Methodological Remarks on Creating *the Corpus of Spoken Israeli Hebrew (CoSIH)*

Regina E. Werum
Emory University, Atlanta, GA

Interdisciplinary research is easier said than done. The challenge in conducting such projects lies in drawing on a common core of methodological, theoretical, and substantive knowledge that makes our work accessible to others. In other words, interdisciplinary research requires intellectual breadth and extraordinary communication skills. The purpose of this chapter is to clarify some of the methodological considerations that have shaped the study design and data collection phases of Hary and Izre'el's Corpus of Spoken Israeli Hebrew (CoSIH). The methodological choices made in designing the CoSIH exemplify the potential strengths of interdisciplinary research projects.

When Benjamin Hary (Emory University) and Shlomo Izre'el (Tel Aviv University) first asked me to serve as methodological consultant to their project, I was flattered yet skeptical. After all, what do I know about sociolinguistics? Though my professional background is interdisciplinary in substance (I am a sociologist with a pronounced comparative-historical bent), my methodological training was strongly rooted in the social sciences, especially quantitative forms of data collection and analysis. But when Hary and Izre'el introduced me to their project and outlined the challenges they were facing, I realized that the potential pitfalls they needed to circumnavigate in designing the CoSIH were, in fact, typical of

interdisciplinary projects. These three pitfalls include nomenclature (disciplinary jargon), methodology (rules of evidence), and thematic focus.

The first pitfall proved to be a classic communication problem that required only a few meetings to overcome. Primarily, it involved defining concepts like "corpus," "dialect," "register," "sample," "unit of analysis," and "([in]dependent) variables." Once we had defined the scope of the project, we began to discuss sampling and analytical issues. Consequently, this chapter is organized as follows:

- Scope of the Project
- Research Design
- Conceptual and Analytical Focus

Though I will emphasize methodological issues, let me begin by providing the academic context for this project from a sociological point of view.

Scope of the Project

The Swiss-born linguist Ferdinand de Saussure (1857-1913) is commonly identified as the field's "founder." Sociologists continue to draw from his fundamental insights regarding the differences between *langue* and *parole*. We conceptualize the former in terms of language as a formal system, focusing on its structural aspects, while envisioning the latter as empirical examples of individual usage. Alternatively, the lay person's definition of both terms tends to revolve around the difference between written and spoken language.

A quarter century ago, Fishman (1974) noted two flaws in existing Hebrew corpus linguistics-based research: (1) it tended to focus on "model language" rather than documenting actual use of language; and (2) existing language studies of Hebrew continued to focus on its historical rather than modern usage. Although the present volume demonstrates that research on linguistic corpora has changed in some respects since Fishman wrote his article (also Berman 1997; Crowdy 1993; McCarthy 1999), de Beaugrande (1999) has recently reiterated that the tendency to favor "ideal language" over "real language" persists among researchers. Remarkably, Hary and Izre'el are the first to heed Fishman's call to create a corpus of spoken Hebrew. In Fishman's words, the main questions are the following:

> What models of "good Hebrew" does the public recognize in the press, in the radio, in literature? What is the popular "cognitive map" *vis-à-vis* the kinds of Hebrew that are spoken in Israel today (qualitative kinds? regional kinds? social class kinds?) [p. 11]

Ragin (1994) points out that most research projects are designed to fulfill one particular research goal. Part of CoSIH's uniqueness derives from the fact that it can, in fact, meet three research goals simultaneously:

- CoSIH will enable us to extract *descriptive* information, e.g., about the ethnic or regional differences in the use of spoken Hebrew, to which Fishman alludes in the above quote. But more importantly,
- CoSIH has the potential to answer *interpretive* questions about contextually significant phenomena, and
- CoSIH data will allow us to examine *causal* relationships regarding issues of social stratification.

These three goals roughly correspond to what Ragin calls "exploring diversity," "interpreting culturally significant phenomena," and "identifying general patterns and relationships" (1994: 32ff). Moreover, they illustrate how the CoSIH manages to bridge discipline-specific goals and methods that have traditionally—and unfortunately—pitted social sciences and humanities against each other (also see Griffin 1992, 1995). To summarize, the CoSIH presents a unique data collection effort that will provide research potential for specialists in various disciplines outside of linguistics, ranging from anthropology and sociology to education and journalism.

Research Design

Methodologically at the cutting edge, this project builds on recent developments (see Ragin 1994, 1998) by blending the logic underlying *qualitative and quantitative data collection methods*. Sections A and B below provide basic definitions and concrete examples related to the design of the CoSIH. Discussing research design leads to the principles of *sampling* and *measurement,* both of which logically precede the actual data collection phase. I elaborate on these issues in sections C and D below.[1]

[1] For more details on technical terms see Babbie (1995), Baker (1994), and Neuman (1997).

A. Methods as Tools

The ultimate goal of the CoSIH is to collect data on spoken Israeli Hebrew from every aspect of life, and to systematize the organization of the information obtained. Several aspects distinguish this corpus from other corpora: First, although linguistic corpora exist for many languages, no one has attempted to construct a corpus of spoken Hebrew. In fact, there is no corpus of a language *comparable* to Hebrew (Izre'el in this volume). Nonetheless, Hary and Izre'el were able to model aspects of this corpus following existing spoken language corpora (Crowdy 1993). Second, linguistic corpora are usually designed by/for grammarians and etymologists (Biber et al. 1994). Words or phrases have tended to be the main phenomena of interest around which other corpora have been designed. In contrast, the two linguists who have designed the CoSIH are primarily interested in the social context in which Hebrew is spoken. As a result, the CoSIH remains unusual regarding

- the size of the corpus
- the combination of sampling procedures used, i.e., the way in which cases or "cells" (segments of recorded speech) are selected for inclusion in the corpus
- the demographic and contextual variables employed to make the data accessible to social scientists.

Underlying theoretical and substantive concerns have shaped Hary, Izre'el and the other members of the team's[2] methodological choices in designing the CoSIH. Clearly, the substantive goal revolves around collecting a comprehensive representation of spoken Hebrew by recording speech used by different speakers in varying situations (monologues vs. dialogues, phone vs. face-to-face, TV, etc.) But a fundamentally sociological premise underlies these recordings: The authors theorize that situational context influences the use of speech. The three theoretically important axes—or "contextual variables"—along which the researchers have designed the corpus deal with the relationship between the speakers, the discourse structure, and discourse topic[3] (see Hary and Izre'el [this volume]; Crowdy 1993; O'Barr 1982). Section D below discusses these issues in more detail.

Eventually, the corpus team plan to make the CoSIH available to researchers from various disciplines, including social scientists interested in explaining causal relationships between, e.g., people's social background (SES, ethnicity, age, gender) and their use of specific idioms. Alternative investigations might deal with interaction effects between individual background, social context, and the use of language.

B. Researchers as Toolmakers

All methods of data collection, from non-participant observation to structured interviews, are subject to specific rules. Most of the time, we use well-established methods of data collection, "canned" survey questions and scales, mathematical formulas for testing linear or non-linear effects of one phenomenon on another. More importantly, few researchers design projects that bridge the classic gap between *deductive and inductive* research approaches. Nor do we typically employ *quantitative and qualitative* sampling procedures simultaneously. In this case, CoSIH team needed to invent some new tools to realize their vision of a corpus of spoken Hebrew.

The researchers plan to test specific hypotheses based on their premise that situational context impacts the use of spoken language. While we regard such theory-testing efforts as *deductive endeavors*, the data collection process itself involves recording qualitative data that will create a wealth of multi-level information about Hebrew language use in various settings. Coding those speech segments that are randomly selected and form the main corpus (n=950 cells) as *situations* rather than disembodied words or phrases will enable other researchers down the road to engage in theory building, e.g., based on conversational or linguistic patterns. Interestingly, we usually associate this type of research with *inductive reasoning*. But frankly, their decision to include additional speech segments selected via purposive sampling into a sub-corpus (n=50 cells) introduces an additional inductive element into their data set. Finally, their plan to code all "cells" or speech segments by situational rather than purely linguistic criteria partly arises out of necessity—as no comparable data set exists and categories stemming from other linguistic corpora can only be transposed to a limited degree.

[2] For a full list of the CoSIH team, see introduction in this volume.

[3] The authors define these dichotomous variables elsewhere in this volume. To be brief, the "interpersonal relationship" variable distinguishes conversations between relatives/friends from others; "discourse structure" involves the presence/absence of power differentials between speakers; "discourse topic" distinguishes conversations dealing with personal matters from conversations about other topics.

The innovative fashion in which their research agenda combines deductive and inductive reasoning parallels their choice to combine two *sampling procedures* some researchers may view as contradictory rather than complementary. Given the ambition of a project this size (1,000 cases of recorded speech segments at ca. 30 minutes each, for a total of five million words), the project's designers could have let methodological convenience drive the scope of this corpus. For instance, it would have been easier to code 1,000 segments of speech already recorded (e.g., broadcast news segments, sermons, lectures), or to have 1,000 people read the same text and code the data in terms of differences in dialect or intonation. Instead, the CoSIH team has turned to an unusual set of complementary "tools." They combine one sampling procedure typically associated with quantitative research (*random sampling*) with another sampling procedure employed in qualitative research (*purposive sampling*). The resulting data collection and coding phases are bound to prove extremely labor intensive, as they require a significant amount of fieldwork. Yet, once completed, the CoSIH promises to be invaluable to quantitative and qualitative researchers across the spectrum.

C. Sampling

Deciding on a sample required determining from whom, what, when, and where to collect data. This decision involved delineating alternative *types of sampling* (sampling procedures) and finding agreement on the "*unit of analysis*." While the former is important for the data collection phase, the latter informs both data collection and subsequent coding schemes.

1. Types of Sampling

When discussing the purpose and scale of this corpus, it became clear that Hary and Izre'el did not seek to provide a comprehensive "cognitive map" of the *universe* of spoken Hebrew. Instead, they grappled with the question of *sample* construction, i.e., which instances of spoken Hebrew to collect—and how many (Biber 1990).

Recent studies involving other linguistic corpora have been based on vastly different samples, collected in markedly different ways. For example, in an article examining the use of compliments in Egypt and the U.S., the authors designed questionnaires after completing a pilot study of 40 in-depth

interviews. By soliciting volunteer participants among their students, the researchers created what we call a "*convenience sample*" for both the qualitative (interview) and quantitative (survey) aspects of their project. Ultimately, the authors identified 60 different compliments on the basis of n=240 questionnaires each from Egypt and the U.S. (Nelson, El Bakary and Al Batal 1993). Another study has made innovative use of n=338 questionnaires and essays each. In this case, the author constructed a so-called "*random sample*" of teens in one particular town, intended to be representative of the population at large. He used the information they provided to conduct linear regression analyses of class-based linguistic patterns in Iceland (Thorlindsson 1987).

Sociologists define the universe as the "population" about which we wish to generalize. This issue is theoretical rather than empirical. For example: In the studies above, researchers might intend to generalize about all college students or all teenagers in said countries—or they might wish to generalize about the entire population of those countries. Moreover, the term "population" does not just refer to people, but to the totality of elements of interest to the researcher. In the cases cited above, information was collected from individuals. But the "population" may also consist of other units of analysis, such as words, sentences, conversations—or "cells" (segments of recorded speech) in Hary and Izre'el's terms. I will return to the issue of *units of analysis* in a moment.

Sometimes social scientists know their universe/population. In this case, the CoSIH team realizes that, in human terms, the population about whom they wish to generalize consists of the roughly 6 1/2 million inhabitants of Israel.[4] But in theoretical terms, their "population" of interest consists of spoken Hebrew as used in Israel in the early twenty-first century. Notice that the number of individuals practicing the language is irrelevant here; what matters instead are variations in the forms of speech practiced.

The challenge became to construct a sample large and diverse enough to capture accurately most (if not all) varieties of spoken Hebrew—while also keeping in mind that no one sample or corpus is perfect. That is, in the interest of keeping a thematic focus and ensuring that the data could yield

[4] Alternatively, they also could have designated only Israeli citizens (who can belong to various ethnic groups) or only Jews in Israel (who can hold various nationalities) as the human population at the basis of their analysis.

parsimonious coding schemes, the researchers had to decide which forms of speech (dialects, situations) to cut out. After debating which categories should form the core of the corpus, they decided that the sample should be divided into two sub-corpora or sub-samples. Each of them illustrates a different *type of sampling*.

Traditionally, researchers may have viewed each sampling type as representative of contradictory ways of conducting research. The sampling procedure underlying the main corpus seems to follow a deductive line of thinking, typical of quantitative research. The procedure that informs the supplemental corpus appears to follow an inductive logic, commonly associated with qualitative research projects. As the discussion below and their chapter in this volume demonstrate, the CoSIH team's careful design manages to weave both together. As a result, the supposedly quantitative part of the corpus is heavily influenced by qualitative concerns about the "thickness" of the data collected. And the qualitative sub-corpus is designed to meet criteria of interest to quantitative researchers: First, it is designed to minimize concerns about "small n" affecting generalizability and external validity of the corpus data. Second, it takes into account that not all speech takes place/has an impact in small group or interactional settings. Rather, the media and government organs play a significant and increasing role in impacting the use of language.

a. Main Corpus: Systematic Random Sampling

The main corpus of CoSIH comprises 95% of the entire corpus: 950 linguistic "cells," each containing approximately 5,000 words of text.[5] Social scientists distinguish between various forms of random sampling, though all of them seek to create a picture "representative" of the population/ universe.[6] *Systematic random sampling* relies on a pre-set procedure according to which every n^{th} person/case is selected for inclusion in a sample. In this case, the project team plans to randomly sample Hebrew

[5] Editor's note: Some changes have been made in the CoSIH design after Werum wrote her chapter. See Izre'el, Hary and Rahav 2001 and 2002.

[6] In all likelihood, their data collection will also include an element of cluster sampling. Rather than choosing cases for inclusion from as many locations as possible, fiscal and time constraints may cause data collection to take place in a few selected urban and rural areas. For further information about the benefits and disadvantages of cluster sampling, see Babbie (1995) and Kmenta (2000:157).

speakers based on their place of residence. Potential difficulties in obtaining a representative sample might arise during efforts to reach residents or obtain their consent and participation. Thus, the project team has thought about how to ensure the richness of dialects spoken. For that purpose, they intend to review the final number of linguistic cells recorded in light of specific demographic categories thought to influence speech (see Crowdy 1993), and to conduct a supplemental follow-up sample if necessary. The key demographic markers are the following:

- Ethnicity/place of Birth/place of Origin ("EBO")—subdivided into five mutually exclusive categories;
- Age—subdivided into three mutually exclusive categories; and
- Education—subdivided into three mutually exclusive categories.[7]

Altogether, the demographic matrix yields 45 different combinations (5x3x3). But it becomes far more complicated once we take into consideration that the authors strive for diversity in the content of speech recorded. In their project description (this volume) they use the term "contextual categories" to indicate analytically significant distinctions between, e.g., physical/spatial context (monologue vs. conversation, or phone vs. face-to-face interaction), substantive context (topic of conversation), and social context (relationship between speakers). Their term "contextual categories" is analogous to social scientists using the term "independent variables." From an empirical point of view, the main challenge will be to obtain multiple speech segments for as many categories in this matrix as possible (Biber 1990). If all else fails, the authors (and subsequent users) may need to collapse some of the sampling or contextual combinations.

[7] Realizing that the sample could be stratified in additional ways (e.g., by income/SES or sex), the authors decided that their substantive focus did not revolve around the gendered use of Hebrew. While the random nature of their sampling should ensure a roughly equal split between men and women, the project team does not intend to pursue this type of representation. By extension, acknowledging the connection between education and income, the authors decided that educational differences would be more closely related to differences in the use of language than income alone. This is a perfect example of the judgement calls researchers face when designing such studies. The smaller the study, the more agonizing such calls can be—and the more devastating the effects sampling bias can produce. Together with the overall size of the corpus, the team's carefully calibrated approach should reduce potential sampling bias.

The CoSIH team's decision to employ systematic random sampling procedures remains unrivaled when we compare CoSIH to the way other linguistic corpora have been collected. Remarkably, their decision to use random sampling will not necessarily benefit their own research agenda. This design further attests to their foresight, as it increases the utility of their corpus to social scientists interested in examining causal relationships and/or doing so in a quantitative manner. After all, systematic random sampling becomes most relevant when we expect the relationship between "dependent" and "independent" variables to differ among subgroups sharing ascriptive or demographic characteristics.

b. Supplemental Corpus: Purposive Sampling

The supplemental corpus will consist of 50 additional "cells" or instances of spoken Hebrew (also at about 5,000 words each). For this purpose, speech segments are not selected based on demographic criteria, with the goal to provide a "representative picture" of spoken Hebrew. Instead, cases are selected based on explicitly theoretical criteria (Denzin and Lincoln 1994). Knowing that some forms of speech have a tremendous influence on overall speech patterns despite the fact that relatively few people actively disseminate such speech, Hary and Izre'el will sample four broad categories of public speech: popular media (TV, radio), the Israeli parliament (Knesset), and the court system (see Crowdy 1993 and O'Barr 1983 for rationale).

2. Unit of Analysis

Understanding the significance of choosing the appropriate unit of analysis becomes key to any successful research project. In standard sociological terms, using the wrong unit of analysis can lead to so-called ecological fallacies. By this statement we mean that one must avoid making inferences from one unit of analysis to another. We commit an ecological fallacy when we use aggregate data to infer people's individual motivations. Let me employ a famous example based on Durkheim's classic study: If a researcher finds that suicide rates (aggregate level data) are higher in Protestant than in Catholic countries, and concludes that this fact is attributable to differences in attitudes (individual level data), (s)he commits an ecological fallacy.

Similarly, with regard to the CoSIH, given that speech segments form the main unit of analysis, researchers must take care not to make inferences about the use of specific words, sentences, or the individuals whose speech was taped. This caution may turn out to be one of the biggest challenges facing researchers who wish to use the CoSIH data. After all, when 5,000 words comprise a single segment/cell/case, it becomes difficult to analyze a large number of such cases side-by-side. To illustrate this point: The text of the essay you are currently reading contains roughly 5,000 words total. Consequently, I imagine that researchers interested in smaller units of analysis—say, the use of gendered and gender-neutral pronouns in parliament speeches, or TV shows, or phone conversations—might wish to recode parts of the original data. Nonetheless, Hary and Izre'el's initial coding scheme should facilitate subsequent selection of specific cells based, e.g., on demographic or contextual criteria.

D. Measurement

To resolve measurement questions we considered the content of the cases/situations to be sampled: To what degree can the use of spoken Hebrew vary within and across situations? This question required us to think in terms of *variables*—a.k.a. "phenomena of interest," "analytical categories," "indicators" or "explanans et explanandum." Once again, readers may prefer different terms depending on their own disciplinary backgrounds. Because Hary and Izre'el were immersed in the corpus linguistics literature and agreed on the analytically significant dimensions of their project, deciding how to measure variables proved to be one of the least difficult elements in designing the CoSIH.[8] Nonetheless, employing well-established indicators whose dimensions or categories have cross-disciplinary appeal surfaced as a core concern for the CoSIH team, precisely because they intend to make their data set widely available.

Variable construction actually consists of two stages, which we call *conceptualization* and *operationalization*. Both are intertwined (e.g., Babbie 1995; Neuman 1997). But the former poses a theoretical challenge (e.g., how do we define "ethnicity": In terms of religion? Language? Region of origin? Or a combination thereof?) In contrast, the latter poses an empirical

[8] For further discussion of how measurement issues can affect the reliability and validity of the data see social science methods texts such as Neuman (1997) and Babbie (1995).

challenge, one of exact measurement. In this case, the researchers relied on a definition of ethnicity typically used in the Israeli context, which is based on a combination of factors involving religion, place of birth, and family origin. The resulting "EBO" variable consists of five empirical, mutually exclusive categories:

- Israeli-born Jews with fathers from Asia/Africa
- Israeli-born Jews with fathers from elsewhere
- Foreign-born Jews who immigrated before or during1965
- Foreign-born Jews who immigrated after 1965
- Non-Jews (e.g., Muslims, Christians, Druze).

These categories make sense in the Israeli context: They distinguish Ashkenazim from Sephardim, differentiate immigration waves shaped by a crucial change in policy (1965), and identify ethnic minority speakers. Clearly, a corpus of another spoken language would use different ethnic markers. But it is also possible that a corpus of spoken Israeli Hebrew, if collected before the end of the Cold War, would have yielded strikingly different speech patterns to one collected in 2001. One potential weakness in the way in which the authors operationalize ethnicity lies in their failure to differentiate long-term residents from recent immigrants. In particular, the large influx of Jews from former Soviet countries may have had a significant impact on spoken Hebrew in the last ten years. If so, the CoSIH team's variable will fail to capture this ethnic distinction.

By extension, the contextual variables the team uses to sample speech segments also involve theoretical and empirical decisions. For instance, they define "discourse structure" in terms of the relationship between participants recorded in any particular speech segment. Every concept can have multiple "indicators" or variables. But of all the different ways to categorize interpersonal relationships, the team has chosen to stress one specific dimension: They dichotomize discourse structure regarding the presence or absence of a power differential between the speakers. The theoretical assumption underlying this operationalization is that spoken language differs in situations in which the speakers relate to each other in a hierarchical manner (e.g., they might use more formal language, fewer fillers, more gender-specific verbal hedges). Just like the demographic variables employed to stratify their sampling, the contextual variables the researchers have chosen illustrate their commitment to making the CoSIH a

data set of interest to social scientists from various disciplines: Both "ethnicity" and "power" form core concepts in Sociology.

Conceptual and Analytical Focus

At the beginning of this essay I mentioned one final potential pitfall. Familiar to all researchers, it tends to afflict interdisciplinary projects more than others. A lack of conceptual and analytical focus occurs when we strive to be everything to everyone. In the process, data collection efforts run at least two risks: They are never completed, or if so, the data turn out to cover a broad area of substantive issues, as intended. But in an effort to "conceptualize" and "measure" a myriad of phenomena, researchers have to step outside of their area of expertise. Unless they seek substantive help from other professionals, they may risk using indicators that are either empirically invalid, or otherwise flawed. Even if completed successfully, mammoth projects can also make the resulting data set cumbersome to use. Though we expect everyone to be technologically and methodologically sophisticated, user-friendliness—or lack thereof—continues to influence the extent to and way in which data are used. Thus, for maximum impact within and across disciplines, a data set that seeks to do less may actually end up accomplishing more (see Biber 1993).

The CoSIH project is ambitious: The researchers intend to code a total of five million words contained in 1,000 speech segments. This plan raises two questions. First, will they complete the CoSIH project? Their most immediate challenge lies in actually collecting the data. To date, ambiguity persists about the location and time frame for data collection. Research projects requiring this amount of fieldwork, followed by detailed linguistic coding, will require substantial and long-term financial support. If the authors succeed in obtaining the necessary grants, I am confident that they will complete the data collection phase, classify the individual segments according to demographic and contextual variables they have identified, and make the data set available to researchers worldwide.

The ultimate word-by-word coding of each segment, however, is unlikely to occur in the near future. In part, it is difficult to project a time frame for this corpus construction phase because, despite recent advances in software, the degree to which computer-based technology can be used in this effort remains unclear. If word-by-word coding had to be done manually, even if

just in part, it might require a long time to complete the CoSIH. More importantly, the authors have not yet specified the criteria, or even variables along which the content of each segment will be coded. Again, the CoSIH could be used to examine a host of purely linguistic issues related to grammar, syntax, pronunciation, or registers. In all likelihood, future researchers interested in such issues will draw sub-samples of the CoSIH and code those particular segments to meet their own research interests. While this customized coding may decrease the coherence of the data set at large, it may also expedite the use of CoSIH data, thus maximizing its impact across disciplines. In short, I expect the data collection effort itself to produce the desired result, i.e., to meet the empirical goal of providing a reasonably accurate description of variations in spoken Israeli Hebrew. It remains to be seen whether the resulting data coding and analytic efforts will meet the substantive goals set forth elsewhere in this volume.

The second question is whether the authors' own research agenda follows a conceptually and analytically clear line—what are their own theoretical and empirical goals? Remarkably, the authors have succeeded in separating their empirical goal regarding the construction of a corpus of spoken Israeli Hebrew from their own research agenda. As outlined above, the main goal behind the data collection effort remains descriptive: providing an overview of variations in spoken language. They have designed the CoSIH so that it permits them to investigate one particular slice of the data, but is perfectly suited for researchers whose interests differ markedly and who may analyze the data using a variety of methods or "tools" ranging from content analysis to multiple regression.

In contrast, their own research project for which they will use the CoSIH data proves to be far more theory-driven. By focusing on how the social context and different dimensions of "power" impact the use of language, Hary and Isre'el pursue a fundamentally sociological question. Their research logic is theory-centered and combines elements of both deductive and inductive reasoning. Figure 1 illustrates how sociological methods books would portray this type of logic: Representing the research process in the form of a wheel, rather than depicting it as a linear progression from theory/hypothesis over data collection to analysis (e.g., Neuman 1997).

Conclusions

While the phenomena of interest to sociologists tend to remain noticeably different from those of sociolinguists, we employ similar yet constantly evolving concepts and theories. For instance, several eminent sociologists have capitalized on de Saussure's early work, perhaps most notably Basil Bernstein (1971) and Pierre Bourdieu (1982/1991). Both researchers synthesized de Saussure's framework with more mainstream sociological theory by focusing explicitly on

- class differences in the actual use of language, and
- language as a form of "cultural capital," used as a status marker by members of the elite.

Given my own area of specialization,[9] exposure to research inspired by this tradition has been limited to social scientists examining how race, ethnicity, class, or gender have shaped the use and meaning of language (also see, e.g., Calhoun 1994; Lakoff 1990; Lareau 1989; Lareau and Horvat 1999; Thorlindsson 1987; Thorne et al. 1983; and Zaretsky 1994). To use research on the connection between gender and language as an example: Johnson (1994) demonstrates that situational context, especially difference in status between group members, has a stronger effect on conversation patterns than gender per se. In contrast, non-verbal behaviors are strongly influenced by the gender of conversation participants, regardless of their formal position in the group (also see Carli 1990; Kollock et al. 1985). Further research indicates that gender differences in the use of language are

[9] Due to blind spots in my own training, I could not possibly provide an accurate portrayal of de Saussure's influence in areas of literary or critical theory. But cultural sociologists and anthropologists have also drawn on de Saussure and his contemporaries, especially Emile Durkheim—one of Sociology's founding fathers. For example, Wendy Griswold (1992, 1993, 1994, 2000), who is best known for her work in the sociology of literature and "pop" culture, explicitly focuses on historical and cross-cultural differences in written language. Similarly, Mary Douglas (1984, 1994, 1996 and with Steven Tipton in 1983), who is perhaps most famous for her work on cultural practices and religion, has made significant contributions to the field by tackling the social construction of class and ethnicity. For further discussion of connections between Durkheim and de Saussure—including possible parallels concerning the Durkheimian dichotomy between "society" and the "individual", and de Saussure's distinction between the structural and idiosyncratic elements of language—see, e.g., Dinneen (1990), Gane (1983), Reckwitz (1997), Wagner (1990). For more general discussion in this area see, e.g., Thorne et al. (1983), Lakoff (1990); Heritage (1984), and Ochs (1992).

not limited to adults, but are also manifest among children (Maltz and Borker 1982).

To summarize, what can the CoSIH offer to Sociologists? The paragraph above implies that our discipline pursues a distinct line of inquiry. The "sociological imagination" (Mills 1959) pictures individuals both as products of their environment and as producers of social reality. By definition, our research involves a more or less explicit comparison between various population groups, such as segments differing by ethnic, class, or gender characteristics (Ragin 1994). Both in terms of size and complexity, the corpus data should be useful for sociologists interested in language as a form of cultural capital—a "currency" whose value depends on, e.g., the personal background of the speaker and the situation/context in which it is used.

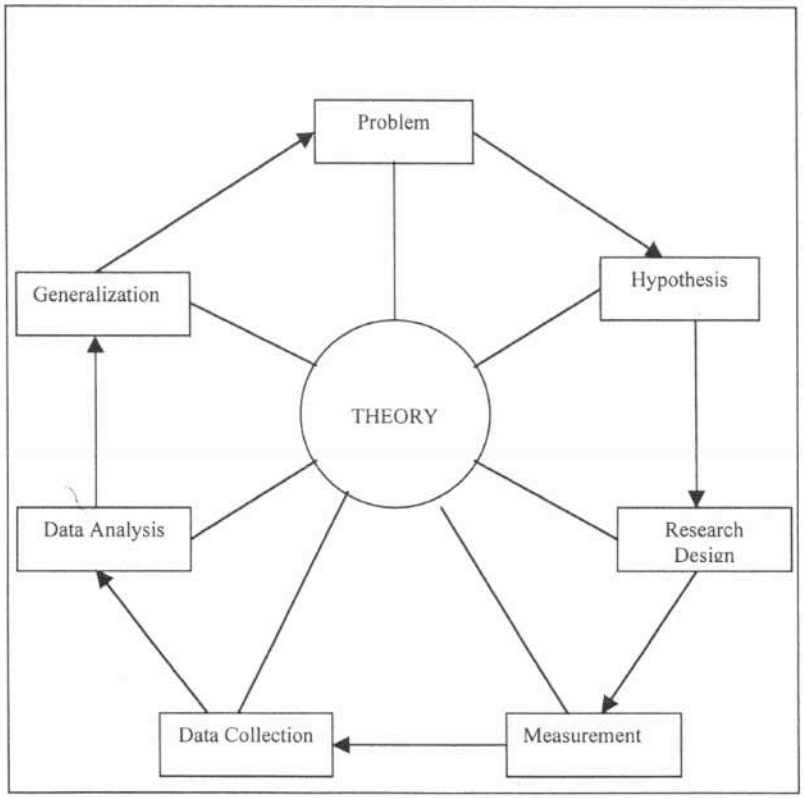

Figure 1

References

Atkins, Sue, Jeremy Clear, and Nicholas Ostler. 1992. Corpus Design Criteria. *Literary and Linguistic Computing* 7: 1-16.

Babbie, Earl. 1995. 7th ed. *The Practice of Social Research*. Belmont: Wadsworth.

Baker, Therese. 1994. 2nd ed. *Doing Social Research*. New York: McGraw-Hill, Inc.

Berman, Ruth. 1997. Modern Hebrew. In *The Semitic Languages*. R. Hetzron [ed.]. London: Routledge. 312-333.

Bernstein, Basil. 1971. *Class, Codes and Control*. London: Routledge and Kegan Paul.

Biber, Douglas. 1990. Methodological Issues Regarding Corpus-Based Analyses of Linguistic Variation. *Literary and Linguistic Computing* 5: 257-269.

_____. 1991. *Language and Symbolic Power*. Edited by J.B. Thompson. Cambridge, MA: Harvard University Press.

_____. 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing* 8/4: 243-257.

Biber, Douglas, Susan Conrad, and Randi Reppen. 1994. Corpus-Based Approaches to Issues in Applied Linguistics. *Applied Linguistics* 15/2: 169-189.

Bourdieu, Pierre. 1982. *Ce que Parler Veut Dire: L'Économie des Échanges Linguistiques*. Paris: Fayard.

_____. 1991. *Language and Symbolic Power*. J. B. Thompson [ed.]. Cambridge, MA: Harvard University Press.

Calhoun, Craig. 1994. *Social Theory and the Politics of Identity*. Oxford: Blackwell.

Carli, Linda. 1990. Gender, Language and Influence. *Journal of Personality and Social Psychology* 56: 565-576.

Crowdy, Steve. 1993. Spoken Corpus Design. *Literary and Linguistic Computing* 8: 259-265.

de Beaugrande, Robert. 1999. Linguistics, Sociolinguistics, and Corpus Linguistics: Ideal Language versus Real Language. *Journal of Sociolinguistics* 3/1: 128-139.

Denzin, Norman and Yvonna Lincoln [eds]. 1994. *Handbook of Qualitative Research*. Thousand Oaks: Sage Publications.

Douglas, Mary [ed.]. 1984. *Food in the Social Order: Studies of Food and Festivities in Three American Communities*. New York: Russell Sage.

_____. 1994. The Stranger in the Bible. *European Journal of Sociology/Archives Européennes de Sociologie* 35: 283-298.

_____. 1996. *Purity and Danger: An Analysis of Concepts of Pollution and Taboo*. New York: Praeger.

Douglas, Mary and Tipton, Steven [eds.]. 1983. *Religion and America: Spiritual Life in a Secular Age*. Boston: Beacon Press.

Dinneen, Francis. 1990. Ferdinand de Saussure. *Georgetown Journal of Languages and Linguistics* 1/1: 31-53.

Fishman, Joshua. 1974. Introduction: The Sociology of Language in Israel. *International Journal of the Sociology of Language* 120 (January): 9-13.

Gane, Mike. 1983. Durkheim: The Sacred Language. *Economy and Society* 12/1: 1-47.

Griffin, Larry. 1992. Temporality, Events, and Explanation in Historical Sociology: An Introduction. *Sociological Methods and Research* 20:4:403-427.

_____. 1995. How is Sociology Informed by History. *Social Forces* 73/4: 1245-1254.

Griswold, Wendy. 1992. The Sociology of Culture: Four Good Arguments (and One Bad One). *Acta Sociologica* 35/4: 323-328.

_____. 1993. Recent Moves in the Sociology of Literature. *Annual Review of Sociology* 19: 455-467.

_____. 1994. *Cultures and Societies in a Changing World*. Pine Forge Press.

_____ 2000. *Bearing Witness: Readers, Writers, and the Novel in Nigeria*. Princeton, NJ: Princeton University Press.

Izre'el, Shlomo, Benjamin Hary, and Giora Rahav. 2001. Designing CoSIH: The Corpus of Spoken Israeli Hebrew. *International Journal of Corpus Linguistics* 6/2: 171-197.

_____. 2002. Towards Establishing CoSIH. *Lešonenu* 54: 265-287. (Hebrew)

Johnson, Cathryn. 1994. Gender, Legitimate Authority, and Leader-Subordinate Conversations. *American Sociological Review* 59: 112-135.

Kmenta, Jan. 2000. 2nd ed. *Elements of Econometrics*. Ann Arbor: University of Michigan Press.

Kollock, Peter, Philip Blumstein, and Pepper Schwartz. 1985. Sex and Power in Interaction: Conversational Privileges and Duties. *American Sociological Review* 50: 34-46.

Lareau, Annnette. 1989. *Home Advantage*. London: Falmer Press.

Lareau, Annnette and Erin Horvat. 1999. Moments of Social Inclusion and Exclusion: Race, Class, and Cultural Capital in Family-School Relationships. *Sociology of Education* 72/1: 37-53.

Maltz, Daniel and Ruth Borker. 1982. A Cultural Approach to Male-Female Miscomminication. In *Language and Social Identity*. J. J. Gumperz [ed.]. Cambridge: Cambridge University Press. 196-216

McCarthy, Michael. 1999. *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.

Mills, C. Wright. 1959. *The Sociological Imagination*. New York: Oxford University Press.

Nelson, Gayle, Waguida El Bakary, and Mahmoud Al Batal. 1993. Egyptian and American Compliments: A Cross-Cultural Study. *International Journal of Intercultural Relations* 17/3: 293-313.

Neuman, W. Lawrence. 1997. 3rd ed. *Social Research Methods: Qualitative and Quantitative Approaches*. Boston: Allyn and Bacon.

O'Barr, William. 1982. *Linguistic Evidence: Language, Power, and Strategy in the Courtroom*. New York: Academic Press.

Ochs, Elinor. 1992. Indexing Gender. In *Rethinking Context: Language as an Interactive Phenomenon*. A. Duranti and C. Goodwin [eds.]. Cambridge: Cambridge University Press. 335-358

Ragin, Charles. 1994. *Constructing Social Research: The Unity and Diversity of Method*. Thousand Oaks: Pine Forge Press.

_____. 1998. The Logic of Qualitative Comparative Analysis. *International Review of Social History* 43: suppl. 6: 105-124.

Reckwitz, Andreas. 1997. Cultural Theory, Systems Theory and the Social-Theoretical Pattern of an Inside-Outside Distinction. *Zeitschrift für Soziologie* 26/5: 317-336.

Thorlindsson, Thorolfur. 1987. Bernstein's Sociolinguistics: An Empirical Test in Iceland. *Social Forces* 65/3: 695-718.

Thorne, Barrie, Cheris Kramarae, and Nancy Hendley [eds.]. 1983. *Language, Gender, and Society*. Rowley: Newbury House.

Wagner, Gerhard. 1990. Emile Durkheim and Ferdinand de Saussure. Some Remarks on the Problem of Social Order. *Zeitschrift für Soziologie* 19/1: 13-25.

Zaretsky, Eli. 1994. Identity Theory, Identity Politics: Psychoanalysis, Marxism, Post-Structuralism.) In *Social Theory and the Politics of Identity*. Craig Calhoun [ed.]. Oxford: Blackwell. Chapter 7, 198-215.