January, 2006

AKNOWLEDGMENTS

The transcription conventions for Hebrew are contributed by the Berman Lab, Department of Linguistics, Tel-Aviv University. Funding for transcription was provided by grants to Ruth Berman from the following agencies and sources:

1988 – 1991:
- The German-Israel Binational Science Foundation (GIF) and the Deutsche Forschungsgemeinschaft (DFG) [with Jürgen Weissenborn, Max-Planck Institute for Psycholinguistics, Nijmegen on the crosslinguistic study of early language acquisition in French, German, and Hebrew]

1994:
- Tel Aviv University "Keshet" (upgrading of infrastructure), grant for Hebrew child language data-base coding and analysis

1997 – 2000:
- Spencer Foundation, Chicago, Illinois, Major Research Grant, project on “Developing literacy in different contexts and in different languages”

2004 – 2007:
- German-Israel Foundation for Research and Development [GIF] Grant for study of advanced learner language: Cross-linguistic perspectives, with Christiane von Stutterheim, University of Heidelberg

Additional funding:
- Brian MacWhinney, director of the CHILDES Laboratory at Carnegie Mellon University
- Wolfgang Klein, director of the Language Acquisition section at the Max-Planck Institute for Psycholinguistics

The conventions documented below were prepared by Bracha Nir-Sagiv. They were originally devised by Ruth Berman, with contributions from Sharon Armon-Lotem (1996), Nurit Assayag and Dalia Cahana-Amitay (1998), and Bracha Nir-Sagiv (1998-2006). Most current adaptation is the result of intensive cooperation between the Berman team at Tel Aviv University and the von Stutterheim team at Heidelberg University.

Hebrew Transcription Conventions

The following transcription procedures are adapted to the phonological and morphological structure of Israeli Hebrew.

This broad phonetic, or phonemic, transcription is intended to represent the way utterances are generally pronounced by speakers. It does not represent abstract underlying historical forms, but it does take into account distinctions still manifest in the current orthography of Hebrew in order to facilitate disambiguation of homophones and homographs.

TEXTLINE BROAD PHONETIC TRANSCRIPTION CHARACTERS

Alphabet	**Symbol**	
Vowels		Names of Relevant Diacritics
	a	*qamac*, *patax*, *xataf-patax*
	e	*segol*, *tsere*, **and** *schwa*
	i	*xiriq*
	o	*xolam*, *qamats qatan*
	u	*qubuts*, *shuruk*
Diphthongs		
	ay	e.g., *pnay, alay*
	oy	e.g., *noy, goy*
	ey*	at end of words, e.g. singular *more*, plural *morey*
	uy	e.g., *panuy*, *asuy*
Consonants + Names of Letters		
א, *alef*		Underspecified = Vowel e.g., *raa* = (he) saw, *tire* = (you/she) will-see, *af* = nose, *ered* = (I) will-go-down, *hevi* = (he) brought, *nasu* = (they) carried
ב, *bet*	b	e.g., *bolet*, *aba*, *hilbisha*

* For purposes of consistency, use *ey* in common closed class items (function words), irrespective of how they are actually pronounced: e.g., *eyfo* = *איפה, eyze* = *איזה, eyn* = *אין*

ב, *vet*	v	e.g., *ahuv, shovav, misaviv*
ג, *gimmel*	g	e.g., *gamar, nagar, boreg*
ד, *daled*	d	e.g., *delet, yidrosh, raqad*
ה, *heh*	h	e.g., *halax, herim, yaharos* (but not word-finally, e.g. *raca* ‘ wanted’ ~ ‘(she) runs’ – except for disambiguation, e.g. *lo* ‘to him’ / *loh* ‘no’, *bo* ‘in him’ / *boh* ‘come’)
ו, *waw* [*vav*]	v	e.g., *ve#, viter, givun, vav*
ז, *zayin*	z	e.g., *zar, muzar, brogez*
ח, *xet*	x	e.g., *xalom, maxaq, maxar* (‘tomorrow’ but also see under *xaf* below), *patax*
ט, *tet*	t	e.g., *teva’*, *matbe’a*, *moret*
י, *yod*	y	e.g., *yarad, tayar, siyer*
כ, *kaf*	k	e.g., *kelev*, *maka*, *kol* ‘all’
כ, *xaf*	x	e.g., *nixva*, *maxar* (‘sold’), *maxbesa*, *masax*
ל, *lamed*	l	e.g., *lilmod, melax, nafal*
מ, *mem*	m	e.g., *maca, limco, lit’om*
נ, *nun*	n	e.g., *nudniq, anaxnu, qatan*
ס, *samex*	s	e.g., *soleax, pasim, namas*
ע, *ayin*	'	[placed according to orthography] e.g., *'*omed, *'*elbon, ne'emad, masa*'*, tav'a (she drowned), tava*'* (he drowned)
פ, *pe*	p	e.g., *pe, pitput, hop*
פ, *fe*	f	e.g., *fibreq, filosof, mefic, mafria’, xacuf*
צ, *tsade*	c	e.g., *codeq, maca, hecic*
ק, *kof*	q	e.g., *qol* ‘voice’, *qafac*, *maqom*, *codeq*
ר, *resh*	r	e.g., *rosh, para, tofer*
שׁ, *shin*	sh	e.g., *shama’, xashuv, lavash*
שׂ, *sin*	s	e.g., *sone, masua*
ת, *taf*	t	e.g., *tafas*, *menateax*, *naxat*
[In loan words]	j	e.g., *juk, jins, pijama*
	ch	e.g., *chupar, richrach*
	zh	e.g., *bezh, garazh*

UNIFYING TRANSCRIPTION [FOR SEGMENTATION AND LEXICAL SEARCHES]

As a rule, certain forms of pronunciation should be corrected in the "text line" (which represents an utterance – a word, a clause, or a turn). This should be done in order to make the transcription as consistent as possible across speakers, for applying computerized lexical searches and (semi-automatic) coding. For example, in all word-initial cases of ***beged kefet*****,** use the stop [*b, p, k*], not fricative [*v, f, x*], even if the fricative is the standard. The **%pho:** tier can be used to indicate how the words were actually pronounced, e.g.:

*SBJ: ec ve# bayit.
%pho: u# vayit

*SBJ: kibasti et ha# xulca.
%pho: xibasti

Definite articles and orthographically prefixal prepositions should be separated from the next word by a double bar [keyboard pound key #] followed by a space, e.g. *ha# bayit sheli, be# Tel-Aviv, ba# bayit, le# maqom axer, la# horim sheli, me# ha# bayit, mi# Yerushalayim.*

Conjunctions that are orthographic prefixes should also be separated from the next word by a double bar [keyboard pound key #] followed by a space, e.g. *ve# ani xoshev she# hu meod nexmad kshe# hu yashen.*

This convention allows for (1) correct representation of disfluencies and repairs that occur frequently in speech, e.g. *ve# eh em atem meod nexmadim*, (2) easy retrieval of function words (closed-class items), (3) ready disambiguation (e.g., *mi#* 'from' vs. *mi* 'who'), and (3) inclusion or exclusion of these items from lexical counts [by use of the command freq -s"*#" *.cha].

[**Note**: The double bar is the CHAT convention for prefixation. This is not to suggest that articles, prepositions, and conjunctions in Hebrew should be analyzed as morphological prefixes]

Bound affixes (inflectional and derivational) are written as part of the word, including *le* as a marker of the infinitive, e.g., *lalexet, lefazer, livdoq; halaxti, mefazrim, bodeqet, yagid,* and inflected (non-nominative) pronouns, e.g., *li, alay, itanu, mimenu, eclam; hitkashrut, maxresha, havana, raqdan.*

Formulaic and other fossilized **bound expressions** (e.g., prefabs and collocations) are written as single items, e.g., *bimqom* vs. *ba# maqom, habayta* vs. *ba# bayit, bircinut* vs. *be# cura recinit.*
In cases where multi-lexemic expressions are written as two or more separate items in standard orthography, the underscore _ can be used, e.g., *axar_kax, sof_kol_sof, be#sofo_shel_davar.* [Note that in previous transcriptions the tilde ~ was used for these purposes, e.g. *sof~kol~sof*. This is a very specialized use of the CHAT convention for cliticization; In general FREQ output these items will appear as one consecutive string, e.g. *sofkolsof*]

Criteria used for identifying multi-lexemic expressions:
Syntactic inseparability and non-exchangeability, that is, you cannot (1) insert any other word or item into the expression or substitute another word for it in that same context, or (2) change the order of the elements in the expression.
In line with these criteria, the following expressions should be considered as multi-lexemic: *loh~kol~she#ken, af~'al~pi~she#, lamrot~zot, ve#xayoce~ba#ze.*

Hebrew **compounds** (*smixut xavura* 'bound construct state') are written as single items using the CHAT convention of joining the two words with a plus sign, e.g. *beyt+sefer*, *na'al+bayit*, *shulxan+ktiva, trumat+ha#nadiv*.

Acronyms, Abbreviations, and Numbers: CHAT does not allow numbers [digits] on the text line (unless they are inserted in square brackets), and so these should be written out in words, e.g. *axat*, *shtayim*, *shalosh*. Acronyms and abbreviations also need to be fully spelled out for word counts, e.b. ב"יס = *beyt+sefer* [but note instances such as *ve#xu* or *ha#nal*].

General comments:

- Use **capital letters** [upper case] for start of proper name, e.g. *Asaf, Hagar, Tel-Aviv*

- Use ***lo*** = לו 'to him, for it', ***loh*** = לא 'no, not' in order to differentiate between these frequently used items

- Use the replacement notation [: text] for a unified analysis of lexical items, e.g.

  *SBJ: ta# [: et ha#] telefon.
  *SBJ: vaksha [: bevakasha].
  *SBJ: ma ztomeret [: zot~omeret]?

  The FREQ command automatically replaces the material preceding the replacement notation with the material inside the square brackets.

NOTE:
With the shift to Arial Unicode font, Hebrew orthography can now be transcribed as such in CHAT, and WORD files in Hebrew orthography can also be converted into CHAT format. However, this requires careful conversion/transcription procedures, and does not allow for ready disambiguation of homographs, e.g. ספר 'book', 'hairdresser', '(he) counted', '(he) told a story', 'border' etc. In addition, the Hebrew MOR analyzer does not support Hebrew orthography at this stage.

For questions, comments, and suggestion, please contact Bracha Nir-Sagiv at brachan@post.tau.ac.il