THE PREPARATORY MODEL OF *THE CORPUS OF SPOKEN ISRAELI HEBREW (CoSIH)*[1]

Benjamin Hary
Emory University

Shlomo Izre'el
Tel Aviv University, Israel

Rationale

The study of Semitic languages has always been based on empirical research. From the outset of Semitic linguistic studies, medieval Hebrew and Arabic grammarians based their studies on corpora. The most famous corpus has always been the Hebrew Bible. Sa'adya Ga'on, a tenth-century Jewish grammarian and philosopher, based his grammatical treatises on the Bible. Centuries later, European scholars compiled concordances of the Hebrew Bible. Dictionaries of Hebrew and other Semitic languages are predominantly based on written corpora, but also, albeit more rarely, on spoken corpora.

---

[1] This article is a description of the CoSIH model presented at the conference in February 2000. It is to the version presented here that the papers by Rahav and Werum refer. Since then, some changes have been made in the CoSIH design. For the latest model see Izre'el, Hary, Rahav, 2001 and 2002.

Project Team—Core team: Shlomo Izre'el (project director); Benjamin Hary (principal investigator); John Du Bois (corpus analyst); Mira Ariel (discourse analysis and pragmatics); Giora Rahav (statistics and sociology). Advisory team: Eliezer Ben-Rafael, Tel Aviv University (sociolinguistics - sociological aspects); Yaakov Bentolila, Ben Gurion University (sociolinguistics - linguistic aspects); Otto Jastrow, Universität Erlangen-Nürenberg (transcription, phonology, dialectology); Shmuel Bolozky, University of Massachusetts at Amherst (phonology, morphology); Geoffrey Khan, Cambridge University (syntax); Elana Shohamy, Tel Aviv University (language education).

In non-Semitic languages, linguistic studies have also been based on empirical research. An elementary form of corpus-based methodology was frequently used: Studies of child language based on written parental diaries in the late nineteenth and early twentieth centuries and field linguistic methods used by Franz Boas (1940) and later linguists of the structuralist school are mentioned by McEnery and Wilson (1996: 2-4) as good examples of what can be termed early corpus linguistics. With the advent of Chomskian theories, less emphasis was placed on empirical observations, and research into language structure focused on the subjective study of one's own language, rather than objective corpus collection. Approaches based on Chomsky's theories, which are still considered mainstream in present-day linguistics, do not address certain aspects of language study, notably linguistic variation. Variation in language is observable in geographical and ethnic dialects, in written and spoken varieties, and in distinct speech patterns among different genders, socioeconomic classes, age groups, professions, and the like. Hence, theoretical and descriptive linguists nowadays regard linguistic variation as an inherent feature of human language.

During the last two decades, much attention has been given to the methodology of corpus linguistics. Using this methodology, linguistic description and theory are based upon statistical performance measures and observation of language use in real life. The first task is the compilation of a body of texts, a corpus. Recent developments in computer science and the enormous increase in computer storage capacity have greatly enhanced corpus studies throughout the world. A few examples of extant corpora are:[2]

First Generation Corpora

- The London Lund Corpus (spoken British, based upon The Survey of Spoken English that was launched in 1959) <http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX.HTM>
- The Brown Corpus (written American, 1961) <http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM>
- The Lancaster-Oslo/Bergen Corpus (written British, based on materials from 1961) <http://khnt.hit.uib.no/icame/manuals/lob/INDEX.HTM>

[2] See also Edwards 1993: chapter 10; <http://www.ruf.rice.edu/~barlow/corpus.html#Corpora>; http://www.icp.grenet.fr/ELRA/home.html.

Second Generation Corpora

- The British National Corpus (BNC) is a corpus of about 100 million of written (90 million words) and spoken (10 million words) British English, available on cd-rom. <http://info.ox.ac.uk/bnc>
- COBUILD - The Bank of English (British and American English, both written and spoken. Constantly growing corpus; now includes about 400 million words.) <http://titania.cobuild.collins.co.uk>

Recent Developments

- The International Corpus of English (ICE) is a collection of corpora, each representing spoken and written assortments from a national variety of English. Each corpus contains about 1,000,000 words. <http://www.ucl.ac.uk/english-usage/ice/index.htm>
- The Wellington Corpora of New Zealand English (written, spoken, and the NZ component of the ICE project; each contains one million words). <http://www.vuw.ac.nz/lals/corpora.htm>
- The Corpus of Spoken American English (1,000,000 words). <http://linguistics.ucsb.edu/research/sbcorpus/default.htm>
- Corpus of Written Estonian. <http://psych.ut.ee/gling/en/corpusb>
- The Swedish spoken language corpus at Göteborg University. <http://spraakdata.gu.se/lb/parole_org.html> <http://www.ling.gu.se/SLSA/SLcorpus.html>
- Corpora from Zimbabwe (Shona, Ndebele). <http://spraakdata.gu.se/lb/parole_org.html>

While corpora have been compiled and continue to be compiled for many languages all over the world, there is still no available comprehensive corpus for modern Hebrew. Moreover, research on modern Hebrew, especially on its spoken varieties, suffers greatly from the lack of descriptive studies, which is, among other issues, the result of a shortage of data (cf. Kaddari 1984). A corpus is a preliminary desideratum for much larger projects that cannot otherwise be achieved, be it a grammar of modern Hebrew, a comprehensive dictionary, or any other theoretical or applied inquiry. The research potential a corpus represents is enormous, and includes linguistic, cultural, sociological, and technological aspects, many of which will be listed

below. Voices have been raised in favor of compiling a spoken Hebrew corpus. Bentolila (1989) has described the corpus of Montreal French ("Le corpus Sankoff-Cedergren du français parlé à Montreal"), calling for a similar project for spoken Hebrew. In a review of Glinert's *Grammar of Modern Hebrew* (Glinert 1989), based on the grammatical judgments of six informants, Blau (1991) has also called for a comprehensive grammar based not only on competence judgments of a few native speakers, but also and mainly on a large corpus of both written and spoken varieties of the language. It is important to note here that introspection of the kind on which Glinert's grammar is based (aiming at reaching linguistic competence, rather than actual performance) cannot cope with a real, comprehensive analysis of a language that includes all its continuum of varieties. Kaddari (1996) noted the urgent need for compiling a corpus of the living literary language. Indeed, Yaacov Choueka is already in the process of compiling a computerized corpus of literary modern Hebrew. This corpus, *The Bar Ilan Corpus of Modern Hebrew*, includes (at the end of 1999) 26 million words in the following categories (Choueka 2000):

- Books (belles letteres and professional literature): 199 books with 11.5 million words;
- Meticulous journalism: 3 million words;
- Local newspapers: 3.5 million words;
- Knesset proceedings: 3.5 million words;
- Internet IDF radio news: 1.0 million words; and
- Daily newspaper news: 3.5 million words.

Hebrew was reintroduced as a full-fledged language in the twentieth century.[3] With the advent of the Zionist movement in the nineteenth century, large waves of Jewish immigration to Palestine resulted in the use of Hebrew as a spoken language. From meager beginnings in the late nineteenth century, Hebrew took its place as the common daily language among the Jewish population in Palestine, and as the national language of the newly established State of Israel in 1948. A hundred years of Hebrew speech have passed, and the academic world has lost a unique opportunity to record the emergence of a language as a full-fledged communicative system. Hebrew is still changing rapidly because of immense waves of immigration and swift

[3] See Izre'el's article in this volume.

changes in Israeli society. Consequently, a corpus of Israeli Hebrew is needed.

With this background in mind, we have decided to work toward the compilation of a comprehensive corpus of Israeli Hebrew. Since literary and most other varieties of the written language are accessible, there is no urgency to record them at this early stage of corpus compilation. Furthermore, compilation of a literary corpus of Hebrew is already under way (see above). Hence, we have chosen to begin this ambitious project by compiling a corpus of the spoken varieties of Hebrew.[4]

The Goals of CoSIH

1. To create a corpus of spoken Israeli Hebrew to facilitate research in a range of disciplines concerned with the Hebrew language and with the general methodology of Corpus Linguistics.
2. To disseminate this corpus publicly 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 CoSIH will be available to any potential user for her or his needs.

The Nature of the Corpus

The Corpus of Spoken Israeli Hebrew (CoSIH) will include a representation of most varieties of spoken Hebrew as it is presently used in Israel. We intend to include a representative sample of both demographically and contextually defined varieties. Demographic varieties are those associated with different groups of speakers: geographical, ethnic, socioeconomic, and social (age, education, gender, sexual orientation, profession, etc.). Contextual varieties refer to situationally defined varieties such as

[4] We have considered compiling, in addition to the spoken corpus, a sub-corpus of written texts, which may be regarded as semi-spoken in nature. The medium of computer correspondence and chats has expanded widely in recent years and no doubt will become more prominent in the future. This type of text is produced spontaneously and intuitively, much like the spoken medium. However, it also displays characteristics of the written language to such an extent that compiling a corpus of both spoken and semi-spoken language seems to be unwarranted, at least at this stage. Despite the fact that computer correspondence and chats tend not to be preserved, we have decided to consider the compilation of a corpus of semi-spoken Hebrew separately.

conversation (face-to-face, telephone) or speeches (spontaneous, prepared, and scripted).[5]

We should take into consideration the unique structure of Israeli society, which has a 1:1 estimated ratio between native and non-native speakers of Hebrew among the current Hebrew-speaking population. Because of this ratio, and the complex history of modern Hebrew, it is necessary to include within the corpus representative samples of all Hebrew speakers, native and non-native alike. Important Israeli figures (e.g., Nobel Literature prize laureate, S. J. Agnon, and Nobel Peace prize laureate, former prime minister Shimon Peres) are non-native speakers. Furthermore, Israeli society is constantly being augmented by a huge influx of immigrants, resulting in a highly transient linguistic structure, which must be recorded over time. In view of this situation, native speakers alone cannot accurately reflect what constitutes contemporary Hebrew as it is actually spoken. Limited, small samples definitely cannot reflect the complex sociolinguistic situation in Israel. Ignoring non-native speakers will distort most types of linguistic, especially sociolinguistic, research based upon the corpus.

It is also important to synchronize CoSIH as much as possible. Since by its very nature language, and all the more so Israeli Hebrew, is in constant flux, it is necessary to record CoSIH within a reasonably short time, and repeatedly over time.

As explained above, extant corpora vary in size. When a corpus includes written and spoken texts, usually the former comprises over two thirds of the corpus. This ratio is a distortion of the true statistical distribution of written and spoken language in real life. This distortion seems difficult to avoid, however, given the relative ease of collecting written texts. While CoSIH is intended to include only spoken transcripts, it should still be large enough to represent accurately the current linguistic state of Hebrew in Israel so that it can enable a variety of potential research. Consequently, the goal is to compile a corpus of five million words.

Large corpora may be presented in various forms. A corpus may be recorded, transcribed and translated in different ways. It may also be tagged, i.e., marked for syntactic, morphosyntactic, morphological, or morpho-phonological features. It may be recorded either by audiotape or videotape,

[5] These terms are equivalent to the terms "dialects" and "registers", respectively (for which see, e.g., Chambers and Trudgill 1998; Biber 1995).

and include spectrographic representations of the recorded materials. Each technique has its particular advantages and disadvantages. For example, while videotape recording may include extralinguistic features, it also has the disadvantage of distracting the informant, thereby reducing the possibility of natural speech. Video recording also demands far greater resources for the collection and transmission of a given amount of spoken language, relative to the collection of an equivalent number of audio recordings.

The planned corpus will consist of the following elements:

1. Digital audiotaped recordings;
2. Selected digital videotaped recordings;
3. Full synchronized transcripts in Hebrew orthography;
4. Phonetic transcription of selected paragraphs;
5. Glossing of selected paragraphs; and
6. English translation of selected paragraphs.

Potential Research and Applications

1. Linguistic research
   - 1.1. Theoretical Linguistics
     - 1.1.1. Hebrew language and linguistics
     - 1.1.2. General linguistics
   - 1.2. Applied linguistics
2. History of Hebrew
   - 2.1. Standardization and variation
3. Education
   - 3.1. Hebrew as a first language
   - 3.2. Hebrew as a second language
     - 3.2.1. For new immigrants
     - 3.2.2. For people abroad
4. Publishing
   - 4.1. Dictionaries and thesauruses
   - 4.2. Grammars
   - 4.3. Word frequencies
   - 4.4. Phraseology
   - 4.5. Translation
5. History and Identity
   - 5.1. Nationalism

5.2. Cultural diversity and pluralism
5.3. Intercultural communication
5.4. Language interference
6. Language engineering
6.1. Telecommunication
6.2. Machine-human interface
6.3. Automatic speech recognition

Corpus Size

The total size of the corpus will eventually reach 5,000,000 words, organized as 1000 cells. The cells will be recorded segments with about 5,000 words per cell. A representative 5% of the cells will be accompanied by a video recording. As explained below, both the number of cells and the number of words within each cell are subject to change according to criteria related to population representation.

A corpus of five million spoken words seems to be of sufficient size to represent well both the overall structure and the specific features of most linguistic varieties of potential interest.[6] Many of the extant spoken corpora include significantly less than five million words, and scarcely address the issue of true representation in their data collection. Larger corpora have addressed this issue rather broadly.[7] The compilation of a spoken language corpus larger than five million words, however, seems to be an unrealistic goal. Written corpora are relatively simple to design and can be compiled with relative ease using scanning techniques and Internet materials. Spoken language corpora, however, are more difficult to design and to distribute for use, due to complexities both in data collection and in data transcription.

[6] Based on linguistic-feature counts conducted with 1,000-word textual sub-samples of three of the early English corpora (both written and spoken), Biber (1990: 261) concludes 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."

[7] For example, the ten-million word spoken corpus of the BNC includes two equally sized parts: a demographic part, containing transcriptions of spontaneous natural conversations by members of the public, and a context-governed part, containing transcriptions of recordings made at specific types of meetings and events.

Representation

The goal is to produce a representative sample, which will not only take into account demographic criteria, but also account for contextual varieties. The data should therefore be collected according to two distinct types of criteria:

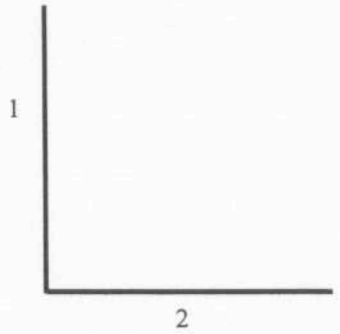
1. demographic criteria; and
2. contextual criteria.

The design of CoSIH intends to compile two complementary corpora: a main corpus and a supplementary corpus. The main corpus will form the bulk of CoSIH and will comprise about 95% of the entire collection. The supplementary corpus will include about 5% of the collected data.

Divisions of CoSIH:

mc = main corpus (95%); sc = supplementary corpus (5%)

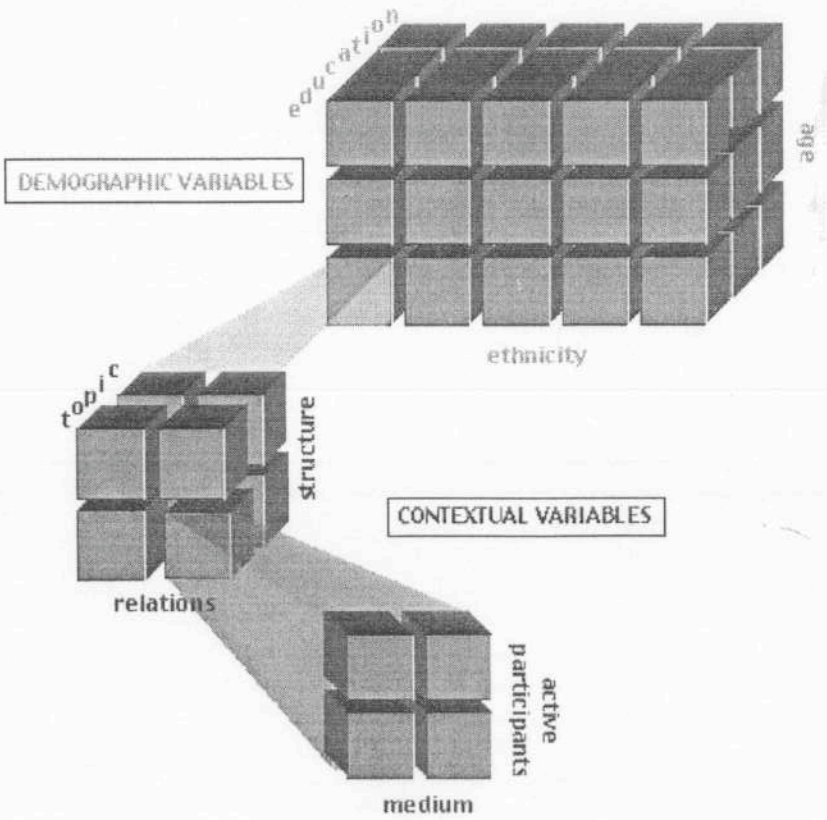
Recordings for the main corpus will be organized according to both demographic and contextual criteria. As a conceptual tool we will use a multidimensional cellular matrix, the main axes of which are demographic and contextual.



1 = Contextual Variables
2 = Demographic Variables

The two respective axes are themselves multidimensional: Each of the 45 (5x3x3) cells of the demographic matrix can potentially be augmented by 8 (2x2x2) distinct cells in the contextual matrix, with each of the latter cells multiplied by four (2x2) optional cells. More details of this structure are discussed below.

The supplementary corpus will be compiled according to contextual criteria, with some attention paid to demographic features.



CoSIH, including both the main and supplementary parts of the corpus, will be housed in an extensive database. Every researcher will be able to retrieve data attached to either distinctive variables, or to any combination of variables. For example, if researchers wish to retrieve data spoken only by young university-educated native speakers of Hebrew whose parents are of North-African origin, the data will be readily retrievable.

We must distinguish between sampling criteria and analytic criteria. The main challenge in sampling a population of over six million people[8] is to select valid criteria and, at the same time, limit the corpus to a manageable size. The task is even more complicated, because the relevant demographic criteria are further stratified according to a variety of contextual categories. While we plan to obtain a random sample of the population, further demographic and contextual information will allow a far more complex analytical approach.

While sampling according to statistical measures will be quantitatively representative of the entire population, collecting data according to analytic criteria is not necessarily statistically representative. A user of a database that includes a large number of sociological and sociolinguistic data will be able to retrieve linguistic data according to any sociolinguistic criteria programmed in the database. However, while data retrieved this way will be representative from the sociolinguistic point of view, they will not be representative quantitatively. In other words, drawing textual materials according to sociolinguistic variables at choice, if there are relatively many variables, may result in a small, unbalanced corpus of idiolects, rather than in a representative sub-corpus. As explained below, we will try to reconcile the virtually infinite variation of the Israeli Hebrew speech community with its representative corpus by categorizing variation according to both demographic and contextual terms.

In any case, it seems advisable to use sampling criteria different from those criteria that will later be used for the analytical process. Sampling procedures will take into account demographic data and will operate according to accepted statistical measures. Sampling populations is the task

[8] This figure includes the entire population of Israel. We will, however, exclude the language of children from the corpus. Research on child and youth language suggests that before age 16 the distinction between spoken and written ways of expression is not yet fully developed (Berman, in press; Berman and Ravid 1999). Hebrew child language has been extensively collected and analyzed by Ruth Berman of Tel Aviv University.

of statisticians and sociologists. All information regarding the process of sampling, while important for design, should be of little interest to the user. In contrast, criteria that will be applied in the analytical process, i.e., by people who will use the corpus, must be amenable to linguistic and sociolinguistic investigations. As explained below, the final design of CoSIH's main corpus will be different from the ideal design of the cellular matrix drawn above, and will depend on the textual sample to be collected randomly. While it will reflect the actual tendencies of language-type usage in any of the respective segments of the population, the design will meet the needs of end-user by defining analytical criteria that are applicable to linguistic investigation.

The Main Corpus

Analytical Criteria

The planned size of the entire corpus is five million words, and each of the represented cells is designed to include 5,000 words. This plan means that the entire corpus is designed for1000 cells of recorded text (transcript).

A cell is defined as the basic sociolinguistic unit of CoSIH. It is a recorded segment (or segments) in which the approximately 5,000 words comprise coherent, continuous text (or texts). Each cell is selected from recordings of people previously sampled according to demographic criteria, and will represent a single type of contextual setting. For example, a cell can consist of a single text produced by a 20-year-old Ashkenazi Jew with 12 years of education, recorded during a formal presentation. As another example, a cell might comprise the speech of several people of different sociolinguistic backgrounds engaged in conversation.

The main corpus, which will represent 95% of CoSIH, is pre-designed to include 950 cells. Of these cells, nine hundred are designed to accept randomly selected texts, dissected into segments. The segments will be assigned according to pre-designed demographic and contextual criteria, by a method explained below. In addition, we will collect data on 50 extra cells, following a non-statistical sampling, to represent special groups and specific contextual varieties which we hypothesize to be significant to the Israeli linguistic situation. For instance: Kibbutzim; ultra-Orthodox; gays and lesbians; people who have spent long time periods outside of Israel.

Demographic Categories

In sampling the population we strive to obtain sociolinguistic information about speakers who differ with regard to criteria such as place of birth; native/non-native status; ethnicity; place of residence; type of settlement (urban, rural, kibbutz, etc.); age; gender; socioeconomic status; profession; occupation; military service; religious affiliation; time spent outside Israel; and language(s) spoken at home. This sampling procedure can result in an unmanageably large corpus. Therefore, sampling the population for data will be controlled according to predefined and accepted statistical measures; a preliminary sociolinguistic interview with each informant will yield as many relevant sociological data for linguistic and sociolinguistic research. All the solicited data will be admitted into the sociolinguistic database.

Fragmenting the sample into too many subgroups will result in a mere collection of idiolects, instead of a workable corpus representing the entire speech community. Therefore, to enable retrieval of textual material that will be statistically representative of the Hebrew speaking population, the designed CoSIH database will structure the textual data according to a relatively small number of key variables.

The working hypothesis takes into account three major demographic criteria that we consider to be the most prominent for linguistic diversity in Israel: (1) Ethnicity, place of birth, and place of origin (EBO); (2) age; and (3) education. Each criterion is further specified with the following categories:

(1) Ethnicity, place of birth, and place of origin (EBO) (five categories):
   1. Jewish, Israeli born, father from Asia-Africa
   2. Jewish, Israeli born, others
   3. Jewish, foreign born, immigrated before or during 1965
   4. Jewish, foreign born, immigrated after 1965
   5. Non-Jewish: Muslim, Christian, Druze

(2) Age (three categories):[9]
   1. Young
      1a. 16-19 years old
      1b. 20-27 years old
   2. Middle-adulthood (28-50 years old)
   3. Adulthood (51 years old and above)

[9] See note 8 above.

(3) Education (three categories):

1. 0-8 years
2. 9-12 years
3. 13or more years

The demographic axis of the matrix is itself multidimensional. It includes categorization by EBO combined with information on age and education.

Contextual Combinations

Similar to the demographic axis, the contextual axis also consists of a multidimensional matrix. For the contextual combinations we have considered five variables (or categories), three main and two secondary.

Main Variables:

(a) Interpersonal relations: intimacy vs. distance (±intimacy). In variable (a), personal relations are reflected. When interlocutors are related personally, namely, when they are either relatives or friends, we identify it as +intimate.

(b) Discourse structure: role driven vs. interaction (±role driven). In variable (b) the structure of the conversation is considered. When the interaction is through a structured conversation, e.g., when there is a power role, we indicate it as +role driven.

(c) Discourse topic: personal vs. impersonal (±personal). In variable (c) the conversation topic is considered. If the topic deals with personal matters or daily issues, the conversation can be marked as +personal.

Secondary Variables:

(d) Active participants: monologue vs. dialogue (±monologue), and

(e) Medium: phone vs. face-to-face (±phone).

The three main variables (a)-(c) are indicated in all eight possible combinations (23 instances) as illustrated below. The secondary variables (d)-(e) are applied only to part of the matrix, since they occur much less frequently in most speech communities.

While monologue-type discourse may potentially be found in any of the recordings, variable (d) will be monitored only in two of the most prominent situations where monologues occur, i.e., in those cases where monologue rather than dialogue is an essential trait of the respective speech variety.

Telephone conversations will be admitted into the corpus only if they are distinctly different from face-to-face interaction.

Table 1. Classification Matrix of Contextual Combinations

Combination	a. Intimacy	b. Role	c. Personal	d. Monologue	e. Telephone
1	+	-	+	-	-
1e	+	-	+	-	+
2	+	-	-	-	-
2e	+	-	-	-	+
3	+	+	+	-	-
4	+	+	-	-	-
5	-	+	+	-	-
5d	-	+	+	+	-
6	-	+	-	-	-
6d	-	+	-	+	-
6e	-	+	-	-	+
7	-	-	+	-	-
8	-	-	-	-	-

Note: 'd' following a number indicates a monologue; 'e' following a number indicates a telephone conversation.

Examples of settings according to the classification matrix in Table 1:

1 family/friends normal conversation
1e family/friends normal telephone conversation
2 family/friends non-personal discussion, e.g., about politics
2e family/friends non-personal telephone discussion
3 traditional family normal conversation, e.g., about business
4 traditional family non-personal discussion, e.g., about politics; informal about a university class
5 therapy session; consultation with a rabbi
5d therapy session; story telling
6 business meeting; job interview
6d university speech; political speech
6e job interview via telephone; business telephone conversation
7 while waiting at a doctor's clinic
8 non-personal conversation between two customers at a supermarket

The resulting cellular matrix of the main corpus of CoSIH is multidimensional. An ideal corpus would be comprised of demographic representatives recorded in all contextual combinations. As this ideal does not exist, we have devised a mechanism where the contextual combinations are measured by the number of speakers and their potential influence on language use.

The demographic criteria consist of 45 combinations (5 EBO categories X 3 age categories X 3 educational categories). Any single demographic combination can be multiplied by the eight contextual combinations, and each of the resulting contextual combinations can potentially include four extra cells of the secondary variables (monologue or dialogue; face-to-face or telephone conversation). Since some of the contextual combinations cannot be elicited within some of the demographic combinations, we will modify the overall matrix accordingly.[10] We will assign relatively more weight to the contextual combinations that occur more often to speakers of the language, or that influence more the linguistic life of the community. The table below represents a hierarchy of the combinations defined in Table 1, and the number of corresponding cells in the corpus. The combinations at the top of the chart are the most typical, and will be represented in CoSIH by four cells each. The contextual combinations with minimal significance will not be represented in the bulk of the main corpus, but only among the extra cells annexed to the statistically represented cells in the sample.

Table 2. Corpus Cells Assigned for Contextual Combinations

Significance	Context Combination #*	Cells per Combination	Total cells in corpus
Maximal	1, 3	4	8
Major	2, 4, 6, 6b	2	8
Minor	1b, 2b, 5a, 6a	1	4
Minimal	5, 7, 8	separate representation	separate representation

*As defined in Table 1.

[10] For example, an adult Jewish person, Israeli born, whose father is of Asian or African origin with minimal education is unlikely to be recorded in contextual combination 1, since he or she is considered to be part of a traditional family.

In total, each of the 45 cells of the demographic matrix will include within itself 20 contextual cells (8+8+4). Altogether, these combinations comprise the 900 cells that form the bulk of the main corpus of CoSIH.

Sampling Approach

Demographic Sampling

The purpose of creating any of the represented cells is to provide sufficiently rich data for sociolinguistic and linguistic research. We will attempt to compile a corpus, which will not only capture the general structure of the language, but also represent linguistic variation.

Dividing the population into subgroups and sampling each of them independently ("stratifying" the sample) may be justified by several specific reasons. The most important reasons are the following:

1. The characteristics of certain sub-groups differ significantly from other sub-groups, particularly if these unique sub-groups are relatively small, or are not well mixed in the population, or both.
2. One or more of the subgroups is relatively small, yet it is important that the sample should not ignore it.
3. One or more of the groups is relatively small, yet we want it to be over-represented.

Not stratifying the sample according to a specific criterion does not mean that this particular criterion will not be represented properly, or that its significance would be neglected. It only means that we allow the random, statistical processes to reasonably represent it. Therefore, initially, the random sampling of the corpus according to residential areas will be applied.

While large segments of the population will be sampled adequately to obtain sufficient linguistic representation, there will still be parts of the population that will not be represented accurately. These parts of the population may include, inter alia, people residing in kibbutzim. This unique group comprises only a miniscule proportion of the population. If kibbutzniks (kibbutz members) comprised 0.2% of the population (which may even be an exaggeration) and the sample size is 900, an "ideal" sample would include 2 kibbutzniks (in fact, 1.8). Such a sample is evidently too small to investigate the language of kibbutzniks, especially because our aim is to include more than one contextual combination for each subgroup. While CoSIH may include some representation of such a rare speech community,

the cell or cells may only indicate the sort of approaches needed for conducting comprehensive research of this kind. For the investigation of such rare, specific subgroups, targeted corpora must be compiled separately. Researchers will be able to use CoSIH for comparisons, for obtaining general knowledge of the entire speech community. As importantly, CoSIH researchers can obtain preliminary information on the type of research needed for each of the particular targeted groups or types of speech. For the major linguistic groups, the five-million-word corpus will suffice for conducting comprehensive investigations into a variety of aspects of Hebrew sociolinguistics and linguistics, and of the Israeli speech community.

Linguistic Sampling

Obtaining a representative corpus in demographic terms is a known and commonly used procedure in sampling populations. This procedure will be followed by random sampling of the Israeli population. Creating a representative corpus in contextual terms, however, is still a vastly challenging undertaking.[11]

To achieve an accurate representation of linguistic data, both in terms of the demographic and of the contextual combinations, we will sample all the data randomly. This sampling will occur after having gathered all collected recordings from the sampled population. Each person, randomly selected for the demographic sample, will be asked to record all of his or her speech activities over three successive days. Each of these three-day-long recordings will be screened, removing segments of silence and passages of long unintelligible speech. From the remaining material, a one-hour recording segment will be randomly extracted.[12] The extracted material will form the basis for the main, statistically balanced corpus.

By following this procedure we hope to achieve reasonable representation of not only the speech population, but also the situations of natural speech that may vary according to context and according to time settings.

---

11 For some attempts towards achieving the goal of representing contextual combinations see Crowdy 1993: 262-263, Berglund 1999: §2.1, <http://info.ox.ac.uk/bnc/what/spok_design.html>, Biber 1995: §3, Čermak 1997: 190-191, Čermak and Sgall 1997: 19, and McCarthy 1998: §1.4.

12 The choice of a time segment rather than a number of words will enable us to make lectal comparison on the basis of speed of speech and other features.

Populating the Cellular Matrix while Satisfying Both the Statistical Requirements and the Analytical Strategies.

The conceptual design of CoSIH considers 20 contextual cells per single demographic combination. Since there are 45 demographic combinations, there is a total of 900 cells. Already during the conceptual design, not all combinations are compatible with each other. The EBO variable 3 (Jewish, foreign born, immigrated before or during 1965) is incompatible with Age variable 1 (16-27 years old) and partly incompatible with Age variable 2 (28-50 years old) regarding people younger than 37. Also, part of the sampled population under Age group 1 (those who belong to Age group 1a) is not compatible with people with high education (Education variable 3). Hence, already during the conceptual design, the number of filled cells will not reach the target of 900. Furthermore, we predict that there will be demographic varieties for which not all contextual combinations will emerge in the sample.

As discussed above, sampling and collecting the data, and arraying it for use are handled by completely different procedures. The first task in compiling the CoSIH, the collection of the recorded information in the form of textual data, will be handled by strict statistical measures, i.e., randomly.

The second task is the organization of the textual material for the intended use of all potential CoSIH users. We will take into account the carefully balanced demographic and contextual variable sets, or, as defined in Table 1, the cellular matrix. We will now have at our disposal (after transcription) randomly selected texts, produced naturally by randomly selected individuals in a variety of contextual situations. These texts will be distributed across the cellular matrix, each according to its respective demographic and contextual cell within the matrix. The textual material will be distributed into cells randomly and quantitatively, i.e., by dissecting the text of the randomly selected recordings into cells according to the actual inventory of textual data, rather than by fitting material and distributing texts into the pre-conceived matrix. This procedure will enable us to learn about the actual distribution of speech patterns in the population, and inform us about quantitative and qualitative patterns relative to both demographic and contextual features. In other words, the information distribution into cells will reveal what specific types of contextual varieties are actually used by the individual segments of the population, and the use ratio of the individual combinations relative to each other. As mentioned repeatedly, all solicited

sociolinguistic data will be available to interested users. To develop meaningful linguistic analysis, the users will have to retrieve data according to the set of variables available in the CoSIH.

Extra Cells

Obviously, there will be combinations, either demographic or demographic-contextual, that will not emerge in the sample at all. Others combinations will appear in tiny representation, inadequate for a substantial linguistic investigation. In most cases, such outcomes will represent the actual demographic strata of the Israeli speech community, or of the inventory of contextual situations used by the respective groups. In some cases, however, corrections will be needed, either due to certain flaws in the sample, or because of the need to over-represent a certain group. In the latter case, especially when there are reasons to believe that a certain group, or groups, has unique influence on the Israeli Hebrew linguistic behavior, an imbalanced sub-corpus will be formed. This sub-corpus will fill up the remaining 50 of the 950 cells of the main corpus. As perceived at this stage of planning, these cells will contain data from such groups as ultra-Orthodox, gays and lesbians, and people who spent long periods of time outside of Israel.

Special attention must be given to the language of the army. Obligatory military service in Israel is three-years for men and 21 months for women. Men serve additional time in the reserve forces, sometimes until the age of 49. Many people serve in the military, or in other defense forces, on a professional basis. Furthermore, Israel being a nation of immigration par excellence, this military service has often served as a melting pot of Israeli society. Due to its extreme significance in the life of the Israeli society, the military is notorious for having an enormous impact on Israeli Hebrew. The military influence is readily observable in the lexicon and phraseology, but definitely reaches far and beyond these domains. Therefore, a sub-corpus of recordings from the military will also be included in the main corpus. Whether it can be extracted from the random sample, or be formed separately, remains to be determined.[13]

[13] This plan depends on the type of sampling taken. Sampling by residential areas will result in a serious gap in informants from the military, who spend their time mostly outside of their home.

The Supplementary Corpus

While the design described above meets the representational requirements of most anticipated speech events, there are still several important domains of spoken varieties that are not covered by the matrix table. Nevertheless, they must be represented in the corpus of spoken Hebrew. These domains are linguistic varieties used in the Israeli parliament (the Knesset), in court, and in the mass media (television and radio). These speech varieties, although not part of the active language of the bulk of Israeli speakers, still have significant impact on the language, as large portions of the population are exposed to them. This is the reason to design a supplementary corpus, which will cover the above-mentioned varieties. This sub-corpus will consist of samples from the categories listed below. For each cell there will be a respective analysis, which considers whether further demographic cross-sectioning is necessary.

Table 3. Classification Matrix of the Supplementary Sub-Corpus

Supplementary Domain	Combination	Categories	Spontaneous	Non-sponta-neous	Non-sponta-neous
				Prepared	Scripted
TV	1a	non-sport broadcast	+		
	1b	non-sport broadcast		+	
	1c	non-sport broadcast			+
	2a	sport broadcast	+		
	2b	sport broadcast		+	
	2c	sport broadcast			+
	3	interview	+	+	
	4	talk show	+	+	

	5	movie			+
	6	commercial			+
Radio	7a	non-sport broadcast	+		
	7b	non-sport broadcast		+	
	7c	non-sport broadcast			+
	8a	sport broadcast	+		
	8b	sport broadcast		+	
	8c	sport broadcast			+
	9	interview	+	+	
	10	talk show	+	+	
	11	telephone program	+	+	
	12	commercial			+
Knesset	13a	speech			+
	13b	speech		+	
	13c	monologue dialogue	+		
Court	14a	speech			+
	14b	speech		+	
	14c	monologue dialogue	+		

According to Table 3, the supplementary contextual corpus is based upon 26 primary cells x 5,000 words, or 130,000 words, about 2.6% of the corpus. It is estimated that further division of the contextual corpus according to demographic measures will either double or triple this corpus, to comprise about 5.0-7.5% of the entire corpus.[14]

[14] In contrast to other spoken corpora (see above), the CoSIH design considers both demographic and contextual criteria for the main corpus. The need to supplement the main corpus with contextual varieties that are not represented in the main corpus design is minimal.

Conclusions

This chapter describes a plan for developing a new corpus of the spoken Israeli Hebrew, called CoSIH, and the design for its construction. The population representation in the corpus must be based on the internal social structure of the speech community for which it is designed. No research has been conducted yet on Israeli society in terms of demographic and contextual representation. Therefore, to achieve a suitable design for the corpus of spoken Hebrew, we plan to conduct the research and corpus compilation in seven consecutive phases (see Appendix). Each phase will enable us to evaluate the results of the previous phases, and will lead us to improve the quality and the final organization of the complete corpus.