A Morphologically Annotated Hebrew CHILDES Corpus

Aviad Albert
Linguistics
Tel Aviv Uni.
Israel

Brian MacWhinney
Psychology
Carnegie Mellon Uni.
USA

Bracha Nir
Communication Sciences
Uni. of Haifa
Israel

Shuly Wintner
Computer Science
Uni. of Haifa
Israel

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.

CHILDES 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 (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. *CHAT*, a system for discourse notation and coding, designed to accommodate a large variety of analyses, while still permitting a barebones form of transcription; 2. *CLAN*, a set of computer programs; 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. While about half of the CHILDES corpus consists of English data, there is also a significant body of transcripts in 25 other languages.

Corpus We focus on the Hebrew section of CHILDES, consisting of two corpora: 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 110,819 utterances comprising of 417,938 word-tokens (13,828 word-types).

Transcription The Hebrew data are transcribed with a Latin-based phonemic transcription (Nir et al., 2010). We use a set of monoglyph Unicode characters (mostly in line with standard IPA conventions) that has already been applied for other complex scripts. In contrast to previous transcription methods, the current transcription reflects phonemic, orthographic and prosodic features. The advantages of our approach in reducing ambiguity are:

- Unlike the standard script, our phonemic transcriptions includes the five vowels of Modern Hebrew, and prosodic information on primary stress location, thereby yielding fewer ambiguities that stem from homographs.
- At the same time, we retain valuable phonemic and phonetic distinctions that are standard in the orthography but are no longer distinct in Modern Hebrew speech (e.g., *t*/*ṭ*, *k*/*q*, *ʔ*/*ʕ*).
- We separate and mark prefix particles, making it easier to recognize them as separate morphemes, which never participate in homographs.

Proceedings of the EACL 2012 Workshop on Computational Models of Language Acquisition and Loss, pages 20–22,
Avignon, France, April 24 2012. ©2012 Association for Computational Linguistics

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.

Morphological Analysis CLAN includes a language for expressing morphological grammars, implemented as a system, *MOR*, for the construction of morphological analyzers. A MOR grammar consists of three components: a set of *lexicons* 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 Hebrew MOR extensively uses all of them in order to account for vocalic and consonantal changes of the stem allomorphs (handled within the A-Rules), and the proper affixation possibilities (via the C-rules and affix lists).

The *lexicon* includes over 5,800 entries, in 16 part-of-speech (POS) categories. 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 morphological alternations involved in affixation.

Results and Evaluation 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. Figure 1 depicts a small fragment of a morphologically-annotated corpus.

To evaluate the coverage of the grammar, we applied it to a new corpus that is currently being transcribed. Of the 10,070 tokens in this corpus, 176 (1.75%) do not obtain an analysis (77 of the 1431 types, 5.3%). While some analyses may be wrong, we believe that most of them are valid, and that the gaps can be attributed mostly to missing lexical entries and inconsistent transcription.

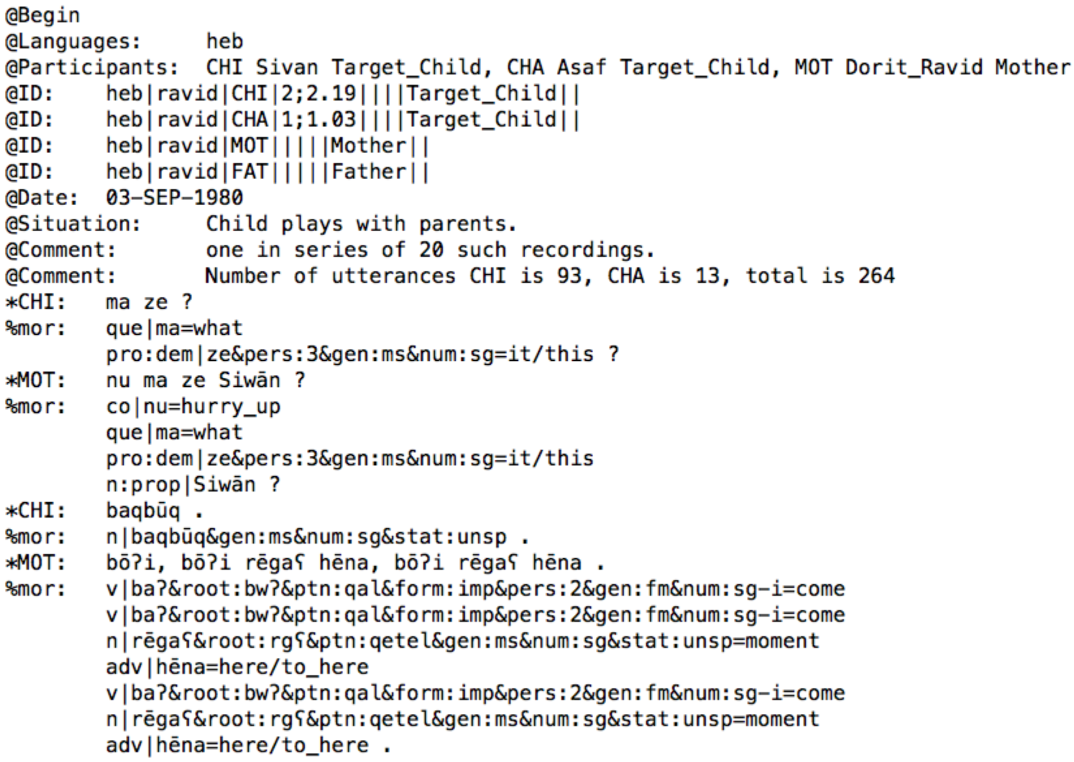
As another evaluation method, we developed a program that converts the transcription we use to the standard Hebrew script. We then submit the Hebrew forms to the MILA morphological analyzer (Itai and Wintner, 2008), and compare the results. The mismatch rate is 11%. While few mismatches indeed indicate errors in the MOR grammar, many are attributable to problems with the MILA analyzer or the conversion and comparison script.

Morphological Disambiguation The MOR grammar associates each surface form with *all* its possible analyses, independently of the context. This results in morphological ambiguity. The level of ambiguity is much lower than that of the standard Hebrew script, especially due to the vocalic information encoded in the transcription, but several forms are still ambiguous. These include frequent words that can function both as nouns, adjectives or adverbs and as communicators (e.g., *yōfi* “beauty/great!”, *ṭov* “good/OK”); verbs whose tense is ambiguous (e.g., *baʔ* “come” can be either present or past); etc.

We manually disambiguated 18 of the 304 files in the corpus, and used them to train a POS tagger with tools that are embedded in CLAN (*POS-TRAIN* and *POST*). We then automatically disambiguated the remaining files. Preliminary evaluation shows 80% accuracy on ambiguous tokens.

Future Plans 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., 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 Arnon Lazerson for developing the conversion script, and to Shai Gretz for helping with the manual annotation.



```
@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 .
```

Figure 1: A fragment of the annotated corpus

References

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

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

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.

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, Valletta, Malta, May 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7.

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`.