The Corpus of Spoken Israeli Hebrew (*CoSIH*); Phase I: The Pilot Study

Shlomo Izre'el and Giora Rahav
Department of Hebrew and Semitic Languages; Department of Sociology
Tel-Aviv University
IL-69978 Tel-Aviv, Israel
{Izreel; grrhv}@post.tau.ac.il

Abstract

The Corpus of Spoken Israeli Hebrew (CoSIH) is, to the best of our knowledge, the first corpus designed to integrate both demographic and contextual variables in its compilation of texts. The suggested design is culturally dependent to suit the structure of the Israeli Hebrew speech community, yet the principles governing this design are such that they would service study of many other speech communities, to the extent that the design itself may be employed in the compilation of other language corpora with the necessary, culture-dependent modifications. A detailed description of the design can be found in Izre'el, Hary & Rahav (2001).

In the paper offered for the workshop, we describe the pilot study of *CoSIH*, its procedures and some of its lessons. The results of the pilot study will bring about some changes in the final model of *CoSIH* and in some procedural strategies. We will address a few of the key issues involved in the construction of the corpus in order to achieve the analytical model we have designed. These are: (1) Demographic sampling and recruiting informants; (2) Evaluation of sequential longitudinal recording: technical matters and ethical issues; (3) Contextual sampling: long- and short-term time sampling, speech sampling; (4) The concept of 'cell'. Lastly, the issue of transcription and annotation will be addressed briefly.

Introduction: The Corpus of Spoken Israeli Hebrew (*CoSIH*)

The objective of the *CoSIH* project is at creating a corpus of spoken Israeli Hebrew in order to facilitate research in a range of disciplines concerned with the Hebrew language, the sociolinguistics of the Hebrew speaking community in Israel, and with the general methodology of Corpus Linguistics. The corpus will be disseminated publicly in multimedia format.

A detailed description of the *CoSIH* project has been published in Izre'el, Hary & Rahav (2001). In this paper we describe the pilot study of *CoSIH*, its procedures and some of its lessons. The results of the pilot study will bring about some changes in the final model of *CoSIH* and in some procedural strategies. This discussion should be viewed as an appendix to the above-mentioned paper, and each section below will be referring to the respective section in that paper (abbreviated IJCL'01). A short summary of the *CoSIH* project is nevertheless in order.

With the outburst of corpus linguistics and the tremendous advance in the use of computers, many spoken corpora have been compiled and disseminated. Some of them have been compiled with attention to demographic and contextual varieties (useful gateways to reviewing such efforts are, among others, the SFB 411 one or the Gateway for Corpus Linguistics on the Internet; further references and some discussion can be found in *CoSIH*'s website and in IJCL'01). *CoSIH* is, to the best of our knowledge, unique in its method to combine both kinds of variables into a single model.

CoSIH is designed to include a representation of most varieties of spoken Hebrew as it is used in Israel today. *CoSIH* 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. For analytical purposes it will use a conceptual tool in the form of a multidimensional matrix combining demographic and contextual tiers. 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.

Daylong recordings of 950 informants and 50 other linguistic events (mostly from the media) will be collected within one year along with respective sociolinguistic data. These recordings will be evaluated, and a sample from each will be transcribed, to set up a five-million-word corpus.

CoSIH will be a basis from which research in many diverse areas will be launched, including, inter alia, theoretical and applied linguistics, sociolinguistics and cultural studies, communication studies, corpus linguistics, computational linguistics, translation studies, and many more.

While these are mostly long-term objectives, our immediate objectives are: Analysis of the Israeli linguistic community: its ethnolinguistic distribution, its linguistic and sociolinguistic behavior and attitudes; the study of demographic and contextual varieties ('dialects' and 'registers') as related to corpus compilation. Setting up a spoken corpus constitutes the initial phase of a major change in our view of language, i.e., as a multi-variant and dynamic continuum. Looking at language differently, as a multi-variant dynamic continuum is a primary target that 21st century linguistics should adopt, and the compilation of corpora is a necessary initial stage for such an endeavor.

The Corpus of Spoken Israeli Hebrew (*CoSIH*) is, to the best of our knowledge, the first corpus designed to integrate both demographic and contextual criteria in its compilation of texts. The design is highly innovative in this respect, and it is expected that its implementation will be a significant contribution to the discipline of corpus linguistics.

The suggested design is culturally dependent to suit the special structure of the Israeli Hebrew speech community and thus includes both native and non-native speakers of Hebrew. Yet the principles governing this design are such that they would service study of many other speech communities, to the extent that the design itself may be employed in the compilation of other language corpora with the necessary, culture-dependent modifications.

CoSIH Phase I: Pilot study

The pilot study, which included also the first steps of a pretest, aimed at achieving the following goals:

(1) To review a variety of linguistic groups among the Israeli Hebrew speech community in terms of linguistic and sociolinguistic behavior.
(2) To study issues involved in random sampling of the population.
(3) To study procedures involved in recruiting informants, eliciting natural recordings and sociolinguistic data.
(4) To study differences in attitude towards cooperation of informants from different sections in the population.
(5) To study issues involved in a sequential longitudinal recording by informants.
(6) To study technical tools, data recording techniques, and transcription issues.
(7) To make preliminary observations as regards linguistic contexts as related to different types of population.

Procedures taken were as follows:

(1) Recruiting informants in quota sampling and getting their preliminary consent to take part in this research.
(2) Instructing each informant as regards recording.
(3) Sequential recording by informant in a variety of time spans.
(4) Tapes collection from informants; inquiry about settings and conditions of the recordings made.
(5) Conducting a sociolinguistic interview.
(6) Questioning the informants as regards technical issues and problems encountered.
(7) Signing consent forms granting us permission to use the recordings.
(8) Preliminary organization of raw data (recordings and written forms).
(9) Evaluation of recordings: quality, language use, sufficient data, etc.
(10) Time sampling and selection of recorded samples to be included in the corpus.
(11) Data organization, registration in database, storage.
(12) Hebrew transcripts.
(13) Selection of segments for expanded analyses: phonetic transcription; glossing; English translation.
(14) Analysis of recordings for evaluation of distinct linguistic varieties for demographic and contextual variation.
(15) Evaluation of Phase I as a whole.

Some Key Issues: Goals, Alternatives and Lessons Gained

In this part of our paper we would like to address a few of the key issues involved in the construction of the corpus in order to achieve the analytical model we have designed. These are: (1) Demographic sampling and recruiting informants; (2) Evaluation of sequential longitudinal recording; (3) Contextual sampling; (4) The concept of 'cell'.

(1) Demographic Sampling and Recruiting Informants (IJCL'01: §§5.1.1, 5.2.1)

While the representative informants for *CoSIH* will be recruited by a probabilistic procedure, we have used quota sampling for the pilot study, trying to reach a wide coverage of the main socio-demographic groups in the population. Recruiting informants for this phase was made by three data collection agencies (a university associated agency and two well recognized, reputable commercial agencies). Each of the agencies was asked to collect data from 16 informants according to the demographic categories presented in Table 1:

Age	Edu-cation	Ashke-nazi	Mizra-hi	Arabs	Special groups
young	≤high school				
	>high school				
old	≤high school				
	>high school				

Table 1: Demographic categories

The three first groups (=columns) were set to fit, mutatis mutandis, the major demographic sections of the Israeli Hebrew speaking community: Jews of European or other Western ethnic origin ('Ashkenazi'); Jews of Asian or African ethnic origin ('Mizrahi'); non-Jews, of which the majority are Arabs, comprising ca. 20% of the Israeli population. The fourth column, 'special groups', was set to consist of three demographic sections for which we hypothesized to show significant differences in their use of language and in their linguistic structure: ultra-orthodox, soldiers and members of other security forces, and recently-arrived immigrants. Each agency was assigned one of these latter groups.

Of the three major ethnic groups, each agency was assigned to recruit four informants: two young (<20) and two old (>50), two with high education, two without. Lastly, each agency was instructed to recruit men and women in equal numbers, irrespective of any of the other criteria.

By choosing to hire a data collection agency we followed the procedure of BNC. We hired three agencys at the pilot phase in order to study procedures and pave the way to select one or more for the larger project, and we now have some idea about the pluses and minuses of each.

This decision has proven right. Academics in general, and linguists in particular, are not the ideal people to knock on doors and persuade people to join them in their research. Survey employees have enough patience and experience to do that, given that they themselves are persuaded by the need to conduct such a research. Money too is a factor in recruiting informants, although not of any kind. The rich or yuppies would not be tempted to be exploited for less than $50. Others would be too shy to do so in any case, or too short of self-confidence. People like me, who tend to get annoyed from answering commercial phone calls, will also tend to decline this generous offer…

The rate of consent to take part in such a burdening undertaking is an important factor for achieving a reasonable representative sample of the population. The B.I. and Lucille Cohen Institute for Public Opinion Research at Tel-Aviv University conducted a telephone survey for us on this issue, asking the following question:

- Would you be willing to take part in the future in a unique research in which you will be asked — for payment — to record all your daily activities during one day?

This survey was conducted on 1170 people, who consisted a representative sample of the Jewish population. A representative sample of 1170 individuals was surveyed. 40% of the respondents answered this question positively. However, as the response rate was about 55% of the households (or apartments), this means that the rate of positive responses may be as low as 22% of the whole population. Among those, only half, viz., 11%, are expected to eventually agree to full cooperation.

Among the Arab population, the consent rate was 24% out of 150 people who were asked. This means a much lower rate of consent than in the Jewish population. As expected, there are differences in consent tendencies among various sections in the population. For example, consent is lower among men than among women in the Jewish sector, while it is higher among men than among women in the Arab sector (the difference seems to be lower among Arab women with high education). There is further a problem of language use among Arab women of lower education, as they do not tend to use Hebrew in daily life, if they speak the language at all. Arabs tend in general to be less open to take part in research of the type proposed, due to their more prominent concern regarding privacy, as well as due to some political anxiety. Concern for privacy is shared by other sectors in the population, notably ultra-orthodox and soldiers. Political anxiety may also be found among new immigrants from the Former Soviet Union. We therefore expect problems in sampling in some sectors of the population, and may need to resort to quota sampling if the random sampling will result in lesser representation of some sectors.

(2) Evaluation of Sequential Longitudinal Recording (IJCL'01: §5.2.2)

(a) Technical Matters

We used Sony TCD-D100 DAT recorders with Sonic Studios stereophonic DSM-1S/L microphones. Each cassette has a capacity of four hours of quality recording. The acoustic output is excellent. However, the recorders seem to have caused difficulties in technical handling, especially at the point of replacing cassettes. Therefore, our first recruited informants from each agency recorded 8 hours each, and the last got to a full 24-hour span, with four or five cassettes each. This enabled the agency's representative to study the technical issues and the ways to overcome them with the informants. Unfortunately, many of the cassettes came back either empty or not fully recorded. In other cases, the sound of the recording person, i.e., our informant, who is closest to the microphones, was distorted. This has proved to be an especially unhappy situation, since it seems that the blame went to our representatives, who failed either to instruct or to supervise quality recordings, with the result being that our targeted informant was not recorded properly. One other crucial point was the physical connection between the microphones and the recorder, which caused distorted recording and at times even loss of some. One last problem is power supply. Two internal lithium batteries are good for some six or seven hours of recording. Still, for the informant's convenience and in order to ensure recording continuity, their replacement should have been made along with the replacement of a cassette. We used instead a battery sled assembly, which holds four C-type batteries. This power supply is good for 24 hours, so that batteries would not need to be replaced during any single recording session. This usually proved to be the case, yet informants have complained on their weight.

Fortunately, time heals in this case, and recent developments in digital recorders will enable us to use quality long-term hardware recorders with no operational complications. At this time, 6-hour sequential recording seems feasible, but when we get to the larger project, we may be able to use still better, more convenient equipment. Hardware recording with no mechanics may further lead toward some other solution regarding power supply.

(b) Ethical Issues

During a whole day our informants meet with people, with which they may have more or less meaningful conversations. The recording equipment was put into a pouch that was carried on the belt or in a bag. The microphones were attached to a device that was carried on the informant's neck, so that the microphones were located one at each respective side of the informant's head, close to the ears. The cable connecting between the microphones and the recorder was hidden beneath the informant's clothes. This way, the recording hardware would not attract any attention of either interlocutors or the surrounding people. This, indeed, proved to be the case in most instances.

The Israeli law does not prevent recording of a third party by a person who either takes part in the conversation or where it is clear that the speaking individual is aware of the attendance of that person. Whereas the recording informant signs a consent form allowing the *CoSIH* project to use the recorded data for research purposes, the other recorded people do not. Still, we are concerned with keeping the privacy not only of our informants but also of their interlocutors. Therefore, our own obligation, expressed explicitly in the consent forms as in other written forms handles to our informants, is to erase personal names and other betraying data of either the informants or their interlocutors from both the transcripts and the respective sound data.

Eliminating personal names in transcripts is an easy task, and the procedure taken is replacing the names with other names that are similar in form and in their socio-cultural setting. Being a multi-cultural nation, Israel has diverse traditions of name giving. Also, name giving to the newborn is a matter of changing fashion, and can indicate age and origin of the person carrying that name. As for name elimination in the sound files, this is a more complicated matter. One way of doing this is putting a weak beep instead of the name. However, this cause problems in understanding, especially in discourse passages that include many names. Therefore, we have devised an alternative method in which only the consonants are eliminated, so that both the syllable structure and the prosody remain intact. This method is still under examination.

Apart from this procedure, we allow informants to object to the inclusion of any part of the recording retroactively, as well as refraining from handing down to us anything they deem sensitive, or even all the recorded materials.

The last issue to be dealt with in this section is awareness to recording. This is an important matter to look at, be it on the part of the informants' interlocutors, in case they know about the recording, and especially on the part of the informants themselves. Change in speech form can take place in front of any microphone, all the more so if the recorded person knows that the goal of the research is linguistic study. We tried to overcome this latter problem by avoiding preliminary awareness of the linguistic goals of the research. When an informant is approached by our representative, s/he is being told that the goal of the research is "recording the daily life of Israeli inhabitants". Although this is not the whole truth, it is the truth, and nothing but the truth. When our representative comes to collect the recordings and before working on the sociolinguistic questionnaire, then our representative tells the informant that the recordings will be used for the compilation of *CoSIH* and requests the informant's consent to use the data.

Our impression is that in most cases speech style and language use is very similar all the way. This will, however, have to be checked in a thorough linguistic research. We had asked our informants to try tell us orally during the recording any information we could use later about the interlocutors or the setting of the recording at any new session. Only some informants kept to this procedure. We will have to double-check the wisdom of this procedure. Of course, whenever informants refer to their being recorded or recording, wherever meta-language is used to describe settings and circumstances of the recording or the recorded interlocutors, this not what we would like to see as part of the natural linguistic behavior of our informants. It is a matter to decide whether such chunks can be included as an integral part of the corpus, albeit in a separate section.

(3) Contextual Sampling (IJCL'01: §5.2)

CoSIH has been designed to be a fully representative corpus, integrating both demographic and contextual variables into a single database. Representativeness is achieved by sampling, and *CoSIH*'s design plan included a main corpus comprising 90% of the data of which both speakers' population and speech events will be selected randomly.

Obtaining a representative sample of the individuals in a group, or society, is known and commonly used. The various forms and methods of survey sampling pride a good representation of the individuals in society. For *CoSIH* this will be done by the use of a statistical sample of the Israeli population (IJCL'01: §5.2.1). However, reaching the goal of having a fully representative corpus in contextual terms too is still a vastly unexplored area (for some examples of spoken corpora aimed at representativeness not only in demographic terms but also in contextual terms see IJCL'01: §5.2.1). By 'contextual sampling' we mean sampling time and speech situations in order to get a representative sample of speech events in various environments. Thus, contextual sampling involves three issues: (i) long-term time sampling; (ii) short-term time sampling; (iii) speech sampling.

(i) Long-term time sampling

By 'long-term time sampling' we mean sampling of the recorded sets, i.e., all daylong recordings made by our informants, throughout the data collection period. Season may well influence language use, definitely in the lexical domain, but also in other domains. This is notably expected to occur in the holidays seasons. As the data-collection period is expected to last throughout a whole year (IJCL'01: 175), we expect long-term time sampling to come as a byproduct of this procedure. Eventually, we may nevertheless have a slightly imbalanced sample.

One particular problem is recordings on Saturday, the Jewish Sabbath, and on religious holidays. First, Saturdays and religious holidays are not working days in Israel. Therefore, we will have problems in asking our representatives to go and ask people to start recordings on Saturday or on a holiday. Also, a high percentage of Israeli Jews (estimated to be anywhere between 20% and 50%) would not operate a recorder on Sabbath or on a holiday because of religious constraints or out of respect to tradition. Even if we eventually find techniques to overcome these problems, we should expect under-representation of Sabbath and holiday recordings, definitely among Jews with religious or traditional restrictions.

(ii) Short-term time sampling

By 'short-term time sampling' we mean drawing a sample of 5,000-word units from each of the recruited daylong recordings. Language use change along the day, notably due to change in environment and interlocutors, but perhaps also due to other reasons, which one cannot predict at this time, like fatigue, attentiveness, and even mood.

The sampling procedure of recorded segments will be a statistically representative selection of one-hour recorded segments from each 24-span recording made by each individual informant. This will follow a procedure of elimination of long silent periods and long unintelligible speech passages (IJCL'01: §5.2.2). Hopefully, as with long-term time sampling, time distribution among the hundreds of daylong-recorded sets will produce a good sample of time within the day. However, if on a large-scale pretest (which we aim at conducting at the beginning of our data-collection year) we will see that this procedure results in imbalance, we will try the following alternative sampling procedure: We will sample time points along all raw daylong recordings to see whether they are located in the midst of a substantial speech event. In case the answer is negative, we will try another time point, until we find one that fits our demands. This or another procedure will have to be checked in the pretest.

(iii) Speech sampling

This sampling procedure can result either in hour-long speech events, or, in the majority of cases, in recorded segments shorter than an hour. While these segments may include substantial materials for inclusion in the corpus. In order to obtain our one-hour recorded segments we will need, in these cases, to make another step in order to reach this goal. This will be done by collapsing shorter speech segments into a single one by further elimination of silent periods that are too short to be eliminated in the first procedure. By using this latter procedure we will have samples of both long and short speech events, which may well represent different types of speech and language patterns. It should be born in mind that distinction is made between sampling and analytical procedures, so that the requirements from sampling, although they may converge with the requirements set for compiling the analytical unit, viz., the cell, are not the same.

(4) The Concept of 'Cell' (IJCL'01: §5.1)

As an end product, *CoSIH* will consist of "cells". A "cell" is an analytical unit. A cell is the basic sociolinguistic unit of *CoSIH*. It should aid the user to conduct research based on sociolinguistic data supplied in the *CoSIH* database, data that include both demographic features and contextual settings of the textual data included and compare it to data of other cells.

Word count is a basis upon which the size of corpora is usually defined. We kept to this tradition, and in our initial design of *CoSIH* a cell was defined as a recorded segment designated to include 5,000 words of coherent continuous text. Each cell was meant to consist of one or more texts produced by one or more speakers classified according to both demographic and contextual criteria. To illustrate this, it was said that "a cell may include a single 5,000-word text extracted from a university lecture given by a female 50-year-old native Israeli speaker of Western-European origin or two face-to-face conversations between two 20-year-old soldiers of Russian origin, one comprising 2,000 words, the other 3,000; or a cell may consist of several shorter phone conversations between a boss and employees. In all of these cases, each of the included sections will be a coherent continuous text" (IJCL'01: 190). The selection procedure of textual data to be included within a cell was to be made from the one-hour recorded segments extracted from the daylong recordings at the sampling stage.

Based on our experience gained so far, the above setting looks unachievable. In our initial design we did pay attention to speech rate and to uneven distribution of contextual setting and text types among different types of the population. However, we did not give enough thought to the fact that simple sampling procedures will not yield the 5,000-word segments to conform to our strict demographic and contextual criteria, and that we will need more complex sampling procedures. Furthermore, our aim was to have each cell consist of the speech of individual speakers about whom we can supply precise and accurate sociolinguistic data, viz., our recording/recorded informants. Linguistic materials gained from the speech of any other recorded individuals, be they interlocutors of our informants or other people, although they may be suitable for general linguistic analyses, are not good enough for lectal investigations, either linguistic or sociolinguistic.

In most sampled one-hour segments from our pilot recordings, the informant (i.e., the recording person) did not speak enough to lend us our 5,000-word cell we strived for. In order to achieve this target of having 5,000 words from a single informant, we would have to sample much longer recorded segments. Needless to say, one does not speak in empty space, and having the context of one's speech is part and parcel of any speech event. Since the corpus will eventually present the texts in both sound and transcription, we will therefore need to transcribe a lot more than originally expected. Anyone who has ever experienced natural spontaneous language transcription will know that this is not an achievable goal.

We have therefore changed our definition of cell to include segments of speech events of the same contextual category consisting of 5,000 words by all people taking part in these speech events. This definition is subject to one restriction: any cell must include at least 1,000 words in substantial speech uttered by *CoSIH*'s recording informant (or informants sharing the same demographic criteria). By 'substantial speech' we mean that the speech of the informant will not include only brief replicas with no linguistic significance. This change still fits the requirements set in the original design as regards cell capacity in terms of enabling linguistic and sociolinguistic research. As referred to in our IJCL paper (p. 194 n. 5), Biber, relying on linguistic-feature counts conducted on 1,000-word textual sub-samples of three of the early English corpora (both written and spoken), concluded that "the 2,000-word and 5,000-word texts in the standard corpora are reliable representatives of their respective text categories for analyses of this type" (Biber, 1990: 261). It may be noted at this juncture, that although spoken Hebrew is more analytic than the written medium, still the highly synthetic nature of Hebrew and its concise written structure will result in larger chunks than its English parallel by 25% or so.

There are some sets of recordings in the pilot sample that do show an intensive participation of the informant in many of the recorded speech events. From such sets we expect to have at least 2,000 word of the informant within a 5,000-word sample. In the long run, we may consider having a subcorpus of *CoSIH* with all cells consisting a minimum number of 2,000 words of their informants. If so, we will aim at achieving representativeness also in this subcorpus. Table 2 will serve to illustrate some types of speech events with varying percentage of participation of the recording informant. All recorded segments are of an identical length of 30 minutes each.

Recorded segment	**Speakers**	**Total turns**	**Informant turns**	**informant turns %**	**Total words**	**informant words**	**informant words %**
1	3	591	139	23.5%	4301	657	15.3%
2	4	616	121	19.6%	3625	756	20.1%
3	4	329	120	36.5%	2788	1102	39.5%
4	3	513	223	43.5%	4038	1667	41.3%

Table 2: Participation of informants in speech events

One last issue that has not been investigated yet is representativeness in terms of contextual variables. It has been mentioned above that *CoSIH* has been designed to be a fully representative corpus, integrating both demographic and contextual criteria into a single database. Contextual sampling as described above will show the distribution of speech events of varying types among the Israeli population. The collection of recordings as sampled is expected to result in deficiency in contextual categories, which will manifest itself during the procedure of allocation of texts into cells (IJCL'01: §5.3). This may lead to a decision to enhance the corpus by over-representation of some texts of varying contextual variables. This will provide better representation of contexts at he cost of being less representative of times or speakers.

Our contextual categories include three main variables and two secondary ones. The main variables are:

(a) Interpersonal relations: intimacy vs. distance
(b) Discourse structure: role driven vs. non-structured interaction
(c) Discourse topic: personal vs. impersonal

The secondary variables are:

(i) Active participants: monologue vs. dialogue
(ii) Medium: phone vs. face-to-face

A very brief survey of our pilot study already suggests that among dialogues we may expect a fair distribution of texts according to our designed main variables. However, we will probably be short of monologues. While the definition of naturally occurring monologue may be a matter for discussion, we do expect the need to over-represent monologues in some way.

As regards phone conversations, in most cases the person on the other side of the line is not heard at all. Special recording techniques may be used in some cases, e.g., with informants whose work involves many phone conversations. In other cases, a small subset of our corpus may be designed to bring forth telephone conversations.

Transcription and Annotation

CoSIH's designed size, five million words, requires some serious limitations as regards project duration, human power and financing, since five 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 and 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. At this point, our estimate is an average of 250 hours of transcription labor per one hour of recording. 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. Still, in order to illustrate phonetic variation of spoken Israeli Hebrew, we aim at including small samples from each cell in phonetic transcription (IPA). For some notes on transcripts in Hebrew orthography see Izre'el (2004).

The method of transcription follows in principle the one developed by Du Bois et al. (1992, 1993), in that it makes a visual representation of intonation units and includes some annotation of final tones. We are still considering the best annotation system and transcription principles for our texts. One should recall at this juncture that *CoSIH* aims at offering a synchronic presentation of sound and transcription in multimedia format. Some samples of transcribed texts have been published in Izre'el (2002). A preliminary analysis of Hebrew intonation units and final tones is presented in Izre'el (in press) and in Amir, Silber-Varod & Izre'el (2004a).

Finally, Hebrew standard orthography goes from right to left and, more prominently, does not include full and unambiguous representation of vowels. This last feature poses a serious obstacle to automatic analysis of the transcribed text. We are still contemplating the ways to overcome this problem in order to enable automatic analysis that will bring about a possible tagging service for *CoSIH*. A possible solution may eventually be found by adding a parallel pseudo-phonemic, broad transcription to the Hebrew text. A sample of a transcription of this kind can be viewed in Amir, Silber-Varod & Izre'el (2004b). This sample further includes glossing and English translation, two additional features of *CoSIH* that we have not yet given serious consideration.

A Final Word: What Is Next

The pilot phase of *CoSIH* is only the first step in a long road until our ambitious project is disseminated. Our next step is a pretest that will implement the lessons gained by the pilot study, examine their effectiveness, and study issues involved in large-scale informant recruiting and data collection. Unlike the pilot-collected data, data collected in the pretest phase will form part of the final corpus. We plan to use data from *CoSIH* Phase I, our pilot, to compile a mini-corpus on its own with its ca. 45 informants.

Acknowledgements

The pilot study was designed by the authors together with Benjamin Hary. It has been followed and supervised by the other launching members of the *CoSIH* project: Mira Ariel, John Du Bois, Yaakov Bentolila, Shmuel Bolozky, Otto Jastrow, Geoffrey Khan, Elana Shohamy and Eliezer Ben-Rafael. Vered Silber Varod has been a tremendous help in preparing the statistical data. We thank all those who, with care and enthusiasm, have taken part in this initial endeavor.

References

Amir, N., Silber-Varod, V. & Izre'el, Sh. (2004a). Characteristics of Intonation Unit Boundaries in Spontaneous Spoken Hebrew: Perception and Acoustic Correlates. In Speech Prosody 2004. [Scheduled for publication, March 2004.]

Amir, N., Silber-Varod, V. & Izre'el, Sh. (2004b). Characteristics of Intonation Unit Boundaries in Spontaneous Spoken Hebrew: Perception and Acoustic Correlates. A Sound Sample.
<http://www.tau.ac.il/humanities/semitic/sp2004.html>

Biber, D. (1990). Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation. Literary and Linguistic Computing, 5, 257-269.

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

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., Cumming, S., Schuetze-Coburn, 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, New Jersey: Lawrence Erlbaum Associates.

Izre'el, Sh. (2002). The Corpus of Spoken Israeli Hebrew (CoSIH): Textual Samples. *Leshonénu* 64, 289-314. (In Hebrew.)

Izre'el, Sh. (2004). Transcribing Spoken Israeli Hebrew: Preliminary Notes. In D. Ravid, & H. Bat-Zeev Shyldkrot (Eds.), Perspectives on Language and Language Development. Dordrecht: Kluwer. [scheduled publication: 2004].

Izre'el, Sh. (in press). From Speech to Syntax — from Theory to Transcription. In M. Bar-Asher & Ch. Cohen (Eds.), Aaron Dotan Anniversary Volume. (In Hebrew.)

Izre'el, Sh., Hary B. & Rahav, G. (2001). Designing CoSIH: The Corpus of Spoken Israeli Hebrew. International Journal of Corpus Linguistics, 6, 171-197.

Svartvik, J. & Quirk, R. (1980). *A Corpus of English Conversation*. Lund: Lund University Press.

Websites

BNC. The British National Corpus: The Spoken Component.
<http://www.natcorp.ox.ac.uk/what/spok_design.html>

CoSIH: The Corpus of Spoken Israeli Hebrew.
<http://www.tau.ac.il/humanities/semitic/cosih.html>

Gateway for Corpus Linguistics on the Internet.
<http://www.corpus-linguistics.de/corpora/corp_spoken.html>

SFB 411. <http://www.sfb441.uni-tuebingen.de/c1/corpora-engl.html>