)).
2. Computational modeling of child language acquisition (Bannard et al., 2009; Borensztajn et al., 2009; Freudenthal et al., 2010; Waterfall et al., 2010).
3. Diagnosis of language differences and disorders through methods such as automation of IPSYN or DSS scores (Sagae et al., 2004; Miyata et al., 2009; Miyata and MacWhinney, 2011).

4 Morphological Analysis

4.1 Hebrew Morphology

Hebrew is a language with rich morphology both in terms of semantics and of systemics (Berman, 1985; Ravid, 2012). Some notions, such as Number, Gender, Person, and Tense, are obligatorily encoded as part of both content and function word structure; other notions, such as possession and objecthood, can be optionally expressed as morphemes rather than syntactically. Processes of inflectional morphology in Hebrew are realized either through affixation or through vocalic changes to the stem (ablaut).

For example, the verb *katav* "write" appears in (1) in the *base (citation)* form, denoting masculine, 3rd person singular, in past tense. A change of person inflection, to the 1st person, yields the form *katav+ti* "I wrote" (2) by affixation. A change in tense inflection, to the present tense, yields the verb *kotev* "I/you/he/it write/s" (3), illustrating a case of ablaut.

1. *katāv* "he wrote"
2. *katāv+ti* "I wrote"
3. *kotēv* "I/you/he/it write/s"

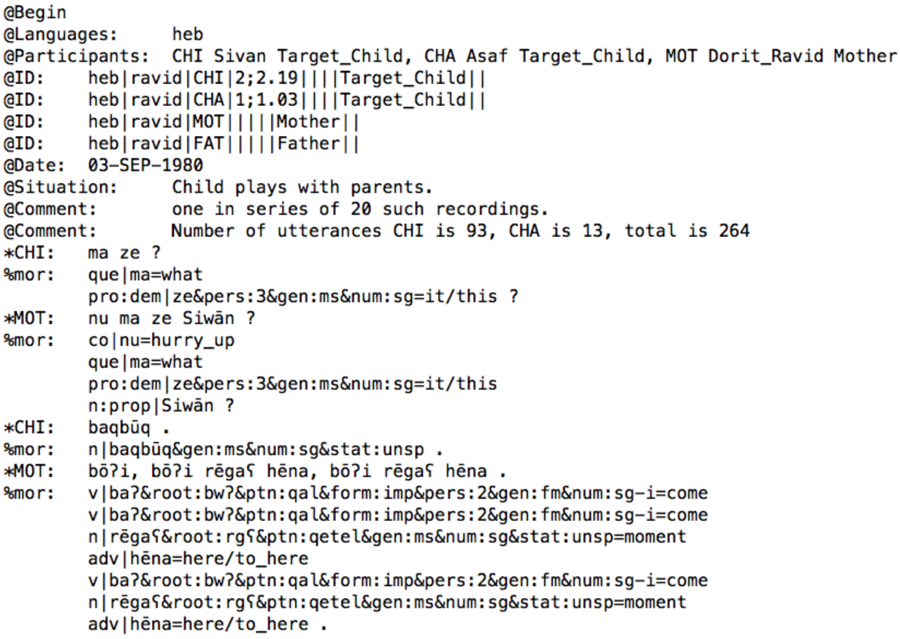
The set of inflectional affixes in Hebrew is limited and mostly stable. In many cases, it is possible to predict which affix attaches to a given stem. However, this predictability is different for the verbal and nominal system, where the former is highly regular while the latter includes numerous cases of lexical exceptions. Thus, for nouns, the assignment of inflectional affixes (such as the two plural suffixes, *+īm* and *+ōt*), as well as stress shifts and morphophonemic changes to the stem are not fully predictable. In many cases, the only way to predict such processes is by resorting to full Hebrew orthography, which preserves some inactive phonological distinctions that may motivate morphological processes. In other cases, only historical accounts may serve to explain and motivate changes that are otherwise completely opaque to speakers.

As for derivational morphology, in accordance with most standard analyses of Semitic languages, it is widely assumed that many Hebrew *stems* (or *bases*) are combinations of two non-linear morphemes: a sequence of consonants, often referred to as *root* or *radic* consonants, and a templatic sequence of vowels (and, optionally, consonants) with empty slots for root consonants, generally referred to as *pattern*, or more specifically *binyan* for verbs and *mishkal* for nouns (Ravid, 2012). As an illustration, consider the consonantal sequence *x.z.r.* The basic verbal pattern *CaCaC* yields the verb *xazar* "return", while the verbal pattern *CiCeC* that denotes transitive activity yields the verb *xizer* "court". Thus, the semantics of different words that share the same consonantal sequence, even when observed within the paradigmatic verbal system, may radically differ.

It should be noted that there is an ongoing debate on the reality of non-linear morphemes as active lexical entities in the grammar of Modern Hebrew speakers. In contrast to many common assumptions, some arguments draw on the universality of morphological systems, and current denominalization processes in Hebrew, to claim that the mental lexicon of Hebrew speakers consists of words, not vocalic patterns and consonantal sequences (Bat-El, 1994; Ussishkin, 1999). At the same time, several new studies show psycholinguistic evidence that the *root* or *radic* consonants are indeed active entities in the mental lexicon of Hebrew speakers (Shimron, 2003, chapter 1 and references therein). Our design of the Hebrew morphological analyzer is intended to be generally useful, regardless of a particular researcher's viewpoint on the above-mentioned theoretical debate.

4.2 MOR Devices

Recall that the CLAN software includes a language for expressing morphological grammars, implemented as a computer program, *MOR*. Applying a MOR grammar to a CHILDES corpus creates a new tier below each main tier, called the *%mor* tier, in which the morphological information for each item in the main tier is listed. The output provides the surface representation of concatenated linear morphemes (stems + affixes) and other morphological (and lexical) information attributed to the surface token. Figure 2 depicts a small fragment of a morphologically-annotated corpus.



```
@Begin
@Languages:     heb
@Participants:  CHI Sivan Target_Child, CHA Asaf Target_Child, MOT Dorit_Ravid Mother
@ID:    heb|ravid|CHI|2;2.19|||Target_Child||
@ID:    heb|ravid|CHA|1;1.03|||Target_Child||
@ID:    heb|ravid|MOT||||Mother||
@ID:    heb|ravid|FAT||||Father||
@Date:  03-SEP-1980
@Situation:     Child plays with parents.
@Comment:       one in series of 20 such recordings.
@Comment:       Number of utterances CHI is 93, CHA is 13, total is 264
*CHI:   ma ze ?
%mor:   que|ma=what
        pro:dem|ze&pers:3&gen:ms&num:sg=it/this ?
*MOT:   nu ma ze Siwān ?
%mor:   co|nu=hurry_up
        que|ma=what
        pro:dem|ze&pers:3&gen:ms&num:sg=it/this
        n:prop|Siwān ?
*CHI:   baqbūq .
%mor:   n|baqbūq&gen:ms&num:sg&stat:unsp .
*MOT:   bōʔi, bōʔi rēgaʕ hēna, bōʔi rēgaʕ hēna .
%mor:   v|baʔ&root:bwʔ&ptn:qal&form:imp&pers:2&gen:fm&num:sg-i=come
        v|baʔ&root:bwʔ&ptn:qal&form:imp&pers:2&gen:fm&num:sg-i=come
        n|rēgaʕ&root:rgʕ&ptn:qetel&gen:ms&num:sg&stat:unsp=moment
        adv|hēna=here/to_here
        v|baʔ&root:bwʔ&ptn:qal&form:imp&pers:2&gen:fm&num:sg-i=come
        n|rēgaʕ&root:rgʕ&ptn:qetel&gen:ms&num:sg&stat:unsp=moment
        adv|hēna=here/to_here .
```

Fig. 2 A fragment of the annotated corpus

A MOR grammar consists of three components: a set of *lexical files* specifying lexical entries (base lexemes) and lists of affixes; a set of rules that govern allomorphic changes in the stems of lexical entries (*A-rules*); and a set of rules that govern linear affixation processes by concatenation (*C-rules*).

Different languages vary in their requirements and their need to utilize these MOR devices. The Dutch grammar, available on the CHILDES database (http://childes.psy.cmu.edu/morgrams/), includes a short list of possible allomorphic changes and several concatenation rules; the grammar for Italian includes a more elaborate A-rules file and a relatively short C-rules file; while the grammars for Cantonese and Mandarin rely almost exclusively on lexicon files. The Hebrew MOR extensively uses all of these devices: bases and affixes are noted in the lexicon files; the rich system of vocalic and consonantal changes of the stem allomorphs is handled within a set of A-Rules; and the proper affixation possibilities are allowed (or restricted) via the C-rules. We now introduce each of these devices in detail; for more information, see MacWhinney (2000, 2008).

4.2.1 Lexicon

The Hebrew grammar is focused on inflectional morphology, rather than derivational morphology. In this sense it is also in line with other Hebrew morphological analyzers, such as the MILA morphological analyzer of written Hebrew (Itai and Wintner, 2008). Consequently, lexical entries are headed by lexemes (rather than root consonants, a design option that is possible in MOR) in a way that reflects the organization of printed dictionaries. As a result, lexical lists are uniform, more human-readable, and require fewer tags to correspond to the correct A-rule.

Example 1 depicts four lexical entries, one per line. Each entry is introduced by the base (citation) form. Then, in curly brackets, lexical features are specified. These include the main syntactic category (scat) of the item, as well as features such as root, pattern (ptn), grammatical gender (gen), etc. Also listed are features that are necessary for proper generation of the inflected forms of the entry. For example, Hebrew adjectives inflect for gender, and feminine variants are realized with a suffix. But the choice of suffix is lexically determined, and has to be specified explicitly; hence the value a for the feature fem of the adjective *xām*. In the same way, it is important to specify that *māyim* "water" is a mass noun, to prevent an erroneous attachment of the plural suffix. Finally, some features (e.g., vchng) specify information that is used by the A-rules to generate the correct set of allomorphs; as noted above, this set, especially in the case of nouns, involves information that cannot be deduced from the form of the lexeme. In contrast, note that the verb system is so regular that no such information is needed; the entire inflectional paradigm of *katāv* "write" is determined by its form. Lexical entries can also specify an English gloss, between '=' symbols.

Example 1 (Lexical entries)

```
katāv {[scat v][root ktb][ptn qal]} =write=
māyim {[scat n][gen ms][infl mass][vchng 1e0-3-0]} =water=
daq {[scat adj][fem a][root dqq][vchange 0-spir]} =thin=
dey {[scat adv]} =quite=
```

4.2.2 A-rules

A-rules specify how to generate allomorphic variants from stems. An A-rule first specifies the lexical entries to which it applies. This can be done by specifying the actual surface form of the lexical entry (using `LEXSURF`), or by listing some morpho-syntactic features, via `LEXCAT`. In Example 2, the `LEXCAT` specification constrains the rule to apply only to verbs whose pattern (`ptn`) is `qal`. Additionally, the `LEXSURF` specification indicates that this rule is only applicable to lexical entries whose surface form is given by the pattern `$Qa$Tā$L`. This pattern makes use of *variables*, such as `$Q` and `$T`. Variables are defined separately, and their values can be constrained. In the Hebrew MOR grammar, the variables `$Q`, `$T`, and `$L` range over all consonants; other variables range over vowels, stressed vowels, gutturals, etc.

Example 2 (A-rules)

```
LEXSURF = $Qa$Tā$L
LEXCAT = [scat v], [ptn qal]
ALLOSURF = $Qa$Tā$L
ALLOCAT = LEXCAT, ADD [tense past], ADD [allo p1-2]
ALLOSURF = $Qa$T$L
ALLOCAT = LEXCAT, ADD [tense past], ADD [allo p3]
```

Then, an A-rule specifies one or more allomorphs for the lexical entry. Example 2 shows two such allomorphs, one for first and second person past tense, and one for third person past. In the former, the surface form of the allomorph (indicated by `ALLOSURF`) is exactly that of the lexical entry. In the latter, the second vowel of the lexeme is reduced. The `ALLOCAT` specification adds feature–value pairs to the ones that are associated with the lexical entry. This particular example, when applied to the lexeme *katāv* “write” (Example 1), yields two allomorphs: *katāv*, with the feature–value pairs {`[scat v]` `[root ktb]` `[ptn qal]` `[tense past]` `[allo p1-2]`}; and *katv*, with the feature–value pairs {`[scat v]` `[root ktb]` `[ptn qal]` `[tense past]` `[allo p3]`}. In order to generate the final forms, however, affixes may have to be attached to these allomorphs. This is achieved by C-rules.

4.2.3 C-rules

All the linear inflectional morphology in Hebrew is handled within the C-rules, allowing a better division of labor in which the A-rules do not include linear affixation of inflectional categories, and surface representation of linear morphemes are readily visible to MOR.

C-rules use a lexicon of affixes, which specifies both the surface form and the morphological features of affixes. Example 3 depicts the lexicon entry of the suffix *ti*. Its syntactic category is determined to be `pastsfx`, as it is a past-tense suffix. It also carries with it the information that it is first person, singular, and unspecified for gender. Finally, it includes the specification `[allo p1-2]`. This is part of a *lock-and-key* mechanism that controls which affixes combine with which allomorphs. Refer back to Example 2, and note that one of the allomorphs (but not both!) was also associated with `[allo p1-2]`. Now, Example 3 also shows part of the specification of the rules that govern the combination of verbs with their past-tense suffixes. The rule states, through the `MATCHCAT` specification, that the allomorph and the suffix must have matching values for the feature `allo`. For the suffix *ti*, this is indeed the case with the allomorph *katāv* of Example 2, but not with *katv*. Hence the final form that will be generated by this specific rule when it applies to *katāv* is *katāvti*. This form will be associated with several feature-value pairs, as we show below.

Example 3 (C-rules)

- Affix entry
  ti {`[scat pastsfx][allo p1-2][pers 1][gen unsp][num sg]`}
- C-rules excerpt
  ```
  RULENAME: v-past
  NEXTCAT = [scat pastsfx]
  MATCHCAT [allo]
  ```

4.2.4 Putting it all together

When MOR is invoked, it passes each lexical form past the A-rules. Importantly, the A-rules are strictly ordered. If a form matches a rule, that rule fires and the allomorphs it produces are generated. Then MOR moves on to the next lexical form, without considering any additional rules. This means that it is important to place more specific cases before more general ones in a standard bleeding relation. The C-rules, in contrast, are not ordered, and all possible combinations of affixes and allomorphs are attempted. Those combinations that survive `MAATCHCAT` specifications are propagated to the output.

As illustrated here, the lexical and morphological information that appears in the output of MOR can be added in various stages of the output derivation: it can be specified in a feature-value pair in the lexical entry; it can be added through an ADD operation in an A-rule; or it can be modified through an ADD or DEL operation in a C-rule. Some of this information is only required for proper derivation of correct forms, but is not specified in the output of the analyzer. The types of morphological information that are propagated to the output analysis are separately defined in a special file, the *output file*. This is crucial for the analysis of Hebrew, as it allows us to state features according to the specific needs and theoretical assumptions of the researcher. For example, information such as the root consonants is stated in the lexicon, and can be propagated to the output of MOR, thus allowing researchers the choice to decide whether they consider non-linear morphemes as separate morphemes or not. This choice has a significant impact on assessment of child language development, for example on the computation of mean length of utterance (MLU).

Example 4 shows the output of the morphological analyzer given the surface form *katāvti*. Each string on the MOR line begins with the main part of speech category; this is followed by the morphemes forming the surface form, and then by a sequence of feature–value pairs, separated by '&' – depending on the category being analyzed. Each string is terminated by the gloss, which is introduced by '='.

Example 4 (Output format)

```
v | katāv-ti & root:ktb & ptn:qal & tense:past
    & pers:1 & gen:unsp & num:sg = write
```

The analysis presented in Example 4 reflects the various morphological features relevant for Hebrew. Features that are listed in the output analysis (`root`, `ptn`, `tense`, `pers`, `gen`, and `num`) are individually stated in the output file. Note that the feature `allo`, which is added in the A-rule excerpt of that example, does not occur in the output analysis, as it is only needed for the inner-workings of MOR (more specifically, it is used to properly match stems and affixes within the grammar).

The Hebrew MOR grammar displays the following features in the output analysis:

`root` The underlying root (when it is lexically specified, e.g., `ktb`)
`ptn` The underlying pattern (when it is lexically specified, e.g., `nifal`)
`form` Infinitive and imperative forms of verbs (`inf` or `imp`)
`tense` Tense of verbs (`past`, `present` or `future`)
`pers` Person of verbs, pronouns and pronominal clitics (`1`, `2`, `3`, or `uns`pecified)
`gen` Gender of verbs, nouns, adjectives, etc. (`ms`, `fm`, or `unsp`)
`num` Number of verbs, nouns, adjectives, etc. (`sg`, `pl`, or `unsp`)
`pl` Type of plural suffix, which can be regular (*ōt* for forms that are morphologically marked as feminine, *īm* for unmarked forms) or irregular
`poss` Possessive inflections of nouns (e.g., `2femSG`, `3mascPL`)
`src` The lexical source of content words (e.g., `deverb`, `denom`, `foreign`)
`stat` The status of nouns with respect to construct state morphology (`bound`, `free`, or `unsp`)

4.3 A Hebrew MORphological Grammar

This section illustrates the operation of the Hebrew MOR grammar and provides examples for solutions that were required in order to handle challenges that emerged from the structure of the language.. We begin (Section 4.3.1) with an overview of general design decisions. We then detail the grammars of two major word categories: verbs (Section 4.3.2) and nouns (Section 4.3.3). We review general phonological and orthographic processes in Section 4.3.4, and discuss the overall organization of the grammar, including the minor part-of-speech categories, in Section 4.3.5.

4.3.1 General Design Decisions

The major challenge in devising a morphological grammar for Hebrew lies in accounting for the various changes that Hebrew stems undergo when they are inflected. In many cases, the vocalic patterns of stems change when affixes are added (as well as in cases of "pure" ablaut). These are morphophonemic changes within a given template – not to be confused with the change of templatic patterns, which is a derivational process. Standard Hebrew orthography, which includes mostly consonants (and very few vowels), is usually oblivious to these changes in the stem, yet our Hebrew transcription convention, which includes phonemic, vocalic and prosodic information (such as vowels and stress assignments) is able to represent them. For the Hebrew MOR, such state of affairs shifts much of the word analysis burden into the A-rules, as it is the list of rules that governs the various stem allomorphs.

A-rules correspond to surface forms of lexical entries. Surface forms within a rule are either stated explicitly, or, more generally, via pre-defined variables. For example, stressed and unstressed vowels: `{ā|ē|ī|ō|ū} / {a|e|i|o|u}` correspond to two different variables ($O for stressed and $V for unstressed). These variables are required to properly identify lexical entries via the position of stressed syllables (only primary stress in Hebrew), and to account for stress shifts under inflection. In the next set of examples we demonstrate how the use of variable groups allows us to treat some of the morpho-phonological phenomena in Hebrew within the A-rules mechanism.

4.3.2 The Verbal System

As noted, the Hebrew verbal system is highly regular. Each verb is derived from one of five major vocalic templates (termed *binyanim*).[3] Most of the morphophonemic alternations in the verbal system can be predicted from the phonemic and orthographic quality of the segments involved in the conjugation. Within MOR's A-rules, this systematic behavior is readily handled via the unique surface forms of lexical entries (including segmental features captured by variable sets) and the precedence effect of rule ordering (the first A-rule that fits an entry is in charge of all the derived stem allomorphs, and blocks any other subsequent rule from operating on that entry again).

However, the default behavior of MOR's rule ordering is problematic when there is one list of A-rules, where only one rule can be active, since a full inflectional paradigm must be generated for each unique verb. We divide verbal paradigms into 5 subsections reflecting the distinctive tense/form divisions (past, participle/present, future, infinitive and imperative). If, for example, two verbs differ only in some of those subsections (say, only under concatenation of prefixes in the future tense inflections), two different A-rules will have to account for that, and those two A-rules will be completely redundant in all the

[3] There are two other verbal templatic patterns, the passive counterparts of two of the five major *binyanim*. These are fully predictable from their active counterparts.

subsections excluding the future tense (i.e., past, participle/present, infinitive and imperative).

Consider, for example, the verbs *qašār* "tie" and *ʔasār* "forbid/imprison" in Example 5. Their forms in either the past or present tense are similar, but they differ when inflected for future tense or when in infinitival or imperative form, due to the unique interaction of guttural consonants (in this case, represented by the {*ʔ*} in the first consonantal position of the verb *ʔasār*) with their environment.

Example 5 (Inflectional subsections)

Rule excerpts (fitting both *qašār* and *ʔasār*):

past tense

```
ALLOSURF = $Qa$Tā$L (+ti/+ta/+t/+nu/+tem/+ten)
ALLOSURF = $Qa$T$L (+ā/+ū)
```

present tense

```
ALLOSURF = $Qo$Tē$L (+et)
ALLOSURF = $Qo$T$L (+īm/+ōt)
```

Rule excerpts (fitting only *ʔasār*):

future tense

```
ALLOSURF = $Ae$Tō$L (ʔe+/ne+/te+/ye+)
ALLOSURF = $Ae$T$L (te+/ye+ BASE +ī/+ū)
```

imperative form

```
ALLOSURF = $Ae$Tō$L
ALLOSURF = $Ai$T$L (+ī/+ū)
```

infinitive form

```
ALLOSURF = $Ae$Tō$L (le+)
```

Rule excerpts (fitting only *qašār*):

future tense

```
ALLOSURF = $Q$Tō$L (ʔe+/ni+/ti+/yi+)
ALLOSURF = $Q$Te$L (ti+/yi+ BASE +ī/+ū)
```

imperative form

```
ALLOSURF = $Q$To$L
ALLOSURF = $Qi$T$L (+ī/+ū)
```

infinitive form

```
ALLOSURF = $Q$Tō$L (li+)
```

The past and present tense inflections in Example 5 exhibit redundancy in the list of rules. To resolve this issue, we introduced a change in MOR's protocol: instead of one ordered list of A-rules, which is able to allocate one rule per lexical entry, we now allow multiple ordered lists of A-rules, each stored in a separate file, still governed by rule-ordering. This system can then allow more than one rule allocation per lexical entry (as many as one rule per list). Consequently, we divided all the verbal A-rules into five lists (corresponding to the five possible verbal forms mentioned above). Each such list consists of the same section of a paradigm from all the five templatic patterns (i.e., an A-rules list such as "Verbs-future" includes the section of future tense inflections from all the verbal templates). Since all lexical entries are full words, and the verbal templates differ in their surface forms, the various templates are immediately distinct. Each template (within a given paradigm section) has its A-rules ordered as follows (the ordering of rules is crucial only within each of these sub-groups):

1. Rules that correspond to standard citation forms (i.e., the uniform standard vocalic pattern), and can be properly ordered via variables, are ordered as "unmarked". For example, lexical entries such as *qašăr* "tie" and *ʔasār* "forbid/imprison" correspond to the standard *CaCāC* template. Within the A-rules list of future tense inflections, the two different rules that account for the different behavior of these verbs are properly identified by the unique variable sets and rule ordering (as in Example 5 above; note that the rule that accounts for *ʔasār* must be ordered before the general rule accounting for *qašăr*).
2. Rules that correspond to non-standard citation forms (i.e., feature a deviation from the standard vocalic pattern in a given template), and can be properly ordered via variables, are ordered as "special-forms". For example, within the *CaCāC* (*qal*) template there are some unique verbs that appear with a different base form, such as *gar* "dwell" (base form=*CaC*) and *qanā* "buy" (base form=*CaCā*). Any such unique deviation from the base form would fire a different A-rule, due to the different `LEXSURF` values. Each special form is (potentially) a separate ordered list of rules.
3. Rules that cannot be properly ordered via variables (due to morphologically conditioned alternations, and/or inability to capture a phonological condition), are listed under "marked". Take, for example, the verb *qaṭān* "diminish" (again, in the *CaCāC* template). In the infinitive form, this verb behaves according to the general rule, but in the past, present and future tense and in the imperative form this verb behaves differently. For the given example, the lexical entry would bear unique and general features (e.g., `part 57`, `fut 4`) as follows:

   Example 6 (Marked verbal entries)
   lex entry: qaṭān
   `[scat v][root qṭn][ptn qal][part 57][past 57][fut 4][imp 4]`

   Since the unique behavior of this verb is not predictable from the segmental quality of the verb (there are no consonants that correspond to any pre-defined set of variables), it is necessary to add unique feature-value matching pairs to both the lexical entry and the corresponding A-rule.[4]

Only rules that are listed as "marked" have to bear unique features that would allow the right lexical entry to correspond to an A-rule. Consequently,

[4] Note that the features `scat`, `root` and `ptn` are the general verbal features that are propagated to the output (syntactic category, consonantal root and pattern, respectively). Unlike these, the features `part`, `past`, `fut` and `imp` are only required for the proper A-rule match within the designated subsections (participle/present tense, past tense, future tense and imperative forms, respectively).

the vast majority of verbal lexical entries do not need to bear any unique features beyond the general lexical information that each verb is given.

4.3.3 The Nominal System

Compared to the verbal system, the nominal system in Hebrew is more irregular. While nouns in Hebrew obligatorily inflect only for gender (when gender is biologically determined, see Ordan and Wintner (2005)) and number (except for some irregular cases) through suffixation, some highly frequent lexical exceptions do occur, where a noun that is morphologically marked as feminine is assigned the plural suffix that commonly fits masculine nouns, and vice versa (Ravid et al., 2008). Like other entities in MOR, suffixes bear feature-value pairs that ensure proper matching with other MOR entities. Thus, these cases are handled by specifying unique features in the A-rules and on entries in the lexicon files. In addition to the irregularity of suffixation, the changes in the vocalic pattern (and stress assignment) that nominal stems undergo when inflected can only be partially predicted by phonological or orthographic regularities. These morphophonemic alternations can often be explained only diachronically and they are unpredictable on the surface.

All the affixes in the nominal system are suffixed to the base. Initially, it seemed reasonable to associate each inflectional category with its own distinctive feature, and assign the same value to any set of suffixes that would always fit the same allomorph together. Consider, for example, the set of (optional) singular possessive suffixes (Example 7). When inflected for some singular possessive conjugation, a regular noun, such as *kidōn* "spear", would have one allomorph (*kidon*) that fits all the different singular possessive suffixes (both (a) and (b) in Example 7). However, in many other cases the inflecting nouns require different allomorphs to match (a) and (b). Such is the case of the noun *šomēr* "guard": the allomorph *šomr* fits (a) while the allomorph *šomer* fits (b).

Example 7 (Nominal suffixes)

Affixes list excerpt:

Singular possessive affixes (a):
`+ī, +ēnu, +ēḳ, +āh, +ō, +ān, +ām`
Singular possessive affixes (b):
`+ḳā, +ḳēn, +ḳēm`

In order to handle this issue, we first assigned the feature `suff1` to the set of singular possessive suffixes (plural possessive suffixes receive the feature `suff2`). This allowed us to give sets (a) and (b) two distinctive values, say `a` and `b`, respectively. This works well for the A-rule corresponding to *šomēr* "guard", since there are two different allomorphs that attach to each set (see Example 8).

Example 8 (Preliminary nominal attachments)

Variables:

```
X = (anything)
```

Rule excerpts:

```
LEXSURF = $X$O$L
```

(Fits base-form nouns, such as: *šomēr* “guard”)

```
ALLOSURF = $X$L
ADD [suff1 a]
```

(Matches the allomorph *šomr* and set (a) of the singular possessive suffixes)

```
ALLOSURF = $X$V<O$L
ADD [suff1 b]
```

(Matches the allomorph *šomer* and set (b) of the singular possessive suffixes)

However, this architecture is not appropriate when applied to nouns where only one allomorph attaches to both sets of possessive suffixes, as in the case of noun *kidōn* “spear”. The problem is that both set (a) and (b) share the same feature (`suff1`) and allomorphs in the A-rules can state a certain feature only once (it is impossible to state both `ADD [suff1 a]` and `ADD [suff1 b]` for the same allomorph).

To solve this issue, we switch the lock-and-key unification mechanism of feature-value pairs: we group all the subsets of suffixes that always attach together to the same allomorph, and create a different feature that identifies that group alone. Instead of assigning distinctive values to a shared *feature*, we assign distinctive features to a shared *value*. Therefore, we can assign the feature `suff1a` to set (a) and the feature `suff1b` to set (b). We also assign the value `11` to both features. This allows the allomorph *kidon+* to state both `ADD [suff1a 11]` and `ADD [suff1b 11]` for the same allomorph. It is thus possible to allow any allomorphic stem in the A-rules to state any group of possibly fitting suffixes, since they never repeat the same feature.

Consider Example 9. Three allomorphs of the stem *sēfer* “book” are listed, each stating different matching groups of suffixes, with different features: `[suff0]`, `[suff1a]`, `[suff1b]`, `[suff2a]`, `[suff2b]`, and `[suff3]` (here, 0 indicates plural, 1 indicates singular possessive, 2 indicates plural possessive, and 3 indicates construct state). All allomorphs contain the same value, `[11]`, which is shared by all nouns that do not inflect for gender and take the plural suffix *+im*.

The allomorph *sfar*, the value of `ALLOSURF`, fits two subsets of suffixes, specified under `ALLOCAT`:

`suff0` for plural inflections (e.g., *sfar+īm* “books”)
`suff2a` for plural possessive inflections (e.g., *sfar+āy* “my books”)

Similarly, the allomorph *sifr* fits three subsets of suffixes:

`suff1a` for singular possessive inflections (e.g., *sifr+i* “my book”)
`suff2b` for plural possessive inflections (e.g., *sifr+eyḳēm* “your books”)
`suff3` for plural construct-state inflections (e.g., *sifr+ēy* “books of”)

Finally, the allomorph *sifre* fits one subset of suffixes:

`suff1b` for singular possessive inflections (e.g., *sifre+ḳā* “your book”)

Example 9 (Nominal suffix subsets)

Rule excerpts:
`LEXSURF = $Q$O$T$V$L` (*sēfer*)
(Fits base-form nouns, such as: *sēfer* “book”)
1st.
`ALLOSURF = $Q$Ta$L` (*sfar+*)
`ALLOCAT = LEXCAT, ADD [suff0 11], ADD [suff2a 11]`
2nd.
`ALLOSURF = $Qi$T$L` (*sifr+*)
`ALLOCAT = LEXCAT, ADD [suff1a 11], ADD [suff2b 11], ADD [suff3 11]`
3rd.
`ALLOSURF = $Qi$T$Le` (*sifre+*)
`ALLOCAT = LEXCAT, ADD [suff1b 11]`

Compare Example 9 with Example 10, where only two allomorphs are listed, taking different subsets of suffixes.

Example 10 (Nominal suffix subsets)

Rule excerpts:
`LEXSURF = $X$O$L` (*maqēl*)
(Fits base-form nouns, such as: *maqēl* “stick”)
1st.
`ALLOSURF = $X$L` (*maql+*)
`ALLOCAT = LEXCAT, ADD [suff0b 12], ADD [suff1a 12], ADD [suff2 12]`
(Allomorph for plural inflection (`suff0b`, e.g., *+ōt*); subset (a) singular possessive inflections (`suff1a`, e.g., *+ī*); and plural possessive inflections (`suff2`, e.g., *+otāy*)
2nd.
`ALLOSURF = $X$V<O$L` (*maqel+*)
`ALLOCAT = LEXCAT, ADD [suff1b 12]`
(Allomorph for subset (b) of the singular possessive inflections (`suff1b`, e.g., *+ḳā*))

4.3.4 Variables of Phonological and Pseudo-Phonological Classes

Our system of variables is designed to represent phonological natural classes in Hebrew, as well as pseudo-phonological classes that are only reflected in the orthography of the language (Section 2). Morphophonemic alternations can be often predicted in accordance with phonological natural classes. For example, two seemingly similar stems may be expected to behave differently under inflection, if only one of them contains a (pseudo-)guttural consonant. The following is a list of such variable classes (signs between curly brackets represent our transcript).

Stops and fricatives: {p/b/k}/{f/v/ḳ} The historical spirantization rule in Hebrew alternates between stop and fricative consonants (in general: fricative consonants appear after a vowel, stops appear elsewhere). In Modern Hebrew, a triplet of stop-fricative alternation still exists pervasively with a set of three stop-fricative pairs, each represented by the same letter:

– The Hebrew letter פ denotes a voiceless labial stop {p} or fricative {f}, phonemic /p/ and /f/ (respectively).
– The Hebrew letter ב denotes a voiced labial stop {b} or fricative {v}, phonemic /b/ and /v/ (respectively).
– The Hebrew letter כ denotes a voiceless velar stop {k} or fricative {ḳ}, phonemic /k/ and /x/ (respectively).

In order to allow our analyzer to identify this alternation and to recognize different surface forms as belonging to the same stem, we define two sets of consonants – one set (P) that includes all consonants minus the fricative triplet, and another set (F) that includes all consonants minus the stop triplet. Within the variable definition, all consonants are listed in the same order, while the alternating triplet are located in identical positions in each set. We then use *variable-shift* in the A-rules to determine which set of consonants should appear in each allomorphic occurrence. If non-alternating consonants enter this rule nothing happens – a consonant from one set is "replaced" by the counterpart consonant from the other set (which is of course the same consonant). If, however, an alternating consonant is subjected to the rule, then it is replaced by its counterpart from the other set. We thus obtain an automatic account of spirantization processes in the rules.

Example 11 demonstrates this solution. The A-rule is designed to account for possible stop-fricative alternations in the 1st and 2nd positions, using MOR's ability to switch between variables with the X<Y syntax. The denotation of this operation is: replace the *n*-th member from Y in the LEXSURF with the *n*-th member from X in the ALLOSURF.

The base (or citation) form, which is the form of the lexical entry, is stated as the value of LEXSURF, while ALLOSURF represents one of the stem's allomorphs. Note that an alternating consonant in the first position would appear as a stop (variable P) in the base form (e.g., *katāv* "he wrote"), but as a fricative (variable F) in the inflected form (e.g., *yi+ḳtōv* "he will write"). Likewise, an alternating consonant in the second position would appear as a fricative (variable F) in the base form (e.g., *savāl* "he suffered") but as a stop (variable P) in the inflected form (e.g., *yi+sbōl* "he will suffer"). Crucially, only alternating consonants can alternate with these two variables, leaving all the other non-alternating consonants intact.

Example 11 Lexical spirantization (alternating segments are underlined)

Variables:

```
P = ʔ|ḇ|g|ʝ|d|h|w|z|ž|x|ṭ|y|ḵ|l|m|n|s|ʕ|p̱|c|ç|q|r|š|ṣ|t
F = ʔ|v̱|g|ʝ|d|h|w|z|ž|x|ṭ|y|ḳ̱|l|m|n|s|ʕ|f̱|c|ç|q|r|š|ṣ|t
Q/T/L = (all consonants)
```

A-rule excerpt:

`LEXSURF = $Pa$Fā$L`

(Fits base-form verbs, such as: *ḵatāv* “he wrote” / *sav̱āl* “he suffered”)

`ALLOSURF = $F<P$P<Fō$L`

(Fits most future tense inflections: *(yi+)ḳ̱tōv* “he will write” / *(yi+)sḇōl* “he will suffer”)

Our design of A-rules with spirantization processes also allows us to handle systematic variation effects within verbal paradigms. Hebrew speakers often do not follow the required sound change for a “proper” stop-fricative alternation. This may happen regularly at some specific consonantal positions within verbal paradigms (Adam, 2002). Certain allomorphs are therefore repeated twice such that the targeted position for alternation appears once with the required sound change, and again without change. Consider Example 11 above. The allomorph that the rule produces may change the first and the second consonants, if they belong to the spirantization triplet. The first consonant normally exhibits this change when possible, yet the second consonant tends to vary in that position, and often appears without change. Example 12 below extends the previous example with two different allomorphs. Importantly, although all the allomorphs of an A-rule are active at the same time, the rule does not create duplicate representations, i.e., the allomorph of verbs with no spirantizationable consonant in their second position will not be generated twice.

Example 12 Stop-fricative variation

`ALLOSURF = $F<P$P<Fō$L`

(2nd variable features a sound-change: *(yi+)sḇōl* “he will suffer”)

`ALLOSURF = $F<P$Fō$L`

(2nd variable features no sound-change: *(yi+)sv̱ōl* “he will suffer”)

Guttural consonants: {*x*/*ʔ*/*ʕ*/*ħ*} Another historical phonologically-motivated alternation concerns the class of gutturals, which, in many cases, trigger vowel lowering in their environment. In Modern Hebrew they do not form a natural class, since the pharyngeal consonants (denoted by the Hebrew letters ח {x} and ע {ʕ}) are not pronounced as pharyngeals (/ħ/ and /ʕ/, respectively), and the glottal consonants (denoted by the Hebrew letters א {ʔ} and ה {h}) are often not pronounced at all. As with the spirantization rule, alternations concerning guttural consonants are only fully manifested in Hebrew orthography (not in its phonology).

Note that while most of the letters in this group often denote a null consonant (they are not pronounced in rapid speech), the consonant denoted by the letter ח {x} is always pronounced (as a velar fricative that never participates in spirantization processes). Evidently, {x} also exhibits a different behavior, as it often does not trigger the expected vowel lowering around it, in positions where the other pseudo-gutturals always do. To account for this kind of variation, two classes of guttural consonants were defined: one with {x}, and another one without.

Consider Example 13: the {ʕ}-initial basic verb *ʕaṣā* "make" triggers the first of the two ordered A-rules, which predicts only one allomorph for that conjugation, with a vowel between the first two consonants of the stem. Having fired the first rule, MOR will not proceed to the following rule (which would potentially create a duplicate analysis). At the same time, the {x}-initial word *xacā* "cross" triggers only the second rule, which, for the same conjugation, predicts a variation (two allomorphs) in the appearance of a low vowel (/a/) between the first two consonants of the stem.

Example 13 (Pseudo-guttural distinction)

Variables:
G = x|ʔ|ʕ|h
H = ʔ|ʕ|h
A-rule excerpts:
1st. rule
LEXSURF = HaTā
(Fits base-form verbs, such as: *ʕ̲aṣā* "he made")
ALLOSURF = HaTē
(Fits most future tense inflections: *(ya+)ʕ̲aṣē* "he will make")
2nd. rule
LEXSURF = GaTā
(Fits base-form verbs, such as: *x̲acā* "he crossed")
ALLOSURF = GaTē
(Fits prescribed pronunciation of most future tense inflections: *(ya+)x̲acē* "he will cross")
ALLOSURF = GTē
(Fits standard pronunciation of most future tense inflections: *(ya+)x̲cē* "he will cross")

Sonorants: {y/l/m/n/r} Another class of phonemes that impacts its surroundings within the word it the class of sonorants. Hebrew speakers follow a phonotactic restriction whereby sonorants must be adjacent to vowels. Instead of producing consonant clusters (a phenomenon that is more tolerated in Modern Hebrew compared to, for example, Biblical Hebrew), a simpler syllabic structure is favored when the first member of a consonantal sequence is a sonorant. Thus, potential consonantal clusters that violate sonority-driven phonotactic restrictions in Hebrew may trigger morpho-phonemic changes which we can properly predict.

In order to allow the analyzer to recognize different syllabic structures as belonging to the same word pattern, a class of sonorants was defined. The A-rules are ordered such that the more specific version (the one targeting a sonorant) is the first, and the more general rule follows it. Lexical entries with a sonorant in relevant positions trigger the first A-rule and not the general rule. Other entries trigger only the general rule.

Consider Example 14: the sonorant-initial word *yēled* "kid" triggers the first of two A-rules, which predicts a vowel between the first two consonants in the plural inflection. The obstruent-initial word *kēlev* "dog" triggers only the second rule, which, for the same inflection, predicts a word-initial consonant cluster.

Example 14 (Sonority restriction)

Variables:
`S = y|l|m|n|r`
Rule excerpts:
1st. rule
`LEXSURF = $Sē$Te$L`
(Fits base-form nouns, such as: *yēled* "kid")
`ALLOSURF = $Se$Tā$L`
(Fits plural inflections: *yelad(+īm)* "kids")
2nd. rule
`LEXSURF = $Qē$Te$L`
(Fits base-form nouns, such as: *kēlev* "dog")
`ALLOSURF = $Q$Ta$L`
(Fits plural inflections: *klav(+īm)* "dogs")

Coronal stops: {`d/t/ṭ`} When Hebrew stems end with a coronal stop (/t/ or /d/) they may immediately precede the coronal stop /t/, which is the initial consonant of some verbal suffixes (*+ti*, *+ta*, *+tem*, *+ten*, etc.) In such cases, a phonotactic restriction requires a vowel insertion that would break the identical (or similar) sequence of stops. This is a standard scenario of an *Obligatory Contour Principle* (OCP) effect (Leben, 1973, 1978; McCarthy, 1986) in Hebrew. Again, to allow the morphological analyzer to recognize this complementary distribution, a class of coronal stops was defined. Lexical entries with a stem-final coronal stop trigger the A-rule that identifies that final stem consonant instead of the general rule that comes next.

In Example 15, the coronal-final stem *laḳād* "capture" triggers the first of two ordered A-rules, which predicts a vowel between that consonant and an immediately following suffix initial /t/. The non-coronal-final stem *dafāq* "knock", on the other hand, triggers only the 2nd (general) rule.

Example 15 (OCP effects)

Variables:
`D = d|t|ṭ`
Rule excerpts:
1st. rule
`LEXSURF = $Qa$Tā$D`
(Fits base-form verbs, such as: *laḳād* "he captured")
`ALLOSURF = $Qa$Tā$D`

(Fits 1st person inflection: *laḳād(+eti)* “I captured”)

2nd. rule

LEXSURF = $Qa$Tā$L

(Fits base-form verbs, such as: *dafāq* “he knocked”)

ALLOSURF = $Qa$T?$L

(Fits 1st person inflection: *dafāq(+ti)* “I captured”)

Stridents / Sibilants: {*s* / *š* / *ṣ* / *c* / *ç* / *z*} Strident (or sibilant) consonants trigger a phonological rule of metathesis in a certain position of the *hitpael* paradigmatic template—when it is in the first consonantal slot, immediately following the /t/ of the *hitpael* pattern (*hitCaCeC*). In these metathesis cases, the coronal consonant that is part of the stem’s pattern may not only switch places with the following stem consonant, but also change to one of the other coronal stops, {ṭ} or {d}. Again, these phenomena stem from phonological assimilation processes of Biblical Hebrew that are no longer fully active in Modern Hebrew. The analyzer is able to recognize these alternations as belonging to the same verb pattern within the A-rule itself (with no need for additional rules or for rule ordering) since such metathesis cases are distinct already at the level of LEXSURF (i.e., the surface form of the lexical entry).

Compare the LEXSURF values of the two excerpts in Example 16. The special form of basic verbs that fit the first rule is distinct from the form of verbs that fit the second (more general) rule.

Example 16 (Metathesis)

Variables:

C = s|š|ṣ|c|ç|z

Rule excerpts:

1st. rule

LEXSURF = hi$C$Da$Tē$L

(Fits base-form verbs, such as: *hictalēm* “he/it was photographed”)

2nd. rule

LEXSURF = hit$Qa$Tē$L

(Fits base-form verbs, such as: *hitpazēr* “he/it was scattered”)

4.3.5 Other Grammatical Categories

The discussion above focused on verbs and nouns, because these are the part-of-speech categories for which inflectional morphology is most productive. 335 A-rules were developed for nouns, and 572 for verbs. Yet other grammatical categories require similar solutions as well. We developed 47 rules for adjectives, 26 for prepositions, 6 for adverbs, 5 for pronouns, and 18 for forms of the verb *hayā* “be”. Rules are organized in separate files according to the POS category they apply to; the verbal rules are further divided into five files, one for each tense/form (see Section 4.3.2). Nominal rules are further divided into six files according to the phonological structure of the nouns they apply to.

To illustrate other grammatical categories, consider adjectives. Hebrew adjectives inflect for gender and number. The base form is the masculine singular form; the feminine suffix is lexically determined (and can be *ā*, *et*, *at*, or *īt*). The addition of a suffix can trigger segmental and/or prosodic changes in the stem.

As an example, consider the adjective *gadōl* "big", whose lexical entry is listed in Example 17. Subsequent to the specification of the POS category (`scat adj`), this entry lists the feminine suffix (`fem a`), followed by an indication of a vowel change (`vchange 1`). Example 17 further specifies the rule that generates the allomorphs of this adjective. The variables `$Q`, `$T` and `$L` are matched against the root consonants of the lexeme (here, *g.d.l*). The `LEXCAT` specification matches the features of the lexical entry, and hence this rule applies. Then, two allomorphs are generated: one for the base form, with no change in the surface form; and one for the feminine and plural suffix. In the second allomorph, note the change in the stress pattern, and the `$V<O` specification, indicating that the stressed vowel is replaced by its unstressed counterpart. This is due to the fact that the feminine and plural suffixes carry inherent stress. The resulting forms are *gadōl, gdolā, gdolīm, gdolōt.*

Example 17 (Adjectives)

Lexical entry: gadōl

```
  [scat adj][fem a][vchange 1][ptn qatol][root gdl]
```

Rule excerpts:

```
  LEXSURF = $Qa$T$O$L
  LEXCAT = [scat adj],[fem a],[vchange 1]

  ALLO:
   ALLOSURF = $Qa$T$O$L
   ALLOCAT = LEXCAT, ADD [gen ms], ADD [num sg], ADD [done yes]
  ALLO:
   ALLOSURF = $Q$T$V<O$L
   ALLOCAT = LEXCAT, ADD [femsfx a], ADD [plsfx reg]
```

Very similar rules were also developed for prepositions (which, in Hebrew, can combine with pronominal suffixes, in a very similar way to nouns), pronouns, and the few inflecting adverbs.

4.4 Results

The main result of our work is a properly and uniformly transcribed, morphologically annotated CHILDES corpus of Hebrew. As mentioned above, the corpus includes the Berman longitudinal corpus, with data from four children between the ages of 1;06 and 3;05, and the Ravid longitudinal corpus, with data from two siblings between the ages of 0;09 to around 6 years of age. Together, the corpora consist of 114,632 utterances comprising of 417,938 word-tokens (13,828 word-types).

The MOR lexicon includes approximately 5,200 entries, in 16 part-of-speech categories (and, additionally, 700 affixes).[5] Table 3 lists the number of lexical entries in the main POS categories.

Noun	2377
Verb	1157
Adjective	833
Adverb	369
Pronoun	89
Preposition	88
Other	283
Total	**5196**

Table 3 The number of lexical entries according to part-of-speech

Lexically-specified information includes root and pattern (for verbs mainly), gender (for nouns), plural suffix (for nouns), and other information that cannot be deduced from the form of the word. Over 1,000 A-rules describe various allomorphs of morphological paradigms, listing their morphological and morphosyntactic features, including number, gender, person, nominal status, tense, etc. Lexical entries then instantiate the paradigms described by the rules, thereby generating specific allomorphs. These, in turn, can combine with affixes via over 100 C-rules that govern the the possible combinations of allomorphs involved in affixation.

The corpora include over 400,000 word tokens (about 14,000 types). More than 27,000 different morphological analyses are produced for the tokens observed in the corpus; however, we estimate that the application of the morphological rules to our lexicon would result in hundreds of thousands of forms, so that the coverage of the MOR grammar is substantially wider. The grammar fully covers our current corpus. As noted, the corpora and the MOR grammar are freely available from the CHILDES repository. Figure 2 above depicts a small fragment of a morphologically-annotated corpus.

5 Morphological Disambiguation

As noted in Section 2, the level of ambiguity of our data is much lower than that of the standard Hebrew script, especially due to the vocalic information encoded in the transcription. However, while most orthographic and phonemic ambiguity is resolved, morphological ambiguity still remains, as the MOR grammar associates each surface form with *all* its possible analyses, independently of the context. Such ambiguity arises especially with items that could

[5] The categories are adjective, adverb, communicator, copula, existential, negation, numeral, onomatopoeia, preposition, pronoun, punctuation, quantifier, question, unknown, verb and vocalization.

belong to more than one lexical category (typically in cases of syntactic conversion or zero derivation as is common in English). In Hebrew, such cases involve mostly participial or present tense forms. Thus, the string *šomēr* (see Example 8 above) can stand for both the noun *guard* and the verb *guard*. In spoken data, this type of ambiguity extends even further, since words can function not only as nouns, verbs, or adjectives but also as adverbs and as communicators (e.g., *yōfi* "beauty/great!", *ṭov* "good/OK"). Such items are highly frequent in interaction in general and in child-parent interactions in particular. Another type of ambiguity that remains involves cases where two forms representing different morpho-lexical or grammatical categories share the same pronunciation and spelling. For example, in the future tense paradigm, 2nd person masculine singular and 3rd person feminine singular take the same prefixal and stem forms across all verb pattern conjugations (e.g., *telēḳ* means either *you-Masc-Sg will-go* or *she will-go*). Finally, some ambiguous word forms are the result of completely accidental processes, as in *ʔeqdāx*, which can mean either the noun *gun*, or the verb *I will drill*.

Such ambiguous entries require disambiguation before the data can be applied to analysis. Following the application of MOR, each ambiguous analysis appears with the caret separating the possible analyses, as in Example 18.

Example 18 (Ambiguity)

```
n | ?eqdāx & gen:m & num:sg & stat:unsp =gun
^
?e# v | qadāx & root:qdx & ptn:qal & tense:fut & pers:1 &
gen:unsp & num:sg =drill
```

In CLAN, a dedicated mode is available for manual disambiguation within context, a time consuming process that involves many on-line decisions, some of which are easily handled, and others that result in inconsistencies. CLAN also provides users with an automatic module, POST, that can be trained on a given corpus and used as a part-of-speech tagger on unseen data. In order to train POST, a set of files with unique analyses for every entry is required. As a first step, we developed a set of guidelines for manual morphological disambiguation. This allowed us to make the decisions regarding unclear cases more internally consistent. Following are some of the most frequent guidelines that were used for our data:

– For any word that is ambiguous with a communicator analysis (`co`), that analysis wins whenever that word is not modifying another word, or being modified by one. This is usually the case with many single word utterances, such as *ken* "honest/yes" and *naḳōn* "correct" (also applicable when such words are utterance-initial or utterance-final).
– When a word can be analyzed as either a noun (`n`) or an adjective (`adj`) (e.g., *xaḳām* "sage/smart", *yafā* "pretty/beauty"), it is a noun if it does not modify another word in that utterance.
– When some word is ambiguous between adjectival and adverbial (`adv`) analyses, it is adjectival if it is modifying a noun. Otherwise, it is adverbial.

- When a word is ambiguous between a quantifier (`qn`) and adverbial analyses, it is a quantifier if it is modifying a noun. Otherwise, it is adverbial.
- When the participle form of verbs (`part`) is ambiguous with an adjective, participle is chosen if the given structure can be inflected for tense and retain the same basic meaning.
- When the participle form of verbs is ambiguous with another form of that verb (typically, past), and the context is not enough to determine the suitable analysis, the participle analysis wins.

Following these guidelines, we manually disambiguated 18 of the 304 files in the corpus. This was done by two lexicographers, and all disagreements were consolidated by a third annotator.[6] We used 14 of the manually-disambiguated files to train the part-of-speech tagger with tools that are embedded in CLAN (*POSTRAIN* and *POST*). We then automatically disambiguated the remaining files. The results of this endeavor are a single morphological analysis for each token in the corpus. The accuracy of the morphological disambiguation is evaluated in the next section.

6 Evaluation

As noted above, the MOR grammar fully covers the forms in our corpus. To further evaluate its coverage, we applied it to a new corpus that is currently being transcribed. Of the 10,070 tokens in the new corpus, 176 (1.75%) do not obtain an analysis (77 of the 1431 types, or 5.3%). This provides a rough estimate as to the coverage of the lexicon and the MOR grammar. Of course, there are no guarantees that the provided analyses are indeed correct; but manual inspection of the first 1,000 tokens in the new corpus reveals that over 90% of them are indeed valid. The missing analyses can be attributed mostly to missing lexical entries and inconsistent transcription (this is the case with 66 of the 77 unanalyzed entries). The remaining 11 types are missing in MOR's *lexicon.* As more Hebrew corpora are being transcribed, we are certain that lexical gaps and any remaining MOR omissions and inaccuracies will soon be amended.

As another evaluation method, we developed a program that converts the transcription we use to the standard Hebrew script. While our transcription is rich, and includes both consonantal and vocalic information (some of which, recall, is missing in the standard Hebrew script), such a conversion is not trivial, and in some cases cannot be done deterministically. Yet it generates the correct form in the vast majority of cases. We then submit the Hebrew forms to the MILA morphological analyzer (Itai and Wintner, 2008), a state-of-the-art morphological tool for (standard, written) Hebrew, and compare the results of the two analyzers. The MILA analyzer operates on non-vocalized

[6] We did not measure inter-coder agreement, but we estimate that more than 90% of the ambiguous tokens were identically annotated by both lexicographers. Consolidating the differences was a quick and easy task.

forms, and does not perform disambiguation; we therefore only check that the MOR analysis is *included* in the set of analyses produced by MILA.

For various reasons, the conversion program only works on 340,212 word tokens (for example, child neologisms, fillers, onomatopoeia etc. are excluded). Of those, 38,481 mismatches with MILA (11.3%) are reported. Again, we manually inspected the mismatches in one file (2,023 tokens, 214 mismatched tokens but only 85 mismatched types). While a few mismatches (5 out of 85) indeed indicate errors in the MOR grammar, or discrepancies between the two analyzers (8 due to incompatible part-of-speech assignment, and 10 due to incompatible morphological features assignment), most of the 214 mismatches are attributable to problems with the MILA analyzer or the conversion and comparison script.

Furthermore, 27 of the 85 mismatches are due to forms unique to the CHILDES corpus (e.g., the verb *lefaxēd* "fear", whose standard form is *lifxōd*); and 6 are due to errors in the transcription (e.g., the proper name *ʔēitan* is transcribed *Eitan*, without the initial consonant/letter, and is therefore wrongly converted to the standard script). 8 more mismatches are due to MOR-specific conventions: 5 mismatches are due to MOR's analysis of multi-lexemic expressions (see section 2.3); and 3 mismatches are due to MOR's unique *completion* part-of-speech category.

Finally, to evaluate the accuracy of the part-of-speech tagger (more precisely, the morphological disambiguation module), we trained the tagger on the fourteen manually annotated files (the training corpus), and used it on the other four manually disambiguated files (the test corpus). The test corpus contains 8,871 tokens, of which 1,631 are ambiguous. Of those, the wrong analysis was selected in 306 of the tokens, or 19%, bringing the overall accuracy of the tagger to $1 - 306/8871 = 96.6\%$. Note that we do not distinguish between child- and adult-speech in this analysis; presumably, the adult utterances produce less error. We intend to manually disambiguate the entire corpus in the future, and use it in its entirety as a training corpus.

7 Discussion

We described a properly and uniformly transcribed, morphologically-annotated CHILDES corpus of Hebrew. We believe that the corpus will be instrumental for future investigations of Hebrew child- and child-directed language, language development and psycholinguistics in general. The corpora are all freely available from the main CHILDES repository.

Furthermore, we described a morphological analyzer specifically developed for this corpus. The analyzer, too, is freely distributed from the CHILDES web site; it is already used to process more corpora that are currently being transcribed by several research groups working on Hebrew child language.

For example, the data described in the current paper are used in an ongoing study of the acquisition of noun plurals, examining both the distributions of plural morphemes and the correspondence between child and adult output. Another ongoing project examining the relationship between adult input and child output is focusing on the Hebrew verbal system, with new densely recorded corpora that were transcribed and analyzed following the guidelines specified here. Yet another study involves a developmental analysis of the lexicon of Russian-Hebrew sequential bilingual children, where data were re-transcribed to allow for automatic analysis of lexical categories. And another ongoing study, which examines the development of complex syntax in children's narratives and dyadic conversations and interviews, relies heavily on the morphological analysis provided by MOR in order to locate conjunctions and to trace complex verb phrases.

Our ultimate plan is to add syntactic annotation to the transcripts. We have devised a syntactic annotation scheme, akin to the existing scheme used for the English section of CHILDES (Sagae et al., 2007, 2010), but with special consideration for Hebrew constructions that are common in the corpora. We have recently begun to annotate the corpora according to this scheme.

Acknowledgments

This research was supported by Grant No. 2007241 from the United States-Israel Binational Science Foundation (BSF). We are grateful to Hadass Zaidenberg, Maayan Bloch and Ezer Rasin for their meticulous lexicographic work, to Arnon Lazerson for developing the conversion script, and to Shai Gretz for helping with the manual annotation.

References

Galit Adam. *From variable to optimal grammar: evidence from language acquisition and language change.* PhD thesis, Tel Aviv University, 2002.

Aviad Albert, Bracha Nir, Brian MacWhinney, and Shuly Wintner. A morphologically-analyzed CHILDES corpus of Hebrew. Presented at The International Association of the Study of Child Language (IASCL), July 2011.

Aviad Albert, Bracha Nir, Brian MacWhinney, and Shuly Wintner. A morphologically annotated Hebrew CHILDES corpus. In *Proceedings of the EACL-2012 Workshop on Computational Models of Language Acquisition and Loss*, April 2012.

Colin Bannard, Elena Lieven, and Michael Tomasello. Early grammatical development is piecemeal and lexically specific. *Proceedings of the National Academy of Science*, 106(41):17284–17289, October 2009.

Outi Bat-El. Stem modification and cluster transfer in Modern Hebrew. *Natural Language and Linguistic Theory*, 12:571–593, 1994.

Ruth A. Berman. Lexical decomposition and lexical unity in the expression of derived verbal categories in modern Hebrew. *Afroasiatic Linguistics*, 6: 1–26, 1979.

Ruth A. Berman. Language development and language knowledge: Evidence from the acquisition of Hebrew morphophonology. *Journal of Child Language*, 8:609–626, 1981.

Ruth A. Berman. The acquisition of Hebrew. In Dan I. Slobin, editor, *The crosslinguistic study of language acquisition*, pages 255–372. Lawrence Erlbaum Associates, Hillsdale, NJ, 1985.

Ruth A. Berman. Children's acquisition of compound constructions. In Rochelle Lieber and Pavol Stekauer, editors, *The Oxford Handbook of Compounding*. Oxford University Press, 2009.

Ruth A. Berman and Dorit Ravid. Lexicalization of noun compounds. *Hebrew Linguistics*, 24:5–22, 1986. In Hebrew.

Ruth A. Berman and Jürgen Weissenborn. Acquisition of word order: A crosslinguistic study. Final Report. German-Israel Foundation for Research and Development (GIF), 1991.

Gideon Borensztajn, Willem Zuidema, and Rens Bod. Children's grammars grow more abstract with age — evidence from an automatic procedure for identifying the productive units of language. *Topics in Cognitive Science*, 1:175–188, 2009.

Hagit Borer. On the morphological parallelism between compounds and constructs. In Geert Booij and Jaap van Marle, editors, *Yearbook of Morphology 1*, pages 45–65. Foris publications, Dordrecht, Holland, 1988.

Hagit Borer. The construct in review. In Jacqueline Lecarme, Jean Lowenstamm, and Ur Shlonsky, editors, *Studies in Afroasiatic Grammar*, pages 30–61. Holland Academic Graphics, The Hague, 1996.

Roger Brown. *A first language: the Early stages*. Harvard University Press, Cambridge, Massachusetts, 1973.

Eve V. Clark and Ruth A. Berman. Types of linguistic knowledge: interpreting and producing compound nouns. *Journal of Child Language*, 14(03):547–567, 1987. doi: 10.1017/S030500090001028X. URL `http://dx.doi.org/10.1017/S030500090001028X`.

David Crystal, Paul J. Fletcher, and Michael. Garman. *The grammatical analysis of language disability : a procedure for assessment and remediation*. Edward Arnold, London, 1976. ISBN 0713158425.

Daniel Freudenthal, Julian Pine, and Fernand Gobet. Explaining quantitative variation in the rate of optional infinitive errors across languages: a comparison of mosaic and the variational learning model. *Journal of Child Language*, 37(3):643–69, 2010. ISSN 1469-7602. URL `http://www.biomedsearch.com/nih/Explaining-quantitative-variation-in-rate/20334719.html`.

Roland R. Hausser. Principles of computational morphology. Technical report, Center for Machine Translation, Carnegie Mellon University, 1989.

Alon Itai and Shuly Wintner. Language resources for Hebrew. *Language Resources and Evaluation*, 42(1):75–98, March 2008.

William Roland Leben. *Suprasegmental phonology*. PhD thesis, Massachussettes Institute of Technology, 1973.

William Roland Leben. The representation of tone. In Victoria Fromkin, editor, *Tone: A Linguistic Survey*, pages 177–220. Academic, New York, 1978.

Laura Louise Lee. *Developmental sentence analysis.* Northwestern University Press, Evanston, IL, 1974.

Brian MacWhinney. The CHILDES system. *American Journal of Speech Language Pathology*, 5:5–14, 1996.

Brian MacWhinney. *The CHILDES Project: Tools for Analyzing Talk.* Lawrence Erlbaum Associates, Mahwah, NJ, third edition, 2000.

Brian MacWhinney. Enriching CHILDES for morphosyntactic analysis. In Heike Behrens, editor, *Corpora in Language Acquisition Research: History, methods, perspectives*, volume 6 of *Trends in Language Acquisition Research.* Benjamins, Amsterdam, 2008.

John J. McCarthy. OCP effects: Gemination and antigemination. *Linguistic Inquiry*, 17:207–263, 1986.

J. Miller and R. Chapman. *SALT: Systematic Analysis of Language Transcripts, User's Manual.* University of Wisconsin Press, Madison, WI, 1983.

Susanne Miyata and Brian MacWhinney. The development of parallel language measures: The example of Japanese DSSJ. Presented at The International Association of the Study of Child Language (IASCL), July 2011.

Susanne Miyata, Makiko Hirakawa, Keiko Itoh, Brian MacWhinney, Yuriko Oshima-Takane, Kiyoshi Otomo, Yasuhiro Shirai, Hidetosi Sirai, and Masatoshi Sugiura. Constructing a new language measure for Japanese: Developmental sentence scoring for Japanese. In Susanne Miyata, editor, *Development of a developmental index of Japanese and its application to speech developmental disorders. Report of the Grant-in-Aid for Scientific Research (B)(2006-2008) No. 18330141*, pages 15–66. Aichi Shukutoku University, Nagoya, Japan, 2009.

Bracha Nir and Ruth A. Berman. Parts of speech as constructions: the case of Hebrew 'adverbs'. *Constructions and Frames*, 2(2):242–274, 2010.

Bracha Nir, Brian MacWhinney, and Shuly Wintner. A morphologically-analyzed CHILDES corpus of Hebrew. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 1487–1490. European Language Resources Association (ELRA), May 2010. ISBN 2-9517408-6-7.

Noam Ordan and Shuly Wintner. Representing natural gender in multilingual lexical databases. *International Journal of Lexicography*, 18(3):357–370, September 2005.

Uzzi Ornan. Phonemic script: A central vehicle for processing natural language – the case of Hebrew. Technical Report 88.181, IBM Research Center, Haifa, Israel, 1986.

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

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

Dorit Ravid. *Spelling Morphology: The Psycholinguistics of Hebrew Spelling.* Springer, 2012.

Dorit Ravid, Wolfgang U. Dressler, Bracha Nir-Sagiv, Katharina Korecky-Kröll, Agnita Souman, Katja Rehfeldt, Sabine Laaha, Johannes Bertl, Hans Basbøll, and Steven Gillis. Core morphology in child directed speech: Crosslinguistic corpus analyses of noun plurals. In Heike Behrens, editor, *Corpora in language acquisition research: finding structure in data*, pages 25–60. John Benjamins, 2008.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, pages 1–15, Mexico City, Mexico, 2002.

Kenji Sagae, Brian MacWhinney, and Alon Lavie. Automatic parsing of parent-child interactions. *Behavior Research Methods, Instruments, and Computers*, 36:113–126, 2004.

Kenji Sagae, Eric Davis, Alon Lavie, Brian MacWhinney, and Shuly Wintner. High-accuracy annotation and parsing of CHILDES transcripts. In *Proceedings of the ACL-2007 Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 25–32, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W07/W07-0604`.

Kenji Sagae, Eric Davis, Alon Lavie, Brian MacWhinney, and Shuly Wintner. Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language*, 37(3):705–729, 2010. doi: 10.1017/S0305000909990407. URL `http://journals.cambridge.org/article_S0305000909990407`.

Hollis S. Scarborough. Index of productive syntax. *Applied Psycholinguistics*, 11:1–22, 1990.

Joseph Shimron, editor. *Language Processing and Acquisition in Languages of Semitic, Root-Based, Morphology.* Number 28 in Language Acquisition and Language Disorders. John Benjamins, 2003.

Dan I. Slobin. *The Crosslinguistic Study of Language Acquisition: The data.* The Crosslinguistic Study of Language Acquisition. Lawrence Erlbaum Associates, Hillsdale, NJ, 1985. ISBN 9780898593679.

Adam Ussishkin. The inadequacy of the consonantal root: Modern Hebrew denominal verbs and output–output correspondence. *Phonology*, 16(03): 401–442, 1999.

Heidi R. Waterfall, Ben Sandbank, Luca Onnis, and Shimon Edelman. An empirical generative framework for computational modeling of language acquisition. *Journal of Child Language*, 37(3):671–703, 2010.

Shuly Wintner. Hebrew computational linguistics: Past and future. *Artificial Intelligence Review*, 21(2):113–138, 2004. ISSN 0269-2821. doi: http://dx.doi.org/10.1023/B:AIRE.0000020865.73561.bc.

Shlomo Yona and Shuly Wintner. A finite-state morphological grammar of Hebrew. *Natural Language Engineering*, 14(2):173–190, April 2008.