A Morphologically-Analyzed CHILDES Corpus of Hebrew

Bracha Nir, Brian MacWhinney, Shuly Wintner

Department of Communication Sciences, Department of Psychology, Department of Computer Science
University of Haifa, Carnegie Mellon University, University of Haifa
bnir@univ.haifa.ac.il, macw@cmu.edu, shuly@cs.haifa.ac.il

Abstract

We present 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). We introduce a dedicated transcription scheme for the spoken Hebrew data that is aware both of the phonology and of the standard orthography of the language. We also introduce a morphological analyzer that was specifically developed for this corpus.

1. Introduction

Recent years have witnessed the gradual proliferation of computerized tools for processing natural languages with complex morphology. These tools serve language researchers by providing them with an automatic interface that enables quick and accurate analyses of corpora in different sizes. 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 free 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 a large variety of levels of analysis, while still permitting a barebones form of transcription; (2) a set of computer programs called CLAN (Computerized Language ANalysis), 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, learners of second languages, and bilinguals who have been exposed to language in early childhood. Researchers can directly test a vast range of empirical hypotheses against data from nearly a hundred major research projects.

Although about half of the CHILDES corpus consists of English data, there is also a significant component of transcripts in over 25 languages. The present paper focuses on the Hebrew section, consisting of two sets of files: 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 09;0 to around 6 years of age. The corpora currently consist of 110,819 utterances comprising of 425,471 word-tokens (19,224 word-types).

2. Transcription of spoken Hebrew Data

Data files in the CHILDES system are transcribed according to CHAT format. Since these files are all written renditions of speech samples, it makes little sense to transcribe them according to the orthographic conventions of the language. The decision to write out stretches of vocal material using the forms of written language can trigger a variety of theoretical commitments (MacWhinney, 2000). Moreover, this decision can become a source for complexity in later stages, when the data are subject to computerized analysis. Standard Hebrew orthography is inherently ambiguous and involves numerous homographs (Ornan and Katz, 1995; Wintner, 2004). Any 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). Instead, and in contrast to most existing Hebrew language resources, the Hebrew data in the CHILDES system rely on Latin-based phonemic transcription.

The Hebrew data were collected by different researchers and transcribed by different people for various research purposes. Consequently, these files were highly lacking in terms of the consistency of the transcription methods that were originally used, both within and across corpora. The first major challenge of working with these data was standardizing transcription. To this end, the bulk of the data, consisting of over 350,000 tokens of Child Directed Speech, has been semi-automatically re-transcribed to conform to a newly devised set of CHAT-compatible transcription conventions.

The standard Hebrew orthography leaves most of the vowels unspecified. On top of that, the script dictates that many particles, including four of the most frequent prepositions, the definite article, the coordinating conjunction and some subordinating conjunctions, all attach to the words which immediately follow them. Existing transcription approaches either use a phonemic transcription (Ornan, 1986; Ornan, 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 both on phonemic and phonological features while also taking orthography into account. Unlike the standard Hebrew script, our transcription reflects all vowels. In addition, it reflects consonant distinctions that are present in the standard orthography but not in phonology. This facilitates resolution of homophonic ambiguity, as in כר *kar* ‘pillow’ vs. קר *qar* ‘cold’; or אח *ʔax* ‘brother’ vs. אך *ʔaḳ* ‘however’. Moreover, our transcription explicitly provides the stress patterns for each word type, thus resolving phonological ambiguities such as bíra 'beer' vs. birá 'capital (city)', or racá 'wanted' vs. ráca 'ran'. Since CHAT conventions do not allow the use of special characters (e.g., $, &, #) for representing consonants, a set of monoglyph Unicode IPA characters that has already been applied for other complex scripts was used. Table 1 presents the complete set of transcription pairs.

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	’	p/f	c	q	r	š/ṣ	t	j̣	ž	ç	

Vowels (stressed and unstressed):

á é í ó ú a e i o u

Table 1: Transcription Pairs for Hebrew in CHAT Format

Figure 1 shows a brief example of an interaction between a Hebrew-speaking child and her caretakers transcribed in CHAT.

Our transcription thus conforms to the three major goals which the CHAT format is designed to achieve (MacWhinney, 1996): systematicity and clarity, human and computerized readability, and ease of data entry.

3. Morphological Analysis

An important feature of the Hebrew CHILDES corpus which is the direct result of our transcription method is that it allows for an overt representation of the morphological structure of Hebrew. Hebrew is a Semitic language that has a rich and complex morphology which relies not only on linear affixation but to a large extent on synthetic root-and-pattern combinations (Berman, 1978). Since both the consonants and the (stressed and unstressed) vowels of each pattern and affix are represented in our transcription, these patterns can easily be identified in the transcripts, again resolving a-priori many homophonic and homographic ambiguities.

We developed a morphological analyzer[1] for the Hebrew transcripts, in compliance with the general architecture of previously compiled MORphological grammars (MacWhinney, 2000) for eight other languages. The Hebrew analyzer (HebMor) consists of three major components: an **a**(llophone)-rules file, a **c**(oncatenation)-rules file and a set of lexicon files.[2] The *a-rules* file specifies the various forms a root or stem can take, whether in terms of inflectional or derivational morphology. The *c-rules* file consists of higher level rules, allowing concatenation of the different categories. The *lexicon* is organized by grammatical category (e.g., adverbs, function words, nouns, pronouns and verbs) and divided into several files that contain roots, stems, or whole words. The current lexicon includes over 7,000 entries, distributed across the different parts of speech as listed in Table 2. Of course, creating a lexicon is an open-ended task; we focused on adding the entries that occur in the corpora we had, and will continue to extend the lexicon as needed, given more corpora.

POS	Number of entries
Noun	3,700
Verb	1,200
Adjective	800
Adverb	370
Preposition	100
Pronoun	80

Table 2: Number of lexicon entries per part-of-speech

Figure 2 illustrates a fragment of an a-rule that is used for analyzing forms in the *hitpael* verb pattern. The operation of this rule is dependent on the specification of consonantal roots in the lexicon. For example, consider the lexical specification of the root *g.b.r*, whose category is regular verb (`rv`):

```
gbr4 {[scat rv][gen gbr]} =overcome=
```

Applying the rule to this root, the analyzer associates the following analysis with the string *hitgabárti* 'I overcame':

```
v|P4:1SG:US:PAST&GBR=overcome
```

This analysis should be read as follows: `V` is the main category (verb), and `P4` indicates the fourth (out of seven) verb pattern (*hitpael*). Then, `1SG` indicates first person singular, `US` an unspecified gender, `PAST` is the tense and `GBR` is the root. Analyses are associated with an English gloss, following the = sign.

Each item analyzed by HebMor is specified for a major part of speech (POS) category (as well as specific subcategories, such as proper noun, modal verbs, etc.) and for all standard features that are consistent with traditional descriptions of Hebrew: gender, number, and template (if relevant) for nouns and adjectives, and gender, number, person, tense, template, and root for all verbs.

Not only open-class but also other categories are specified for morpho-syntactic features which can be used by parsers and other applications. For example, the lexicon also specifies the person, number and gender of closed-class items such as pronouns, the gender of proper names, etc. Consider the inflected preposition *el-ejha* 'to-her, to-it+FEM', which is analyzed as:

```
preppro|el:3SG:FEM=to&her
```

Here, the main category is `preppro`, indicating a combination of a preposition (`el` 'to') with a bound pronoun. The remaining features indicate that this pronoun is third person singular (`3SG`) feminine (`FEM`). Again, an English gloss is provided.

As these examples show, the morphological analysis is adapted to Hebrew-specific grammatical categories, which apply across different lexical classes. Finally, since a

[1] The morphological analyzer presented here is based on the morpho-lexical analyzer originally developed for the CHILDES system by Bracha Nir and Sigal Uziel-Karl.

[2] For a detailed explanation of the syntax of a-rule and c-rule files, see MacWhinney (2000).

```
@Begin
@Languages:  he
@Participants:  CHI Hagar Child, MOT Inbal Mother, GRA Grandmother
GRA: Hagár, ʔat xolá .
GRA: ʔat yodá’at še- ʔat xolá Hagári ?
CHI: ava [:  aval] [*] le- gag .
CHI: gag .
GRA: mi ze ?
CHI: ladow le- gag .
CHI: le e gag .
MOT: ʔímaʔ lo holéḳet la- gag .
CHI: gag gag !
MOT: Hagári .
GRA: ʔat rocá sipúr ?
GRA: bóʔi tavíʔi li sipúr we- ʔaní ʔasapér lak .
CHI: le- gag !
MOT: loʔ meṣaxqím ’aḳšáyw ba- gag .
@End
```

Figure 1: Example of the transcription

```
RULENAME: v-hitpael
LEX-ENTRY:
LEXSURF = $Q$T$L4
LEXCAT = [scat rv]

%past
ALLO:
ALLOSURF = hit$Qa$Ta$Lti
ALLOCAT = LEXCAT, DEL [scat],
  ADD [scat v], ADD [ptn p4],
  ADD [prsn 1], ADD [num sg],
  ADD [gen us], ADD [tense past]
ALLOSTEM = P4:1SG:US:PAST
```

Figure 2: Example of an a-rule

phonemic transcription of Hebrew is highly transparent in comparison to an orthographic representation, a small set of rules is required for producing a relatively extensive number of outputs (above 96,000 items are analyzed by the current version of the analyzer). Currently, 95.4% of all adult word tokens and 61% of all child tokens are analyzed (the lower rate of recognizing child tokens is for the most part due to the fact that these utterances still do not all conform to the current transcription standard). Out of these analyzed tokens, 29.4% of all adult forms and 15.4% of all child forms are still ambiguous. These figures compare favorably with results pertaining to the state-of-the-art morphological analyzer of (adult, written, newspaper-style) Hebrew: Itai and Wintner (2008) report that their analyzer produced the correct analysis for 93.8% of the tokens in their evaluation corpus; and that the average number of analyses per word form is 2.64. Undoubtedly, our good results must be attributed both to the fact that our forms are vocalized and to the special focus on a particular corpus.

4. Future Plans

Several issues still need to be resolved with respect to our CHILDES-based Hebrew corpus. First, only few transcribed corpora of Hebrew have so far been adapted to the new transcription method described here, and so transcription is still inconsistent and much manual post-editing will be required. Moreover, the re-transcribed data are largely Child Directed Speech samples, and re-transcription of all Child Speech data (approximately 157,000 tokens) is still required. These need careful attention since many of the word types are truncated or invented forms and must be re-transcribed while relying on context.

Second, although the new transcription method significantly decreases the level of ambiguity, it does not eliminate all ambiguous cases, which should be resolved during the stage of morphological analysis. One way to handle this is to train a part-of-speech (POS) tagger using automatic tools provided by CLAN. Such a tagger will be able to automatically select the most suitable analysis for every multiple-choice output for a given lexeme. We are currently preparing training material for this task, and will use the POST program that is part of CLAN since it has worked to markedly reduce ambiguity for languages such as English, Spanish, and Chinese.

The output of the POS tagger will serve as a basis for developing a statistical parser for Hebrew CHILDES data of the kind that was developed for the English section of CHILDES (Sagae et al., 2007; Sagae et al., To Appear). We are currently developing a Dependency Grammar-based annotation schema for Hebrew grammatical relations. Once this schema is operationalized, we will address outstanding issues in usage-based analyses of language acquisition data in light of current formal grammars.

Finally, our morphological analyzer still requires evaluation. Since the amount of data is quite extensive, manual evaluation is labor-intensive and time consuming. However, our transcription method allows for a simple and straightforward conversion between Latin-based and Hebrew vocalized script. Therefore, once the data are converted, it should be possible to use other available morphological analyzers for Hebrew (Itai and Wintner, 2008) and compare analyses in order to determine the accuracy of our system.

Acknowledgments

This research was supported by a grant from the United States-Israel Binational Science Foundation (BSF), Jerusalem, Israel. We are grateful to Aviad Albert and Hadas Zaidenberg for their help with the annotation. The morphological analyzer was originally created by Bracha Nir and Sigal Uziel-Karl.

5. References

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

Ruth A. Berman. 1978. *Modern Hebrew Structure*. University Publishing Projects, Tel Aviv.

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

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

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

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

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

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

Kenji Sagae, Eric Davis, Alon Lavie, Brian MacWhinney, and Shuly Wintner. 2007. 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. Association for Computational Linguistics.

Kenji Sagae, Eric Davis, Alon Lavie, Brian MacWhinney, and Shuly Wintner. To Appear. Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language*.

Shuly Wintner. 2004. Hebrew computational linguistics: Past and future. *Artificial Intelligence Review*, 21(2):113–138.

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