5. TRANSCRIBING SPOKEN ISRAELI HEBREW: PRELIMINARY NOTES

SHLOMO IZRE'EL[1]
Tel Aviv University, Tel Aviv, Israel

INTRODUCTION: THE CORPUS OF SPOKEN ISRAELI HEBREW (CoSIH)

In this paper I wish to share with my readers some reflections concerning transcribing spoken Israeli Hebrew (IH), reflections which are the outcome of the need to form a set of guidelines for the transcription of The Corpus of Spoken Israeli Hebrew (CoSIH). These guidelines are derived from the goals, size, features and scope of the corpus.

The Corpus of Spoken Israeli Hebrew (CoSIH) will be compiled in order to facilitate research in a range of disciplines concerned with the Hebrew language and with the general methodology of Corpus Linguistics. The corpus will be disseminated in multimedia format and in print. The multimedia format will be disseminated via electronic means including CD-ROM, DVD-ROM and the World Wide Web, and will present the recorded sound simultaneously with its transcriptions and other extensions, all linked together by software.

The size of CoSIH will be 5 million words. It will consist of the following features:

- Digital audiotaped recordings
- Full synchronized transcripts in Hebrew orthography
- Narrow phonetic transcription of selected passages
- Glossing of selected passages
- Translations (into English) of selected passages

[1] I thank Mira Ariel, John Du Bois and Yael Maschler for reading a draft of this paper and commenting on it.

CoSIH is designed to include a representative sample of speakers and situations. It will consist of two complementary corpora: a main corpus and a supplementary corpus.

The main corpus, which will comprise about 90% of the entire collection, will be sampled statistically, representing both demographic and contextual variation. For analytical purposes it will use a conceptual tool in the form of a multidimensional matrix combining demographic and contextual tiers. We suggest three demographic variables: (1) ethnicity/religion, place of origin, and place of birth; (2) age; (3) education.[2] We suggest three main and two secondary contextual variables. The main contextual variables are: (1) interpersonal relations: intimacy vs. distance; (2) discourse structure: role driven vs. non-structured interaction; (3) discourse topic: personal vs. impersonal. The secondary contextual variables are the number of active participants in the discourse (monologue vs. dialogue) and the means of communication (face to face vs. telephone). The matrix will consist of 900 cells. A cell is the basic sociolinguistic unit of CoSIH. It is a recorded segment designated to include about 5,000 words of coherent continuous text. Each cell may consist of one or more texts produced by one or more speakers classified by the conceptual demographic-contextual matrix.

The supplementary corpus will include about 10% of the collected data, and will add to the statistically-sampled corpus some targeted demographically sampled texts and a contextually designed collection. CoSIH's design is culturally dependent to suit the special structure of the IH speech community and thus includes both native and non-native speakers of IH.

In order to get a more acute representativeness in linguistic data (of both demographic and contextual varieties), we will sample all of the textual data randomly. This will take place after all of the collected recordings from the sampled population are in hand. Each person (randomly) selected for the demographic sample will be asked to make a recording of all his or her activities over a 24 hour period. This span of time will be distributed homogeneously among the informants. Each of these one-day long recordings will be screened to remove long silent periods and long unintelligible speech passages, and from the remaining material, a 5000 words recording segment (between 30 minutes and an hour) will be randomly extracted. This will form the basis for the main, statistically balanced corpus. For a more detailed description of the CoSIH model see Izre'el (Hary and Rahav, 2001).

SPOKEN VS. WRITTEN

In a literate society like ours, where there is a long tradition of reading and writing, as well as a body of linguistic research more than a millennium old, the linguistic models we tend to construe are based on a long tradition of the study of the written medium. This is true not only for Semitic linguistics, for which the study goes back to the Middle

[2] The sex variable will come out as a result of the random sampling. In any case, the corpus database will enable users to retrieve data according to social and sociolinguistic variables at their will.

Ages, but also for general linguistics. Therefore, the study of the spoken language is almost exclusively based on concepts established for the study of the written language. Likewise, metalanguage leans heavily on the written medium. When we speak about components of the spoken language we always use terminology which sprung from the study of the written medium. We speak (and write) about *letters* instead of about phones or sounds,[3] we speak of root letters, functional letters, and so on. We think that we talk in sentences, although the sentence seems not to be a valid unit of the spoken language (see, e.g., Miller & Weinert, 1999: chapter 2).

However, if one takes spoken language as a distinct structural system, disregarding for a moment its relationship with the written medium, one can definitely show that it has a linguistic structure of its own, which is substantially different from the structure of the written language, one that we are so used to see and work with. This is true for every domain of language, in the lexicon and in the grammar, including phonology.

One is accustomed to speak of a phonemic system of a written language, yet a written language has no phones, so that its basic distinguishing components are not phonemes but graphemes, and its graphemic system is only in partial agreement with the phonemic system of the spoken medium. Even if one reads aloud a written text, there is a constant need to transpose the written graphemes into oral phones, and this transposition does not always show a one-to-one correspondence.

The Hebrew script is basically a consonantal one, and vowels are represented only partially and ambiguously. For example, a and e are never indicated in spelling, and they are either not spelled at all or spelled by a so-called *mater lectionis* (e.g., the letter ה, also used for /h/), yet still with no way of telling whether a or e is to be read. Similarly, the vowels u and o are indistinguishable even when indicated by the *mater lectionis* ו, which, on its part, is also used to indicate consonantal /v/ (for further details see, e.g., Bolozky, 1978: §1.1; Schwarzwald, 2001: §1.1).

The difference between the graphemic system of the written language and the phonological system of the spoken medium can be illustrated for IH in several ways and by several features. For example, written IH has two graphemes, ט and ת, which are never distinguished in speech, and must therefore be described as a single phoneme /t/. One other example is the noted pairs of b–v, k–x, p–f. These pairs have a complex way of transmitting them in the written language, as both b and v are written by ב, yet v can also be written by ו; k and x are both written by כ, but k can also be written by ק, and x can be written by ח; p and f are both written by פ. On the phonemic level, these pairs, although they may alternate systematically in some paradigms (e.g., navax "he barked" jinbax "he will bark"; Bolozky, 1997: §17.5.4), still show phonemic distinction, as in the minimal pair hitxabeʁ "joined" – hitxaveʁ" "became friends with" (Kutscher, 1982: 248–249; cf. also Schwarzwald, 1981: 24–30).

[3] Cf. the following statement, quoted from a study of awareness of the phonological system with laymen: "Speakers of English are able to manipulate phonemes only if they can read. The acquisition of the alphabetic representation of language enables the language knower to transfer this way of representation (i.e. sequences of discrete sublexical elements) to speech. In short, we know about phonemes because we know about letters" (Scholes & Willis, 1991: 220).

VISUALIZATION OF IH SPEECH

Since the spoken medium is acoustic, linear and temporally extended, visual transmission is necessary in order to enable any research, except, perhaps, for such as focused on individual, small units. Even in this latter case, one needs to transmit sound into the visual medium in order to publish the results. The linguist must therefore use a transcript of the spoken text.

Transcript types range from texts written in the standard orthography using accepted punctuation to the narrowest phonetic transcription which may include in addition intonational and other prosodic notation.

Here are examples of four types of transcription:

ˈmaaˈkeʃeʁ	ma akeʃeʁ	ma hakešer	מה הקשר
ˈafə̃ xaðlɔbɔˈbaətʃəˈli	af əxað lo babait ʃəli	af exad lo babait šeli	אף אחד לא בבית שלי
ˈəjtənɔˈteðleˈxatɑmafˈtɛax	aiti noteð lexa tamafteax	hayiti notenet lexa et hamafteax	הייתי נותנת לך את המפתח
aˈitʌoˈlex	aita olex	hayita holex	היית הולך

The right hand column represents the standard Hebrew orthography. The left-most column is a narrow phonetic transcription. Next to it comes a semi-narrow transcription, and to its right a rote, intuitive Latin transcription used widely by both professionals and laymen (with some variation, e.g., *š* ~*sh*, *x* ~ *ch*).

Any type of transcription, including the narrowest one, is based on theory, since there is no way of transforming the infinite range of acoustic features into phonetic symbols. Therefore, any type of transcription must be anchored in a theoretical ground. The theoretical ground depends on research goals (Ochs, 1978; Du Bois, 1991; Edwards, 1993: 3–5; Crowdy, 1994: 25; Kennedy, 1998: §2.6.4.2; Blanche-Benveniste, 2000: 63).

As mentioned, the form in which CoSIH will be transcribed must be derived from its goals, size, features and scope. CoSIH is designed to serve many research activities, among them linguistic, sociolinguistic, cultural, educational, and language engineering (see the CoSIH web page <http://www.tau.ac.il/humanities/semitic/cosih.html>). In other words, we have tried to create a corpus model that will not be limited in its scope to specific linguistic investigations, but rather to meet with a wide range of linguistic and extra-linguistic interests. Its size, 5 million words, requires some serious limitations as regards project duration, human power and financing, since 5 million words is a large corpus in terms of spoken corpora (Blanche-Benveniste, 2000: 63). The texts will be recorded in natural settings, which means an often noisy environment and many overlaps between speakers, just to mention two of the most conspicuous problems for transcription. Existing corpora of similar size and scope are all transcribed in the standard orthography, and may include some additional notations, primarily of conversational features or intonation (e.g., Svartvik & Quirk, 1980; Du Bois et al., 1992, 1993). From both our experience in the pilot study and from experience of others we note that one needs many dozens of hours to transcribe one hour of a spoken conversation recorded in a natural setting. Given the above, and since CoSIH will present its texts to the user in both sound and transcript, we have decided to have

CoSIH transcribed not in a phonetic transcription of any kind but in the standard orthography.

A broad phonetic transcription is based on a prior phonological analysis and on phonological assumptions. While some phonological studies have been made on spoken IH (e.g., Bolozky, 1997), these were made in previous times, based on only a few varieties of IH, and, most importantly, were not based on data drawn from a large-scale corpus. There is definitely a need for fresh analyses of IH phonology (or, depending on the inclination of the researcher, phonologies). This will require a narrow transcription of the spoken texts. A narrow transcription endeavors to transmit to the written medium as many features of the spoken utterance as possible if the phonological system of the language is unknown, or, if it is known, to transmit allophonic variation (International Phonetic Association, 1999: §5). Offering a phonetic transcription of our own cannot be based, at least at this stage, on research, and, in any case, any future student of phonology or phonetics will often prefer to transcribe the spoken texts for his or her own needs. Since "any phonetic symbol—for vowel, consonant or prosodic feature—can be applied to a range of sound-types" (Wells and House, 1995: 2), the sound of CoSIH's texts will be available to the end user in digital form. One other point with regard to narrow transcription, especially one that includes in addition intonational and conversational notation, is the constant gap between the need to transmit as many features as possible so that we lose no significant feature and the need to keep some ease in the reading of an overloaded text (Ochs, 1978: 44).

In the case of Hebrew, which is spelled in a script which is not based on the Latin set of characters and goes from right to left, there is still another option, which is a transcription based on native intuitions of phonology, a type of transcription used in several settings and circumstances, as in some email correspondence or telegrams, as well as by some IH linguists (notably in the four IH grammars written in English: Berman, 1978; Rosén, 1977; Glinert, 1989; Schwarzwald, 2001). This option is illustrated in the third column (from the left) of the set of transcriptions above. This solution may be beneficial for scholars who would like to use CoSIH for reference for studies such as comparative corpus linguistics rather than IH, and will also ease handling of the texts in computers, which are more apt to left-to-right English-based script manipulation. However, as much as non-Hebrew linguists form an important sector of users, they are but a fraction of the many users who can manipulate Hebrew. For these users, working with Hebrew characters is much easier than working with transcribed Hebrew. More importantly, I believe we should refrain from using any type of broad transcription, especially one that is based on intuition rather than on a prior analysis of IH phonology, since it may block insights into the phonological structure of the language, by precluding study whether consiously or unconsciously. Transcribing CoSIH in the standard orthography is the most arbitrary transcription available, and thus the least problematic. As mentioned, CoSIH will offer its users selected passages with a (relatively) narrow phonetic transcription of each of its cells, so that some idea of the IH linguistic variation will be available.

Let me illustrate the type of theoretical issues involved in transcription. In the text transcribed above there are three occurrences of the phone [x], two of which

are represented in the standard orthography as ח, one as ך (an allograph for כ at word-end). In some IH ethnolects, these two graphemes are pronounced [ħ] and [x] respectively, and are taken to represent two distinct phonemes (Blanc, 1964; Devens, 1980). Ornan (1974) claims that these two phonemes also exist in standard Israeli Hebrew. There is obviously a theoretical issue involved in the interpretation of the standard pronunciation of these two graphemes, whether it stands for two distinct phonemes or a single one, and scholars differ in their views regarding this subject (for a convenient summary see Waldman, 1989: 238–240). Obviously, any description and any explication of the system is dependent on theoretical background. It may be treated in different linguistic domains, phonological, morphological and morphophonological. Finally, it includes different points of view on language, whether one regards all varieties of IH as reflecting a single phonemic set with demographic variation in producion or as two distinct sets, each defined for a distinct linguistic variety. By the same token, contextual varieties may be taken to represent a continuum within a single system, or separate, tangential systems. One last, related question is whether the written and the spoken media are to be represented as a single system or as distinct systems.

Variation

Variation is inherent to language, and IH is not excluded. There is variation dependent on demographic diversity and there is contextual variation. In the speech of a 63 year old woman of Moroccan origin one finds some neutralization of voicing; e.g.,

(1) baʕaˈla u daˈjak
Her husband is a fisherman

The standard IH pronunciation of [daˈjak] is [daˈjag], which is probably also the phonemic string underlying the speech of the informant in this case, and this is also reflected in the IH orthographic representation: דייג. Word-final devoicing is a feature of other ethnolects as well. Another feature manifest in this informant's speech is the existence of the phone ʕ, interpretable as phonemic in her speech, as it is in the speech of many IH speakers of Mizrahi ("Oriental") origin, those who came from Arab countries. This feature is not necessarily lost in the speech of their descendents born in Israel (Devens, 1980).

Variation, both demographic and contextual, is manifest in the use of many forms with the pairs b–v, k–x, p–f; e.g., xiba ~ kiba "turn off (light)"; bikeʃ ~ vikeʃ "asked for"; yitpos ~ yitfos "he will catch", and so forth. As is expected, substantial variation is manifest mostly in vowels. Besides individual and environment-dependent variation, there is also morphophonological variation as in the cases of *mekir* ~ *makir* "be acquainted with" (Ravid, 1995: 41, 84, 163–165; Bolozky, 2003: §1.1.2).

One other type of variation is fast-speed reduction. Two notable examples are t͡sax ~ t͡saʁix – t͡sxa ~ t͡sʁixa – t͡sxim ~ t͡sʁixim "need (sg m, sg f, pl)" and ta+noun ~ et ha+noun (accusative-marker+definite-article+noun). As against English, IH has

not established norms for writing such forms in their short variants (except for, perhaps, the latter example, spelled תי in representations of speech, mostly in literature). Such reduced forms are easily handled by rules of fast speech (Bolozky & Schwarzwald, 1990; Bolozky, 2003: §1.2.)

Representing variation—of all types and in all domains—will not allow computer searches to be carried out on the texts without prior tagging. Since the Hebrew script is basically a consonantal one, with only a limited capability to denote vowels, it conceals variation and is therefore much more morphologically transparent than any of the phonetic transcriptions. It thus lends itself much more to easy and intuitive search operations.

On the other hand, using the written norm is contradictory to the wish to detach research of the spoken language from leaning on the written standards. Transcription in the standard orthography, precisely because it represents quite a different system from the spoken language, is not misleading as regards interpretation and analysis of the transcribed form, given that users are well aware of its arbitrariness. Such awareness will enable all users to utilize the transcript for their respective interests, after transcribing themselves the texts in a phonetic transcription that will fit their own research goals. Transcripts in standard orthography are not targeted to deal with phonetic or phonological observations, and can be used by interested phoneticians and phonologists only for quick reference. Still, there are some cases where a phonetic transcription seems to be a necessity, notably where homographs are indistinguishable and can cause ambiguity.

Homographs

Hebrew orthography may use vocalization signs and diacritics which are put either above, inside or under the letters, and may enable disambiguation of homographs (for examples 2–4 below, notice שֶׁמָּה ʃe'ma, שְׁתִקי ʃte'ki, כְּבְּסָה xib'sa vs. כִּבְּסָה kib'sa). However, these vocalization marks and diacritics are inserted between the letter characters by computers and will thus inhibit comparison of letter strings for search operations.[4] IPA notation external to the respective Hebrew strings is therefore preferable. Examples:[5]

(2)

ס: ההורים שלהם רוצים שהם יהיו בחגים בבית ||
ע: שמה {ʃe'ma}\
ס: שהם יהיו בבית בחגים ||

S: Their parents want that during the holidays they will be home.
O: That what?
S: That they be home during the holidays.

[4] While the software can be modified to skip such characters while searching, there are still some Hebrew vocalization characters that are inserted insead rather than between existing characters (e.g., וֹ, וּ).

[5] Symbols: | intonation unit boundary with a continuing tone; || intonation unit boundary with a final tone; \ intonation unit boundary with an appeal tone; - truncated word; – truncated intonation unit (cf. Du Bois et al., 1992, 1993).

The letter string שמה can be interpreted either as ˈʃama "there", as ˈʃema "lest" or as ʃeˈma "that what".

Then shut up. אז שתקי‖{ʃteˈki} (3)

The letter string שתקי can be interpreted in two ways: normative ʃitˈki or standard colloquial ʃteˈki.

She washed it כיבסה{xibˈsa}אותו| (4)

The string כיבסה can be read as either kibˈsa or xibˈsa, depending on the idiolect, and in some cases also on the sociolinguistic context.

Examples 3 and 4 are illustrative of our inclination to draw attention to forms that are not accepted as normative, although some may be standard in colloquial speech.

A special type of homographs are ones where the spoken language shows a radical transmutation of the orthographical norm to the extent of forming a different standard. This is the case, for example, in the string ˈbo(ə)na in the following two examples:

Hey,	בוא הנה {ˈbona}‖	(5)
(the price of) this gas	הדלק הזה|	
is tearing me apart	קורע אותי|	
Omer—	עומר|	
a whole lot!	חבל על הזמן‖	
Hey,	בוא הנה {ˈboəna}‖	(6)
every other day I–	אני כל יומיים—	

In these examples one notices a completely different use and perception of the IPA-transcribed string from the respective standard orthographic string. The latter is to be interpreted as "come here". In this case, probably exclusive to the spoken medium, ˈbo(ə)na serves as a presentation adverb, probably indeclinable, and pronounced rather differently than the pronunication reflected in the written form ([bo ˈhena]). When studying this string in the spoken language one will have to look for all its occurrences in the corpus in order to find a definitive answer to its structure, syntax, meaning and pragmatics. CoSIH cannot deal with these issues before being presented to the public, but may find this and other strings worthy of noticing.

Morphological variation is hard to compound by escaping into the standard orthography. Some notable cases are the verbal prefix-conjugation 1st person variants ʔ∼Ø∼j, gender variation in numerals, and the common plural forms with final m for normative n for the feminine. These and other such morphological variants will have to be transcribed in their colloquial forms also in the standard orthography, and some lists of variants will be given for the benefit of search operations. Examples: אשיר {aˈʃiʁ / ʔaˈʃiʁ} ∼ ישיר {jaˈʃiʁ} "I will sing", where the latter form is a colloquial innovation; שלוש ימים {ʃaˈloʃ jaˈmim} ∼ שלושה ימים {ʃloˈʃa jaˈmim} "three days", where the first is colloquial and the second normative and more widespread in educated speech; אתם{aˈtem} "you (pl c)" ∼ אתן{aten} "you (plf)" (see below).

DEVIATING STANDARDS AND TRANSCRIPTION

Given the above limitations, a preliminary principle must always be present in the transcriber's mind: the need to represent any spoken string as accurately as possible. While this principle may seem obvious, it is sometimes hard to implement. There are two main difficulties set by the sociolinguistic system on the way to a straightforward implementation of this principle:

(1) Bridging the structural gap between the written and the spoken;
(2) Rectifying the tendency to draw from the standards of the idiolect.

The structural gap between the written and the spoken languages can be illustrated by the following example:

What	מה	(7)
Don't you (2plf) think	אתן לא חושבות	
What are you (2plf) fucked up	מה אתן דפוקות	
He got angry	התעצבן [6]	

This is the first draft of a transcription made by a transcriber with good awareness of the spoken medium. Still, she replaced the spoken standard form of the second person plural pronoun, which is unmarked for gender (traditionally masculine) אתם, pronounced atem (with a final m), with the second person plural pronoun of the feminine, אתן/ʔaten/ (with a final n), much less used in speech than in writing. This difference between the two media of IH is further enhanced by prescriptive tendencies, still widespread in Israel today (cf. Téné 1996).

Drawing from an idiolect's standard may include imposing lectal standards of the transcriber on his or her perception of the informant's language. This is expected where there are demographic differences between the respective idiolects. Such demographic differences can be of all sorts: ethnic, age, education, sex, and so on. The following example is a text of an elderly, uneducated woman of a low socioeconomical status, of Moroccan origin:

So the woman	אז האישה |	(8)
who is waiting for her to come live with her,	שהיא מחכה שתבוא לגור איתה |	
the capable woman,	האשת חיל |	
the woman who has to come,	האישה שהיא צריכה לבוא |	

In the second line the transcriber missed a 3 f sg pronoun between the nominalizer ש /ʃe/ and the verb תבוא ta'vo "she will come", although a similar construction occurs both immediately before and after: שהיא מחכה ʃi:meχa'ka "nom.+she waits"; שהיא צריכה ʃic(ʁ)i'xa "nom.+she needs". Aided by the assimilation of the vowel e by the following i in the string ʃi: (←/ʃe+hi/; cf., with further shortening, ʃi in ʃit͡s(ʁ)i'xa),

[6] This example is taken from a draft without prosodic notation.

the omission in writing of the pronoun is possibly caused by the difference in syntactic usage of the two dialects, that of the informant and that of the transcriber, who would not usually use a pronoun here. One may note that in Hebrew the subject is inherently expressed in a verbal form, but not in participles and adjectives. Whereas tavo is a verb, meχa'ka and t͡s(ʁ)i'xa are not. Therefore, the insertion of a personal pronoun between /ʃe/ and tavo, which inherently includes the subject in the morpheme /t/, is less expected (at least in the transcriber's idiolect) than between /ʃe/ and meχa'ka or between /ʃe/ and t͡s(ʁ)i'xa.

IMPLICATIONS FOR AN ACCURATE UNDERSTANDING OF LANGUAGE

To conclude my sporadic collection of notes on transcription, I wish to draw attention to one example that illustrates how transcription may bear significant implications for our understanding of the linguistic structure of a text.

But it is obvious that it will- was e – {–e }אה{ i: ɛ'æ } אבל זה ברור שזה י- היה (9)

Both the Hebrew transcription and the IPA one draw from the transcriber's interpretation of the string he has heard. According to his interpretation, there is hesitation between the "future" and "past" forms of the verb "to be" here: -י(/j/) is the 3 sg m personal prefix of the verb, and היה haja indicates the past-tense form "he was". The IPA transcription as presented is similarly interpretative in that it shows a space between i: and ɛ'æ. However, the vocalic segments as heard are not that certain, as they can also be interpreted, at least prima facie, as a production of the form /jihje/: including the segment i: as part of the verb, and therefore interpreted as future. Given that the actual time of the referred situation is known, each interpretation has significant implications for the study of the Tense-Mood-Aspect system in this lect. Surely, both transcriptions are theory-bound. Apart from those theoretical premises which lie behind the type of transcription, there is the question of whether the string ɛ'æ can reflect phonemic /haja/ and whether it can reflect /jihje/. Only a profound linguistic analysis conducted on a full-scale corpus can answer these questions.

CONCLUSION

The way a spoken string recorded for use in a corpus passes from its conception to its being read by the end user is lengthy and full of transmissions. The following is a simplistic representation of it:

message > immediate context > utterance > environmental intermediate (noise) > electronic mediator > hearing > decipherment > transcribing > reading > message

There are further complications in this transmission line, of which the most important one is perhaps the cultural, personal and linguistic background of the informant as well as those of the transcriber and the end user.

As is well known, ambiguity is always on the side of the hearer, almost never on the side of the speaker. Ambiguity results from many reasons, linguistic or extra-linguistic. One example will suffice to illustrate the type of misunderstanding caused by ambiguity and its possible ramifications to linguistic analysis:

a'marti'kax'po'kax'po

This string is interpretable in two ways:

a'mar ti'kax po kax po (10a)
He said: "Turn here, turn here."
a'marti kax po kax po (10b)
I said: "Turn here, turn here."

These two interpretations are possible because the sg m imperative kax can be replaced by the 2 sg m "future" form tikax, and since the first person of the "past" form adds the morpheme -ti to the bare stem, used for the 3 sg m. The first interpretation of the string (10a), submitted to me by a rather experienced transcriber (from a conversation which she herself had participated in!), raises the question whether it is possible for a Hebrew speaker to use both the "future" form and the imperative in such repetitive utterances. The alternative interpretation (10b) suggests a repetition of the same string, which seems more intuitive. Further contextual analysis seems to support the second parsing.

Ambiguity may result from homonymy, linguistic context, speech interaction and linguistic practices (e.g., overlaps), noisy environment, physical dislocation, channel of communication, cultural context, cultural background, personal background, cultural or social gap, and so forth. Moreover, both the transcriber and the end user are removed from the background, the environment and the visual aids which help interlocutors understand each other. When The Corpus of Spoken Israeli Hebrew (CoSIH) will be available, its end users must be urged to always check the spoken medium and take their own stand regarding the offered transcription. Transcription is theory, and theory may be verified or refuted. In any case, it must be tested.

BIBLIOGRAPHY

Berman, R. A. (1978). *Modern Hebrew structure*. Tel Aviv: University Publishing Projects.

Blanc, H. (1964). Israeli Hebrew texts. In *Studies in Egyptology and linguistics in honour of H. J. Polotsky* (pp. 132–152). Jerusalem: Israel Exploration Society.

Blanche-Benveniste, C. (2000). Transcription de l'oral et morphologie. In M. Guille and R. Kiesler (Eds.), *Romania una et diversa: Phologische Studien für Theodor Berchem zum 65., Geburstag*. Band 1: Sprachwissenschaft. Tübingen: Gunter Narr. 61–74.

Bolozky, S. (1978). Some Aspects in Modern Hebrew Phonology. Chapter 2 in: Ruth Aronson Berman. *Modern Hebrew structure* (pp. 11–67). Tel Aviv: University Publishing Projects.

Bolozky, S. (1997). Israeli Hebrew phonology. In A.S. Kaye (Ed.). *Phonologies of Asia and Africa (Including the Caucasus)* (pp. 287–311). Vol. 1. Winona Lake, Indiana: Eisenbrauns.

Bolozky, S. (2003). Phonological and morphological variation in spoken Hebrew. In B. Hary (ed.). *Corpus linguistics and Modern Hebrew: Towards the compilation of the corpus of spoken Israeli Hebrew (CoSiH)*. Tel Aviv: Tel Aviv University; The Chaim Rosenberg School of Jewish Studies. 119–156.

Bolozky, S. & Schwarzwald, O. (Rodrigue) (1990). On vowel assimilation and deletion in casual Modern Hebrew. *Hebrew Annual Review*, 12, 23–45.

CoSIH web page: <http://www.tau.ac.il/humanities/semitic/cosih.html>

Crowdy, S. (1994). Spoken corpus transcription. *Literary and Linguistic Computing*, 9: 25–28.
Devens, M. S. (1980). Oriental Israeli Hebrew: A study in phonetics. *Afroasiatic Linguistics* 7(4), 26–37 (127–141).
Du Bois, J. W. (1991). Transcription design principles for spoken discourse research. *Pragmatics* 1, 71–106.
Du Bois, J. W., Cumming, S. Schuetze-Coburn S. & Paolino D. (1992). *Discourse transcription*. (Santa Barbara Papers in Linguistics, 4.) Santa Barbara, CA: Department of Linguistics, University of California, Santa Barbara.
Du Bois, J. W., Schuetze-Coburn, S., Cumming S. & Paolino D. (1993). Outline of discourse transcription. In J. A. Edwards' & M. D. Lampert (Eds.). *Talking data: Transcription and coding in discourse research* (pp. 45–89). Hillsdale, NJ: Lawrence Erlbaum Associates.
Edwards, J. (1993). Principles and contrasting systems of discourse transcription. In J. E. Edwards & M. D. Lampert (Eds.). *Talking data: Transcription and coding in discourse research*. Hillsdale, NJ: Lawrence Erlbaum Associates. 3–31.
Glinert, L. (1989). *The grammar of Modern Hebrew*. Cambridge: Cambridge University Press.
International Phonetic Association. (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press.
Izre'el, S., Hary B. & Rahav G. (2001). Designing CoSIH: The corpus of spoken Israeli Hebrew. *International Journal of Corpus Linguistics* 6(2), 171–197.
Kennedy, G. (1998). *An Introduction to corpus linguistics*. (Studies in Language and Linguistics.) London: Longman.
Kutscher, E. Y. (1982). *A history of the Hebrew language*. Edited by Raphael Kutscher. Jerusalem: Magnes.
Miller, J. & Weinert R. (1999). *Spontaneous spoken language: Syntax and discourse*. Oxford: Oxford University Press.
Ochs, E. (1978). Transcription as theory. In E. Ochs & B. B. Schieffelin (Eds.). *Developmental pragmatics* (pp. 43–72). New York: Academic Press.
Ornan, U. (1974). Ordered rules and the so-called phonologization of ancient allophones in Israeli Hebrew. In H. Luigi (ed.), *Proceedings of the Eleventh International Congress of Linguistics, Bologna-Florence* (pp. 1023–1036). August 28-September 2, 1972. Vol. II. Bologna.
Ravid, D. D. (1995). *Language change in child and adult Hebrew: A psycholinguistic perspective*. New York: Oxford University Press.
Rosén, H. B. (1977). *Contemporary Hebrew*. The Hague: Mouton.
Scholes, R. J. & B. J. Willis. (1991). Linguists, literacy, and the intensionality of Marshall McLuhan's Western Man. In D. R. Olson & N. Torrance (Eds.). *Literacy and orality*. Cambridge: Cambridge University Press. 225–235.
Schwarzwald & O. (Rodrigue) (1981). *Grammar and reality in the Hebrew verb*. Ramat-Gan: Bar-Ilan Univesrsity Press. (In Hebrew.)
Schwarzwald, O. R. (2001). *Modern Hebrew*. (Languages of the World/Materials, 127.) München: LINCOM EUROPA.
Svartvik, J. & Quirk, R. (1980). *A corpus of English conversation*. Lund: Lund University Press.
Téné, D. (1996). Three notes on Hebrew planning. In: *Evolution and renewal: Trends in the development of the Hebrew language*. Lectures Commemorating the 100th Anniversary of the Establishment of the Hebrew Language Council. (Publications of the Israel Academy of Sciences and Humanities, Section of Humanities.) Jerusalem: The Israel Academy of Sciences and Humanities.
Waldman, N. M. (1989). *The recent study of Hebrew: A survey of the literature with selected bibliography*. Winona Lake, IN: Eisenbrauns.
Wells, J. & House, J. (1995). *The sounds of the International Phonetic Alphabet*. (Booklet and CD.) London: Department of Phonetics and Linguistics, University College London.