Designing CoSIH: The Corpus of Spoken Israeli Hebrew *

SHLOMO IZRE'EL, BENJAMIN HARY
Tel Aviv University *Emory University*

AND GIORA RAHAV
Tel Aviv University

This paper describes the initial design of the Corpus of Spoken Israeli Hebrew (CoSIH). CoSIH will attempt to include a representation of most varieties of spoken Hebrew as it is used in Israel today. CoSIH is designed to 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 combined demographic and contextual design will be capable of showing the distribution of speech types in various subgroups of the population. 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. This 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. Nonetheless, 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 for other corpora with only slight modifications.

KEYWORDS: corpus design, spoken corpus, Israeli Hebrew

INTERNATIONAL JOURNAL OF CORPUS LINGUISTICS Vol. 6(2), 2001. 171–197

1. Introduction

The study of Semitic languages has always been based on empirical research. From the outset of Semitic linguistic studies, medieval Hebrew and Arabic grammarians have based their studies on corpora. In Hebrew studies, the most famous corpus has always been the Hebrew Bible. For example, Saadya Gaon, a tenth-century Jewish grammarian and philosopher, based his grammatical treatises on the Bible (see, for example, *Sefer ha-Egron*, which he wrote around 902 CE and *Sefer Ẓaḥut ha-Lashon ha-Ivrit*, which he wrote about fifteen years later [*Encylopaedia Judaica*, Vol. 14, col. 552–553]). So did later Hebrew grammarians. Centuries later, European scholars compiled concordances of the Hebrew Bible and based their grammatical description on the biblical corpus. Following this long-standing tradition, a limited number of grammatical treatises dealing with later periods of Hebrew were written, especially in the twentieth century. Furthermore, dictionaries and grammatical studies of Hebrew and other Semitic languages have been based on written and, more rarely, on spoken corpora (for the latter see e.g., the work done on Neo-Aramaic dialects [Jastrow 2002] and on the Modern South Arabian dialects [Simeone-Senelle 2002]).

Hebrew has one of the longest recorded histories among languages of the world. The earliest recorded texts go back to the beginning of the first millennium BCE. After over a millennium during which Hebrew was spoken and written, the language ceased to be used as a vernacular until the end of the nineteenth century. At that time, with the advent of the Zionist movement, large waves of Jewish immigration to Palestine resulted in the use of Hebrew as a spoken language. Hebrew was then reintroduced as a full-fledged 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 the national language of the newly established State of Israel in 1948.

A century of Hebrew speech has passed, and the scholarly world has lost a unique opportunity to record the emergence of a language as a full-fledged communicative system. Hebrew is still undergoing rapid change because of massive waves of immigration and swift changes in Israeli society. The reintroduction of Hebrew as a vernacular accelerated an unceasing flow of publications: dictionaries, grammatical studies, textbooks, and many others. All but a few of these publications have been targeted towards language learning.

To date, there have been only a limited number of lexical and grammatical studies based on actual linguistic usage. It would be interesting to investigate the linguistic historiography of research into Modern Hebrew. In any case, the extant studies, in clear contrast to the research conducted on previous layers of Hebrew, have not been based on a full-scale corpus. Consequently, a corpus of Israeli Hebrew is in sorely needed.

Yaacov Choueka of Bar Ilan University has made a promising start toward the compilation of a corpus of modern written Hebrew (Choueka 2000). This corpus, however, is still unavailable to the research community and does not claim to be representative. There has been a call for the compilation of 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, based on grammatical judgments of six informants (Glinert 1989), Blau (1991) also discussed the need to base a comprehensive grammar 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 to reach linguistic competence rather than actual performance) cannot cope with a real comprehensive analysis of a language that includes its entire continuum of varieties. Kaddari (1996) noted the urgent need for compiling a corpus of the living literary language.

While corpora have been and continue to be compiled for many languages all over the world (cf. Edwards 1993; Michael Barlow's website <http://www.ruf.rice.edu/~barlow/corpus.html>), there is still no corpus for modern spoken Hebrew. Moreover, research on modern Hebrew, and especially on its spoken varieties, suffers greatly from the lack of descriptive studies, which is in turn the result of a shortage of data (see Kaddari 1984). A corpus is a preliminary desideratum for larger projects that cannot otherwise be accomplished, such as a grammar of modern Hebrew, a comprehensive dictionary, or any other theoretical or applied research. The research potential of such a corpus is enormous, and includes, inter alia, applications in the following areas: general and theoretical linguistics, Hebrew language and linguistics, applied linguistics, language engineering, education, and cultural and sociological studies.

With this in mind, the Corpus of Spoken Israeli Hebrew (CoSIH) has been initiated. As literary and most other varieties of the written language

are always accessible, there is no urgency to record them at this early stage of corpus compilation. Furthermore, as compilation of a literary corpus of Hebrew is already under way (Choueka's project, see above), it is best to start this ambitious project by compiling a corpus of the spoken varieties of Hebrew.[1]

2. Goals

The goals of CoSIH are as follows:

1. To create a corpus of spoken Israeli Hebrew in order 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 such as CD-ROM or DVD-ROM, and will present the recorded sound simultaneously with its transcriptions and other extensions, linked together by software.

We should emphasize that CoSIH will be available to all potential users, either free of charge or at cost.

3. The nature of CoSIH

3.1 *Content*

CoSIH will attempt to include a representation of most varieties of spoken Hebrew as it is used in Israel today. It is intended to include a representative sample of both demographically and contextually defined varieties. Demographic varieties are those associated with different groups of speakers depending on their geographical location, ethnic grouping, socioeconomic and social status (age, sex, sexual orientation, education, profession, etc.). Contextual varieties refer to situationally defined settings which may affect linguistic varieties such as conversation (face-to-face, telephone), types of interaction (the interpersonal relations involved, the relevant discourse

structure), discourse topic, and types of speeches (spontaneous, prepared, or scripted).

It is also imperative to take into consideration the unique structure of Israeli society, which consists of 79.2% Jews and 20.8% non-Jews (14.9% Muslims; 2.1% Christians; 1.6 Druze; 2.2% others, according to a 1998 estimate). Of the Jewish population, 63.2% were born in Israel (27.4% whose father is Israeli-born, 21.2% whose father was born in Asia or Africa and 14.6% whose father is from Europe or America),[2] 11.7% were born in Asia or Africa, and 25.1% were born in Europe or America.

As such, the Israeli Hebrew speech community exhibits an unusual ratio between native and non-native speakers of Hebrew. Among the Jewish population alone, there are about 61% native speakers.[3] Adding the non-Jewish population, the ratio between native and non-native speakers of Hebrew is about 1:1. This, along with the complex history of modern Hebrew, makes it essential to include within the corpus samples of all kinds of speakers of Hebrew, non-native included. Many prominent Israeli figures, like the Nobel Literature prize laureate, S. J. Agnon, or the Nobel Peace prize laureate, former prime minister Shimon Peres, have not been native speakers of Hebrew, yet as dominant figures in the cultural and political life of Israel, their influence on the linguistic behavior is potentially high. Furthermore, the society is constantly being augmented by a huge influx of immigrants, resulting in a highly variable linguistic structure that should be recorded. Moreover, Arab citizens of Israel are increasingly demanding their fair share of the "Israeli pie," using Hebrew as a vehicle for their cause.[4] In view of this situation, native speakers alone cannot accurately reflect what constitutes contemporary Hebrew as it is actually spoken, and they definitely cannot reflect the complex sociolinguistic situation in Israel. Ignoring non-native speakers would have resulted in distorting most types of linguistic and especially sociolinguistic research based upon the corpus. Since by its very nature, language—and all the more so Israeli Hebrew—is constantly changing, it is crucial to record CoSIH within a reasonably short time.

3.2 *Size*

Extant corpora vary in size. When a corpus includes written and spoken texts, usually the former comprises over two thirds of the corpus. This, it seems, is a distortion of the statistical distribution between written and spoken texts

in real life. This distortion is difficult to avoid, given the comparable ease of collecting written texts. While CoSIH will include only spoken discourse, we would still like the corpus to be large enough to represent the current linguistic situation of Hebrew in Israel accurately so that it enables all kinds of potential research. Consequently, our goal is to compile a corpus of five million words.

The aim is to collect 1000 cells, or recorded segments, of 5000 words per cell (about thirty minutes of continuous speech). A cell of 5000 words seems to be large enough to enable reliable observations on linguistic structure.[5] 5% of the cells will be recorded by video.[6] As will be explained below, both the number of cells and the number of words in each cell will be subject to change according to criteria related to representativeness.

A spoken corpus of five million words seems just large enough to convey both the overall structure and specific features of most linguistic varieties represented within it. Many of the extant spoken corpora include much fewer than five million words and hardly address the issue of representativeness in their data collection. Larger corpora have addressed this issue rather broadly.[7] However, the compilation of a spoken corpus larger than five million words seems to be an unrealistic goal. Written corpora are easier to design and can be compiled with relative ease using scanning techniques and internet materials. Spoken corpora, on the other hand, are difficult not only to design but also to make available for use due to complexities both in data collection and in data transcription.

To conclude, CoSIH will consist of the following:

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

4. Representativeness

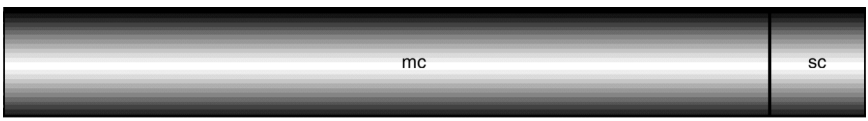
4.1 *The general design*

CoSIH is designed to include a representative sample of speakers and situations. Thus, data will be organized according to two distinct types of criteria: demographic and contextual. Moreover, the design of CoSIH is targeted at two complementary corpora: a main corpus and a supplementary corpus. The main corpus will form the bulk of CoSIH, and will comprise about 90% of the entire collection. The supplementary corpus will include about 10% of the collected data. Figure 1 shows the components of CoSIH.

For the main corpus, we will use a conceptual tool in the form of a multidimensional cellular matrix with demographic and contextual tiers. These two respective tiers are themselves multidimensional: each of the 45 cells of the demographic matrix (5 ethnicity[8] categories $\times$ 3 age categories $\times$ 3 educational categories) can potentially be augmented by eight distinct cells in the contextual matrix (2 interpersonal relations categories $\times$ 2 discourse structure categories $\times$ 2 discourse topic categories), where each of these latter cells may be multiplied by another four optional cells (monologue or dialogue, face-to-face vs. telephone conversation; i.e. 2×2). This structure is discussed in more detail below. Figure 2 shows the conceptual multidimensional matrix of the main corpus.

The supplementary corpus will include two distinct subcorpora, one to be based very much like the main corpus on demographic criteria, yet it will be compiled using non-proportional sampling. A second supplementary subcorpus will be compiled basically according to contextual criteria, with some attention paid to demographic features. Each of the distinct supplementary subcorpora will include about 5% of the entire corpus or 50 cells.

CoSIH, including both the main and supplementary corpora, will be housed in a large, sophisticated database. Every researcher will be able to



mc = main corpus (90%); sc = supplementary corpus (10%)

Figure 1. The components of CoSIH

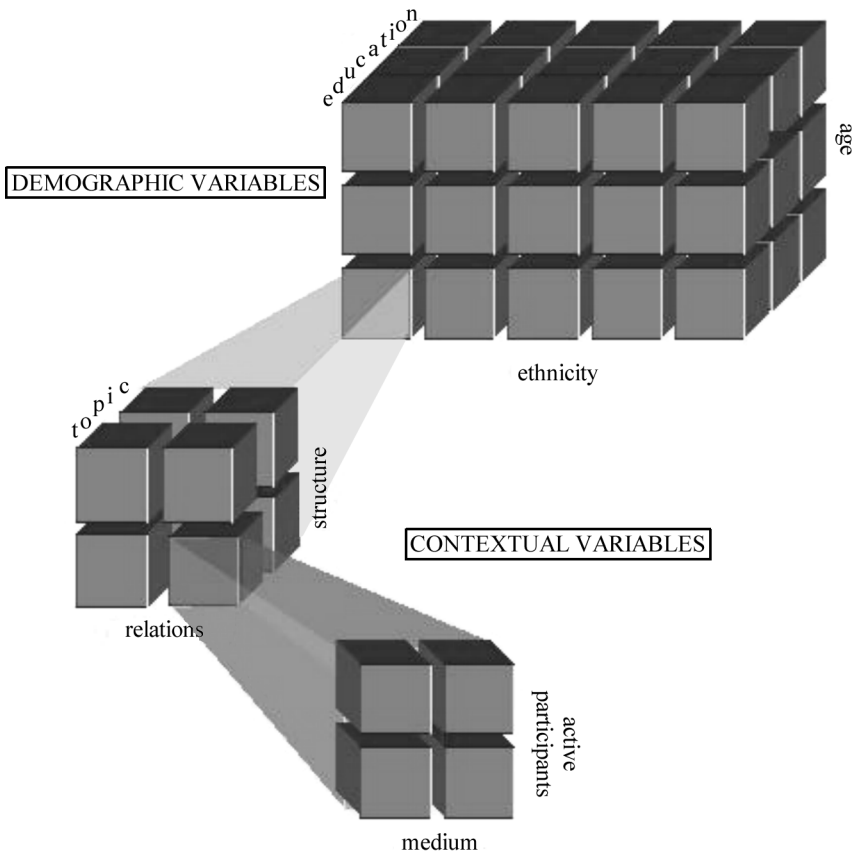


Figure 2. The conceptual multidimensional matrix of the main corpus

retrieve data attached to either distinctive variables or to any combination of variables. For example, researchers will be able to retrieve data spoken only by young university-educated native speakers of Hebrew whose parents are of North-African origin.

4.2 *Sampling and analytic criteria*

The distinction between sampling and analytic criteria warrents mention here. The main challenge in sampling a population of about 4.2 million[9] lies in

selecting valid criteria and, at the same time, keeping the corpus to a manageable size. The task is even more difficult because the relevant demographic criteria are further stratified according to a variety of contextual categories. While a random sample of the population is planned, further demographic and contextual information solicited from each informant will make possible a far more complex analytical approach.

A user of a database which includes a large number of sociological and sociolinguistic data will be able to retrieve linguistic data according to any sociolinguistic criteria installed in the database. However, while data retrieved this way will be representative from the sociolinguistic point of view, they will not necessarily be representative of any small segment of the population. In other words, drawing textual materials according to sociolinguistic variables of choice, especially if the latter are too numerous, may result in a small, unbalanced corpus of idiolects rather than a representative subcorpus. As will be explained below, CoSIH will attempt to reconcile between the infinite variation of the Israeli Hebrew speech community and its representative corpus by categorizing variation in both demographic and contextual terms.

All information regarding the process of sampling, while important for design, should be of no concern to the end user.[10] On the other hand, criteria that will be used in the analytical process, i.e., by those people who will use the corpus, must be amenable to linguistic and sociolinguistic studies. As will be explained below, the data for the main corpus of CoSIH will be collected randomly and later be prepared for the benefit of its users. The cellular matrix format should thus satisfy analytical requirements.

5. The main corpus

5.1 *Analytical criteria*

Totalling five million words, the entire corpus is to comprise 1,000 cells of recorded text, with each cell containing 5,000 words. 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 main corpus, which represents 90% of CoSIH, is designed to include 900 cells and accept randomly selected texts. In addition, data will be collected for 50 extra cells in non-statistical sampling, representing special groups and specific contextual varieties which are hypothesized to be significant to the linguistic situation in Israel. This unbalanced subcorpus of 50 extra cells will be designed after the organization of the balanced corpus of 900 cells is set, and will form part of the 100-cell supplementary corpus (see below, Section 6.1).

5.1.1 *Demographic categories*

In sampling the population for linguistic or sociolinguistic research, the aim is to obtain sociolinguistic information about speakers. This information should be representative of the variability that exists in the real, full-scale linguistic community. Therefore, it should represent the variability that exists due to differences in place of birth, native/non-native status, ethnicity, place of residence, type of settlement (urban, rural, kibbutz, etc.), age, sex, socioeconomic status, profession, occupation, military service, religious affiliation, whether one has spent time out of Israel, and language(s) spoken at home. As explained above (Section 4.2), sampling the population for data recording will be conducted according to accepted statistical procedures, and will be targeted towards a manageable corpus of five million words. All sampled individuals, i.e., the informants, will be interviewed in order to obtain as many relevant sociological data as possible. All the solicited data will be registered in a sociolinguistic database. Fragmenting the sample into too many subgroups will result in a collection of idiolects rather than in a workable representative corpus of the entire speech community. Therefore, in order to enable retrieval of textual material which will be quantitatively representative of the Hebrew speaking population, or specific segments of this population, the CoSIH database will allow the division of the textual data according to a smaller number of variables.

In planning the corpus we have kept in mind a view of Israeli society as comprising several segments, which may be considered speech communities (although at the moment we can only hypothesize about their similarities and differences). Obviously, these speech communities differ in their diversity of demographic features[11] and must be dissected accordingly. CoSIH's working hypothesis takes into account the three major demographic criteria considered

to be the most prominent for linguistic diversity in Israel: (1) ethnicity, place of birth and place of origin (EBO); (2) age; (3) education.

1. Ethnicity/religion, place of origin, and place of birth (EBO) (five categories):
   1 Jews, Israeli born, father from Asia-Africa
   2 Jews, Israeli born, others
   3 Jews, foreign born, immigrated before 1965
   4 Jews, foreign born, immigrated since 1965
   5 Non-Jews (Muslims, Christians, Druze, others)

This rather rough categorization takes into account the ethnic division in Israeli society as a whole between Jews and non-Jews (variables 1–4 vs. variable 5). It further takes into account the difference between what is usually regarded as the major division in Jewish society between Ashkenazi and non-Ashkenazi Jews (variables 1 and 2);[12] distinguishes between native and non-native speakers of Hebrew among the Jewish community (variables 1–2 vs. 3–4); and suggests a dividing line between Jews who immigrated to Israel before and after 1965 (variables 3 and 4), separating immigration during the pre-state years and immediately following the establishment of the State of Israel in 1948 from that of the late 1960s and thereafter, also taking into account immigration frequencies.[13]

2. Age (three categories):
   1 Young (15–27 years old)
   2 Middle adult (28–50 years old)
   3 Senior adult (over 50 years old)

The CoSIH project will sample the Israeli population aged fifteen and above. Most Israeli teenagers begin high school at the age of fifteen. In order to get significant results from the quantitative point of view, it is suggested that the sampled population be divided into three groups roughly similar in size. In the analysis of the data, age group 1 may be divided further into two subgroups, as we hypothesize that there is a linguistic change around the time people leave high school and start their military service or studies as young adults (for the significance of the military in linguistic culture in Israel see below Section 6.1). Therefore, we would suggest that linguistic analysis further recognize two subgroups, one of teenagers (15–18 years old) and one of young adults (19–27 years old), as at age 18 people begin their military service. Age group (2) starts at about the age where young people build their

own families (the average age of marriage being 27) and ends at about the age where their children have grown up and start leaving home. Similarly to the case of the young-age group, we might also suggest that users of the corpus consider further distinction between two generations in the over-50 group.

3. Education (three categories):
   1 People who have not graduated from high school
   2 High school graduates
   3 College or university graduates

5.1.2 *Contextual categories*

As the type of language used in spoken discourse is mostly dependent on specific situations, spoken corpora should aim at capturing diverse types of contextual varieties of their respective languages. Some corpora have indeed been compiled with attention to contextual varieties. Different approaches have been adopted to conform with this requirement. For example, the ten-million word spoken component of the British National Corpus (BNC) consists of two equal parts: a demographic part and a context-governed part. The latter comprises four equal-size broad categories of social context: (1) educational and informative events; (2) business events; (3) institutional and public events; and (4) leisure events. Each category is divided into the subcategories of monologue (60%) and dialogue (40%),[14] with yet further subcategorization within each of the latter parts, taking into account topic of discourse and demographic criteria (Crowdy 1993: 262–263; Berglund 1999: Section 2.1; <http://info.ox.ac.uk/bnc/what/spok_design.html>).

While the mode of classification varies, actual settings or contexts rather than abstract notions are taken to be the most practical way of collecting spoken material (see e.g., the survey of register categories in four linguistic corpora in Biber 1995: Section 3). Indeed, Kennedy (1998: 71) offers a list of contexts as a guide for compiling corpora, although these are grouped according to two broad classes: monologue and dialogue. Within these two broad classes, further subclasses are defined, among which are formal and less formal (within "monologue"), face-to-face dialogue, telephone dialogue, and structured interaction.

Atkins, Clear and Ostler (1992) offer a different approach in an insightful article. The authors suggest that balance in a corpus cannot be achieved (if at all) without fulfilling two requirements: (1) taking into account both

external and internal criteria, or, in our terminology: contextual variables and linguistic variables (or text types; cf. Biber 1995: Section 1.2.1); (2) constant feedback from its end users. Accordingly, the authors offer a large list of attributes that may be documented for every text in a corpus. They suggest that these attributes be added gradually into the corpus database. This list includes criteria for both written and spoken corpora, and includes attributes such as mode of transmission (written, spoken etc.), constitution (a single or composite text), topic, age of intended readership, familiarity with intended readership, and many others.[15]

Still, some corpora have taken into account more general criteria for collecting spoken varieties. Such an approach is more theoretically oriented. As put by Milroy, among the most important contextual factors "is probably the speaker's psycho-social orientation to his or her conversational partner(s) on the dimension of *social distance* and *intimacy*" (Milroy 1987: 36).

Work done on the 800,000-word spoken component of the Czech National Corpus (CNC) has indeed taken into account more theoretically oriented criteria of classification. CNC is composed of both interviews and conversations. It has been suggested that the choice of speakers and the triggering questions take into consideration three factors: different degrees of formality or familiarity among speakers, different emotional situations, and situations of different power structures (Čermák 1997: 190–191; Čermák and Sgall 1997: 19). According to Čermák, in practice the full and graded scale of familiarity could not have been strictly followed, as it would require a much larger corpus and more time. Therefore, only two grades were taken into account: (1) a complete familiarity, or rather intimacy in conversations (both interlocutors had to know each other well); (2) a formal approach in interviews in that it included people who were not familiar, or rather not very familiar with the interviewer prior to the recording.

Another corpus that was designed according to more general conceptual categories is the five-million word CANCODE corpus (Cambridge and Nottingham Corpus of Discourse in English, which is part of the Cambridge International Corpus, developed and owned by Cambridge University Press). CANCODE focuses mainly on unrehearsed, non-formal speech. Emphasizing the need for classificational categories to be discrete and comprehensive, the CANCODE corpus distinguishes four types of relationship between speakers: intimate, socio-cultural, professional and transactional. CANCODE further distinguishes between three types of interaction: non-collaborative, collabo-

rative idea, and collaborative task. The four relationship categories together with the three interactional categories form a twelve-cell matrix unto which actual situations could be matched (Hudson, ms; cf. McCarthy 1998: Section 1.4).

The design of the contextual matrix for CoSIH is based on the theoretical premise that situational contexts depend on three main features: the relations between the speakers, the organization of the discourse, and the topic of the discourse. These will form three of the five variables of which the contextual matrix will consist (variables a–c). Two other variables are more technical and take into account the active participants in the discourse and its medium (variables d–e).

Main variables:

a. Interpersonal relations: intimacy vs. distance (±intimacy)
   Variable (a) reflects personal relations. When interlocutors have personal relations, namely when they are either relatives or friends, we identify the situation as +intimate.

b. Discourse structure: role driven vs. non-structured interaction (±role driven)
   In variable (b) the structure of the conversation is taken into account. When the interaction is structured in the conversation or when there is a power role, we indicate the situation as +role driven.

c. Discourse topic: personal vs. impersonal (±personal)
   In variable (c) the topic of conversation is considered. If the topic concerns personal matters or daily matters, the conversation is classified as +personal.

The above-mentioned three main variables (a–c) are indicated in all eight possible combinations (2^3 instances) as illustrated below, whereas the secondary variables (d–e below) are applied only to a part of the matrix since they are much less frequent.

Secondary variables:

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

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

While monologue-type discourse may potentially be found in any of the recordings, variable (d) will be monitored in only two of the most prominent situations where monologues occur, i.e., in those cases where monologue rather than dialogue is an essential aspect of the relevant speech variety. By the same token, phone conversations will be admitted into the corpus only in cases where they may make distinct text types from the respective face-to-face interaction. Table 1 shows the matrix of contextual varieties listed according to these abstract variables.

Table 1. Matrix of contextual varieties

	Intimacy	Role	Personal	Monologue	Telephone
1	+	−	+	−	−
1t	+	−	+	−	+
2	+	−	−	−	−
2t	+	−	−	−	+
3	+	+	+	−	−
4	+	+	−	−	−
5	−	+	+	−	−
5m	−	+	+	+	−
6	−	+	−	−	−
6m	−	+	−	+	−
6t	−	+	−	−	+
7	−	−	+	−	−
8	−	−	−	−	−

(Note: 'm' following a number indicates a monologue; 't' following a number indicates a telephone conversation.)

Examples for the matrix:

1 family/friends daily conversation
1t family/friends daily conversation on the telephone
2 family/friends non-personal discussion (e.g. politics)
2t family/friends non-personal discussion on the telephone
3 traditional family daily conversation; some business meetings
4 traditional family non-personal discussion (e.g. politics); informal university class
5 therapy session; consultation with a rabbi
5m therapy session; story telling
6 business meeting; job interview

6m university speech; political speech
6t job interview on the telephone; business telephone conversation
7 while waiting at a doctor's clinic; on a flight, between strangers
8 non-personal conversation between two customers at the supermarket

As mentioned above, the resulting conceptual context-based matrix of the main corpus of CoSIH is multidimensional. The demographic criteria consist of forty-five combinations (5 EBO categories × 3 age categories × 3 educational categories). Any single demographic variety can be multiplied by 8 contextual varieties, and each of the resulting contextual varieties potentially includes 4 extra cells of the secondary variables (monologue or dialogue, face-to-face vs. telephone conversation).

An ideal corpus would be comprised of demographic representatives recorded in all contextual varieties. However, some of the combinations do not exist in the real world. Therefore, some of the cells in the larger conceptual matrix will be empty,[16] giving more weight to those contextual varieties that are used by more speakers of the language, and thus can be hypothesized to have more influence on language use and on linguistic development. The chart below represents a hierarchy of the varieties listed above and the number of cells in the corpus, based on a working hypothesis that takes into account expected frequencies of contextual situations. Whereas the varieties at the top of the chart will be represented in CoSIH by four cells each, the varieties with expected very low frequency may not be represented at all. Table 2 shows the hierarchy of contextual variables according to expected frequency.

Thus, each of the 45 cells of the demographic matrix will include in itself 20 contextual cells. Altogether these combinations make up the 900 cells that form the bulk of the main corpus of CoSIH.

Table 2. Hierarchy of contextual variables according to expected frequency

Frequency	Varieties	Cells per variety	Total cells per single demographic cell
High	1, 3	4	8
Medium	2, 4, 6, 6t	2	8
Low	1t, 2t, 5m, 6m	1	4
Very low	5, 7, 8		not represented

5.2 *Sampling*

The purpose of creating any of the represented cells is to provide sufficient data for sociolinguistic or linguistic research. We will try to compile a corpus that will not only capture the general structure of the language, but will also be representative of linguistic variation to a reasonable extent.

The way to do that, we suggest, is by using a representative sample of the spoken language with its variations. In principle, this might be accomplished if we could sample each of the cells in our matrix. The combined sample would represent the language as it is spoken. In practice, however, this is impossible: speakers and speeches do not come tagged with the identification of the cell to which they belong. Moreover, we do not know the relative prevalence of speakers and speeches. The viable alternative, seems to comprise two stages: first, to draw a representative sample of Hebrew speakers in the country; second, to draw a sample of the speech of each speaker.

5.2.1 *Demographic sampling*

Drawing a sample of speakers is relatively easy. One way would be to draw a sample of the names and addresses of all residents, and to screen out those who do not speak Hebrew. According to statistical theory, this process alone would yield a representative sample, provided that the sample is large enough. Alternatively, we may draw a random sample of residences (apartments, homes, etc.) and then sample one person from each. This procedure is technically simpler. It also allows a simple control over the regional distribution of the subjects: we can sample from each region a random sample of the residents, with the number of subjects proportional to the population of this region.

While large sections of the population will be sampled in enough numbers to obtain sufficient material for linguistic representation, there will still be parts of the population that will not be represented accurately. For example, sections of the population may include people residing in Kibbutzim—a unique group which consists of only a miniscule proportion of the population. If Kibbutzniks actually comprise 0.2% of the population (which may even be an overestimation) and the sample size is 900, an ideal sample would include 2 individuals (in fact, 1.8). Such a sample is obviously too small for conducting research on the language of Kibbutzniks, especially as our aim is to

include more than one contextual variety for each subgroup. This means that while CoSIH may include some representation of such a speech community, this will be adequate only for the representativeness of the whole sample. It will not provide an adequate sample for the particular subgroup. As a result, the sampled corpus may not include enough data to enable a thorough linguistic or sociolinguistic study of small subgroups. For this specific kind of investigation, targeted corpora must be compiled separately.[17] Researchers will be able to use CoSIH for comparisons, for obtaining general knowledge of the entire speech community, or, far more importantly, to obtain preliminary information on the type of research needed for each of the individual targeted groups or types of speech. Still, for the major linguistic groups, the five-million word corpus should suffice.

5.2.2 *Linguistic sampling*

Obtaining a demographically representative sample is a known and commonly used procedure in sampling populations. As explained above, this will be done by the use of a statistical sample of the Israeli population. However, reaching the goal of having a fully representative corpus in contextual terms also 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 Section 5.1.2 above).

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 span of time of 24 hours. This span of time will be distributed homogeneously among the informants. Ideally, seven equal one-day temporal units will start respectively on a different day of the week. 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 one-hour recording segment will be randomly extracted.[18] This will form the basis for the main, statistically balanced corpus.

By following this procedure we hope to achieve reasonable representativeness of not only the population, but also of the situation of natural speech, which may vary according to context as well as according to time settings. One other issue to be raised at this juncture is the production of speech by

individuals, which is uneven: some speakers speak very little, some speak much of the time. Accordingly, we will have to make a strategic decision to emphasize either the representativeness of speakers or the representativeness of speech. If we take an equal number of words from each speaker without regard to his or her relative speech production, the representation of speakers will be adequate, but texts produced by "heavy" speakers will be underrepresented. If we emphasize the accuracy of representing the speech, then the "light speakers" will have to be underrepresented. Whichever strategy we choose, a careful research protocol would allow anybody to follow the speech patterns of speakers or of speech by simple weighing procedures. Having in mind the compilation of a corpus which should represent language and linguistic variation more than speech habits, we are planning to follow the first alternative and give equal weight to speech patterns as used by speakers.

5.3 *Reconciliation between statistical requirements and analytical strategies: filling in the cellular matrix*

The conceptual design of CoSIH has 20 contextual cells for each of 45 demographic varieties allowing for 900 cells. At the level of the conceptual design, however, not all varieties are compatible with each other. Thus EBO variable 3 (Jews, foreign born, immigrated before 1965) is incompatible with age variable 1 (15–27 years old) and partly incompatible with age variable 2 (28–50 years old), as regards people younger than 35. Also, part of the sampled population of age group 1 cannot include people who have had higher education (educational variable 3). Thus, at the design stage we know that the amount of filled cells will not reach the target of 900. Furthermore, we predict that there will be demographic varieties for which not all contextual varieties will emerge in the sample.

As we have seen above, there are completely different procedures for collecting the data and putting it to use. The first step in the compilation of CoSIH will thus be the collection of the recorded textual data from a statistically representative sample. The second step is the organization of the textual material for use by all potential users of CoSIH. In this process, we will take into account the carefully balanced demographic and contextual variable sets, or, as we have called it before, the cellular matrix. We will now have at our disposal randomly selected texts produced by randomly selected individuals in a variety of contextual situations. These texts will be

distributed in the cellular matrix according to their respective demographic and contextual cells within the matrix. The allocation of the textual material into cells will thus yield cells with unequal frequencies of texts, which in turn will represent the actual frequency of speaker-situation texts in the speech community. This procedure will enable us to learn about quantitative and qualitative patterns as relating to both demographic and contextual features. In other words, the relative frequency of the cells will be informative as to what types of contextual varieties are actually used by the individual sections of the populations, and the ratio of use of the individual varieties in relation to each other. Thus, 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. As mentioned above, all solicited sociolinguistic data will be available to the user upon searching these cells, however, in order to make the linguistic analysis meaningful, the user may prefer to retrieve data according to the CoSIH set of variables.

6. The supplementary corpus

As mentioned above, the supplementary corpus will consist of two equal-sized parts; one demographically based and one contextually based.

6.1 *The demographically based supplementary subcorpus*

Some combinations, either demographic or demographic-contextual, will not emerge in the sample; others may emerge too infrequently to accomodate substantial linguistic investigation. In most cases this outcome will be representative of the actual demographic strata of the Israeli speech community and of the inventory of contextual situations used by the respective groups. In some cases, however, corrections will be needed, either due to specific flaws in the sample or because of the need to over-represent one group or another. In the latter case, especially when there are reasons to believe that a certain group or groups have a special influence on Israeli Hebrew linguistic

behavior, an unbalanced subcorpus will be formed, with 50 additional cells. As can be perceived at this stage of planning, such cells will include data from such groups as the ultra-orthodox; gays and lesbians; and people who have spent long periods of time outside of Israel.

Special attention is due to the language of the military. Obligatory military service in Israel is three years for men and twenty-one months for women. Men serve further time in the reserve forces, sometime until the age of 49.[19] Many more people serve in the military or in other security forces on a professional basis. Since Israel is a land of immigration par excellence, military service has always served as a melting pot for Israeli society. Moreover, due to its extreme significance for Israeli society, the military is known to have had an enormous impact on Israeli Hebrew. This is mostly observable in the lexicon and phraseology, but definitely goes far beyond these areas. Therefore, the main corpus will also include a collection of recordings from the military. Whether it can be extracted from the random sample or must be formed separately remains to be seen.[20]

6.2 *The contextually based supplementary subcorpus*

While the above design meets the needs of representativeness of most speech events, there are still some important domains of spoken varieties that this matrix does not cover. Nonetheless, they must be represented in a corpus of spoken Hebrew. These are linguistic varieties used in the Israeli parliament (the Knesset), in court, and especially in the electronic media (television and radio). Such varieties, although not part of the active language of the bulk of Israeli speakers, still execute a significant impact on the language, as large portions of the population are exposed to them. This is why we have designed a supplementary corpus that will contain the above-mentioned varieties. This corpus will consist of samples from the categories listed below. Each cell will undergo a check to see whether further demographic cross-sectioning is necessary. Table 3 shows the contextual varieties chosen for the contextually based supplementary subcorpus.

According to the above, the contextually based corpus is based upon 26 primary cells $\times$ 5000 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 enlarge the size of this subcorpus. In addition, texts representing the oral language of outer media such as lyrics, movies,

Table 3. Contextual varieties

			spontaneous	nonspontaneous	
				prepared	scripted[21]
TV	1a	non-sports broadcast	+		
	1b	non-sports broadcast		+	
	1c	non-sports broadcast			+
	2a	sports broadcast	+		
	2b	sports broadcast		+	
	2c	sports broadcast			+
	3	interview	+	+	
	4	talk show	+	+	
	5	movie			+
	6	commercial			+
radio	7a	non-sports broadcast	+		
	7b	non-sports broadcast		+	
	7c	non-sports broadcast			+
	8a	sports broadcast	+		
	8b	sports broadcast		+	
	8c	sports broadcast			+
	9	interview	+	+	
	10	talk show	+	+	
	11	phone-in program	+	+	
	12	commercial			+
Knesset	13a	speech			+
	13b	speech		+	
	13c	monologues; dialogues	+		
court	13a	speech			+
	13b	speech		+	
	13c	monologues; dialogues	+		

the theater (original and translated), standup comedy and other performances will be compiled, as some of these texts are known to have had longstanding influence on the linguistic culture of Israel. The goal of the contextually based corpus is 250,000 words organized in 50 cells, comprising 5% of the entire corpus.

7. Conclusion

This paper describes the initial design of The Corpus of Spoken Israeli Hebrew (CoSIH). To the best of our knowledge, this is the first attempt to construct a single, unified design of a representative corpus with both demographic and contextual variables taken into account according to acceptable statistical and analytical criteria.

The combined demographic and contextual design will be capable of showing the distribution of speech types in various subgroups of the population. This design is culturally dependent to suit the special structure of the Israeli Hebrew speech community. As such it includes both native and non-native speakers of Hebrew, emphasizing the special structure of Israel as an immigrant society par excellence, takes into account the large Arab minority, and pays special attention to such unique speech communities as the Israeli military forces.

We intend to compile CoSIH within a relatively short time in order to capture a synchronic picture of Israeli Hebrew. Still, its design will hopefully serve as a basic template for future corpus compilations. This template can be modified relatively easily to suit the compilation of sectional corpora within Israeli society. Furthermore, it is hoped that the principles upon which this design is based upon can be modified to suit speech communities outside Israel as well.

Notes

* This paper was composed by the above mentioned authors in collaboration with John Du Bois (University of California at Santa Barbara) and Mira Ariel (Tel Aviv University), whose contribution is especially prominent in the design of the contextual matrix. We also thank the other members of the CoSIH project team for their insightful comments. 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).
We are also indebted to Regina Werum, who has helped us with several sociological issues. We thank František Čermák for information on the Czech National Corpus and

Jean Hudson for her kind permission to cite from her unpublished paper (Hudson, ms). Izre'el and Hary have lectured on CoSIH and its design from various platforms, and have received useful comments and suggestions from many people. Special thanks are due to Elena Tognini Bonelli and to John Sinclair of the Tuscan Word Centre for their guidance and constant support.

1. We have considered compiling in addition to the spoken corpus a subcorpus of written texts that may be regarded as semi-spoken in nature. The medium of computer correspondence (e-mails) and "chats" has been widely enhanced in recent years and no doubt will become more and more prominent in the future. This type of texts is produced spontaneously and intuitively very much as in the spoken medium. However, it also has 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.

2. While questions may be raised as to the choice of father rather than mother or both parents, we suggest that the category of persons born in Israel be defined according to the father's country of birth. This is the way current population data has been gathered, and by following this pattern we facilitate comparison of our data with theirs.

3. People who immigrated to Israel in their formative years as regards language acquisition may form a borderline between native and non-native speakers with regard to their linguistic skills.

4. See, for example, the important influence of Anton Shammas' *Arabesques* or the Hebrew spoken by several Arab members of Knesset (the Israeli Parliament).

5. Based on linguistic-feature counts conducted on 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."

6. Audio and video recordings have their respective advantages and disadvantages. For example, while videotape recording has the advantage of including extralinguistic features, it also has the disadvantage of drawing too much of the speakers' attention, thus reducing the possibility of achieving naturalness in speech. It also demands far greater resources to collect and transmit a given amount of spoken language.

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 made by members of the public, and a context-governed part, containing transcriptions of recordings made at specific types of meetings and events (<http://info.ox.ac.uk/bnc/what/balance.html>; cf. the remarks by Berglund 1999: Section 2.1, p. 31–32; see further below, Section 5.1.2).

8. As will be explained below, this category combines more than plain ethnicity, and includes, in fact, reference to ethnicity or religion, to the place of birth and to the place of origin, as well as to the question of immigration.

9. This figure includes the sampled population, i.e. not including children below the age of fifteen (see below, Section 5.1.1).

10. For issues involved in sampling a population for linguistic analysis see Milroy 1987: Chapter 2. Any potential bias resulting from the replacing of randomly selected informants or any other issues which will emerge in the sampling procedures will, of course, be available. A second phase of the project will involve a trial run, or a pretest aiming to evaluate the statistical and related issues. For further details see the CoSIH web site at <http://spinoza.tau.ac.il/hci/dep/semitic/cosih.html>.

11. Different speech communities also differ in their inventory of speech situations (cf. Biber 1995). CoSIH is designed to dissect contextual features in conceptual rather than actual categories (see below, Section 5.1.2).

12. The variables used here account for statistical measures only, as explained in note 6 above. The first two EBO variables are reminiscent of the older distinction between Ashkenazi and non-Ashkenazi (or, rather, Sephardi) Jews. This latter distinction has been used time and again to distinguish between two major dialects in Israeli Hebrew (Ashkenazi or general Hebrew vs. Arabicized Hebrew; e.g. Blanc 1956a: 189, 1956b; Berman 1997: 312–313; Bolozky 1997: 287). We hypothesize that data extracted and analyzed from the corpus will show that this bipartite division is too general to account for actual linguistic variation in Israel.

13. The Jewish population of Israel in 1948 numbered only 650,000 people. The 684,000 immigrants who arrived in the newly established state between 1948 and 1951 more than doubled its population. In the early 1960s the Jewish population of Israel increased to 2,000,000.

14. A biased ratio to counteract the large-scale conversational text collection sampled in the demographic part of the corpus.

15. The design of CoSIH presented in this paper does not address questions raised by Atkins, Clear and Ostler. The aim of our paper is to present the initial design as a preliminary requisite for the compilation of the corpus.

16. For example, an elderly Jewish person, Israeli born, whose father is of Asian or African origin, with minimal education is unlikely to be recorded in contextual variety 1, since he or she is probably a member of a traditional family.

17. In some cases such under-representation may be rectified, at least to some extent, within the 50 cells of the demographically based supplementary subcorpus (Section 6.1).

18. The choice of a time segment rather than a number of words will enable us to compare language varieties on the basis of speed of speech and other features.

19. Some women also serve in the reserve armed forces, for a much more limited time than men.

20. This depends on the type of sampling taken. Sampling by residential areas would result in a significant gap in informants from the military, who spend their time mostly outside the home.

21. A scripted text is a text read aloud from a written version, usually prepared for oral presentation. Prepared speech is a text whose contents or form have been predesigned for oral presentation, yet still takes a free form when presented to the public.

References

Atkins, S., J. Clear and N. Ostler. 1992. "Corpus Design Criteria." *Literary and Linguistic Computing* 7: 1–16.

Bentolila, Y. 1989. "Methods of Speech Research: The Corpus of Montreal French." *Hebrew Linguistics* 27: 13–28. (in Hebrew)

Berglund, Y. 1999. "Exploiting a Large Spoken Corpus: An End-users's Way to the BNC." *International Journal of Corpus Linguistics* 4: 29–52.

Berman, R. A. 1997. "Modern Hebrew." *The Semitic Languages*, ed. by R. Hetzron, 312–333. London: Routledge.

Biber, D. 1990. "Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation." *Literary and Linguistic Computing* 5: 257–269.

Biber, D. 1995. *Dimensions of Linguistic Variation: A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

Blanc, H. 1956a. "Dialect research in Israel." *Orbis* 5: 185–190.

Blanc, H. 1956b. "A Note on Israeli Hebrew 'Psycho-Phonetics'." *Word* 12: 106–113.

Blau, J. 1991. "A Grammar of Modern Hebrew" (= Review of Glinert 1989). *Leshonénu* 55: 149–157. (in Hebrew)

Bolozky, Sh. 1997. "Israeli Hebrew Phonology." *Phonologies of Asia and Africa (Including the Caucasus)*, ed. by A. S. Kaye, Vol. 1, 287–311. Winona Lake, Indiana: Eisenbrauns.

Čermák, F. 1997a. "Czech National Corpus: A Case in Many Contexts." *International Journal of Corpus Linguistics* 2: 181–197.

Čermák, F and P. Sgall. 1997b. "Výzkum mluvené češtiny: jeho situace a potřeby" ("Research of Spoken Czech: Its Situation and Needs"). *Slovo a slovenost* 58: 15–25.

Choueka, Y. 2000. *Bar-Ilan Corpus of Modern Hebrew: Interim Report*. Institute for Information Retrieval and Computational Linguistics, Bar-Ilan University (Ramat Gan, Israel). (in Hebrew).

Crowdy, S. 1993. "Spoken Corpus Design." *Literary and Linguistic Computing* 8: 259–265.

Edwards, J. 1993. "Survey of Electronic Corpora and Related Resources for Language Researchers." *Talking Data: Transcription and Coding in Discourse Research*, ed. by J. A. Edwards, and M. D. Lampert, 263–306. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Encyclopaedia Judaica. 1972. Jerusalem: Keter.

Glinert, L. 1989. *The Grammar of Modern Hebrew*. Cambridge: Cambridge University Press.

Hudson, J. ms. "Categorizing Chaos: Text Types in a Speech Corpus".

Jastrow, O. 2002. "Neo-Aramaic Dialectology: The State of the Art." *Israel Oriental Studies 20: Semitic Linguistics: The State of the Art at the Turn of the Twenty-First Century*, ed. by Sh. Izre'el, 347–363. Winona Lake, Indiana: Eisenbrauns.

Kaddari, M. Z. 1984. "Introduction: The State of the Art of Israeli Hebrew." *Hebrew Books, Articles and Doctoral Theses on Contemporary Hebrew Published in Israel (1948–1980)*. (From the Workshop: Studies and Research for Teachers of Hebrew as a Second Language, 6), ed. by B.-Z. Fischler, 1–16. Jerusalem: Council on the Teaching of Hebrew. (in Hebrew)

Kaddari, M. Z. 1996. "The Pressing Need for a Survey of the Living Literary Hebrew Language." *Evolution and Renewal: Trends on the Development of the Hebrew Language: Lectures commemorating the 100th Anniversary of the Establishment of the Hebrew Language Council*, 127–147. Publications of the Israel Academy of Sciences and Humanities, Section of Humanities. Jerusalem: the Israel Academy of Sciences and Humanities. (in Hebrew)

Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. (Studies in Language and Linguistics.) London: Longman.

McCarthy, M. 1998. *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.

Milroy, L. 1987. *Observing and Analysing Natural Language: A Critical Account of Sociolinguistic Method*. (Language in Society, 12.) Oxford: Basil Blackwell.

Simeone-Senelle, M.-C. 2002. "Les langues sudarabiques modernes à l'aube de l'an 2000: evaluation des connaissances." *Israel Oriental Studies 20: Semitic Linguistics: The State of the Art at the Turn of the Twenty-First Century*, ed. by Sh. Izre'el, 379–400. Winona Lake, Indiana: Eisenbrauns.