Population Sampling for the Establishment of a Representative Corpus[1]

Giora Rahav
Tel Aviv University, Israel

Introduction

The purpose of this article is not to describe what and how much has already been accomplished with regard to building a representative corpus of Israeli Hebrew. Instead, the purpose is to discuss plans and ideas for the future.

The problem is to find a way to represent adequately the spoken Israeli Hebrew as its natural speakers presently use it. The best way to prepare the corpus would be to record whatever anybody says and make it accessible for everybody next year. This method is obviously impossible. The only possible way to prepare the corpus is to take a subgroup of speakers and speeches, or a sample that represents the whole population and its speech patterns. But how can one be sure that the subgroup or sample is actually representative of the whole population?

Directed Sampling

One approach to the sampling problem would be to sample "typical speakers" who use "typical language." The problem with this solution is that "typical speakers" are highly atypical (or, at least, are rare). What is usually

---

[1] This article is based on a transcription of the lecture delivered at the CoSIH conference at Emory University, Atlanta.

considered "a typical speaker" or "a typical speech" may reflect stereotypes rather than the reality of the language and its speakers. It is, therefore, necessary to resort to a methodologically sound form of representing the population. Drawing a statistical sample of people and their speech should solve this methodological problem.

We must first specify what we want to represent: the spoken, natural language. "Performed language," or the language that one speaks to another for a third person (or audience) to hear must be excluded. We shall include only language as it is spoken in bona fide conversations, by participants who are actually engaged in unrehearsed speeches. The language of theater and the language of radio and television are excluded. Sampling language by sampling conversations means that the number of recorded speakers will be larger than the number of recording individuals.

The difficulty is to insure that the sampled speakers and speech adequately represent the whole community of Hebrew speakers and speech in Israel. For adequate representation, we must sample communities, sub-communities, and sub-cultures. The most significant subcultures must be reasonably well represented in the sample.

The sample must also include a variety of speech situations and contexts, since we do not want to limit the corpus to only one type of situation. We want to represent the whole range of situationally modified language, not just two people talking, one as interviewer and the other as respondent.

Finally, perhaps as a sub-category of sampling situations, we must consider sampling periods, or time sampling. Speech situations change with some regularity throughout the course of the day, the week, or the year. For example, speech situations during times of hard work, or of vacation, may be different from those situations in the middle of the day when everyone is engaged in regular work. Speech situations that change regularly would be different from situations that change irregularly, such as "fashions" or "fads" in speech. The issue is how to sample effectively all these speech variations.[2]

To perform effective sampling, we can follow one of two completely different approaches. We can follow a non-statistical approach, or a

[2] This discussion of sampling methodology simplifies the issues. In principle, sampling speech, sampling speech situations, and sampling speakers may be interrelated, and a sample of one may misrepresent the other two.

"directed sample." Those in charge of selecting the sample decide who will be included in the sample. For example, one could decide to select twenty speakers of Polish ancestry, ten of Arab ancestry, two Ethiopians, and so on.

Another way to produce a "selected," non-statistical sample is to let the sample develop itself. For example, if we know the number of people who are fluent in a particular language, or in some sub-population that we want to represent in the final sample, all we have to do is locate and identify some of them. Once we find one of these individuals and interview and record this person, we can ask him or her: "Could you direct me to a number of other people who are members of the same category, of the same sub-culture?" This approach is known as "Snowball Sampling." Similar to the former approach, it yields a directed, non-statistical sample, but it does have some advantages. We can control the sample relatively well, and we can find a large enough sample using limited resources. Much of the work on the corpus so far has followed these methods for the above stated reasons.

Statistical, Random Sampling

In contrast to the directed sampling methods, there is the statistical approach, which requires the introduction of random, chance processes into the sampling procedure. Random sampling does have several advantages. If done properly, it yields a fairly good representation of the whole population as we define it. This aspect is an advantage over the former methods, which at best represent what we believe to be the major trends, or characteristics, of the population. The results of a random sampling are not influenced by our beliefs about that population. A second advantage: though rarely accurate, random sampling does enable the calculation of the size of expected errors. This calculation is a major advantage, because it can tell us the level of confidence one may have in our statements about the data. Finally, a third advantage to random sampling is that this method is acceptable to scientific communities. People find it much more convincing when a sample is generated by statistical methods, and this is often the major reason for using random sampling.

Statistical (or random) sampling is based essentially on two principles. First, every person, every unit in the population must have the same probability of being included in the sample. This principle should be

deployed, at least initially, before we begin the sampling procedure. The sampling must be based on a random process completely non-determined by the investigator. A completely random process must not be affected (or biased) by our beliefs, our wishes, or any other external factor.

The second principle is that the sample must be relatively large. It is difficult to define how large. For typical public opinion surveys in Israel, the sample size is somewhere between 500 and 1,500 individuals. The sample size depends to a large extent on the number of natural or logical divisions to be considered in the sample. For example, if we know from the beginning that we want to compare male speech with female speech, we already know that we need to divide the sample by two. We may want to distinguish between two or three major age groups, which would require that we define additional subgroups, or subdivisions. If we know the sub-divisions in advance, we must include enough individual subjects in each sub-division to represent it well.

Bound by calculations of practicality, we figured that a sample of about 1,000 people would be adequate to represent the major aspects of Israeli speech for most purposes. How do we select the 1,000 individuals? And how do we sample the speech situations and the speech periods?

Sampling periods, sampling times, and sampling situations might be avoided if we sample the whole population, convince every participant in the sample to wear a tape recorder for a week or two, or for a whole month, and record everything in which he or she is engaged. This is clearly not practical. We need more reasonable guidelines to follow.

If we could convince a large enough sample of individuals to wear some sort of recording device and record themselves selectively over prolonged periods (nobody would agree to record everything), it would solve much of the problem of sampling situations. Every individual ordinarily moves through a large variety of situations and roles during the performance of daily activities. Of course, no individual moves through the whole variety of situations possible during a single day, or even a month. But when we consider the whole sample, in the aggregate, it covers a large, fairly representative variety of situations.[3]

[3] To be precise, the sample is designed to represent the population, not the situations. Therefore, the representation of situations may be somewhat biased.

Technical Sampling Approaches

To draw a representative sample of informants, or focal persons for the recording, we can follow one of three approaches. First, it may be possible to sample individuals from the official population registry. This approach is analogous to writing the name of each person in the population on a piece of paper, putting the papers in a hat, mixing them well, and then randomly drawing the number of names that are needed. In practice, the technical procedure of drawing a random sample can be computerized. If the population registry is sampled, we have a good foundation to claim that the sample represents the whole population (at least the initial sample, before attrition). To be sure, this representation is not complete, due to various factors. For example, some individuals may be out of the country, and some are not listed in the register because they are illegal immigrants. Nevertheless, such a sample has explicit validity.

If we can get the Israeli central Bureau of Statistics, with its access to census data, to participate, we can actually predetermine and dictate certain properties of the sample. For example, it is possible to request that the sample comprise a given number of individuals of certain ancestry, from a particular age group, who live in a specific district. Using such criteria, one can insure that the sample composition will be identical (based on these criteria) to that of the population. Alternatively, one may decide that a certain group should be over-represented in the sample. This procedure will guarantee adequate representation of a group, which may be important, yet randomly small.

Despite potential advantages, using an official population registry also has disadvantages. The major disadvantage is that it can be expensive. Once a name is selected, regardless of how random the selection has been, one still has to find that specific individual. If the person has moved, we have to find his or her new residence. If the individual is abroad, we have to wait until he or she returns home, and so on.

A second, much more common approach to sampling is based on area sampling of households, and then of individuals living in the households. Most surveys use this approach. The entire country is divided into localities, or communities. Then, each of the larger communities is divided into sectors, similar to census tracts in the United States. We can then sample, within each community, a number of census tracts. Within each

census tract, we sample several locations. We instruct the interviewers to select one person who lives in the first apartment building. Then move, say, ten apartment buildings down the road and select one person from that building. By repeating this procedure, one can sample five, or ten, or twenty apartments for each location. In each apartment sampled, the interviewer will pick one person, according to predetermined criteria.

This method is almost as effective as sampling from the population registry. To control the composition of the sample, we can precede the interview with a screening questionnaire (e.g., "How old are you?" "Where were you born?" "Where were your parents born?" and so on) and decide on the basis of the answers whether or not to include the individual in the sample.

Finally, the third method, which combines telephone interviews with the kind of interview we wish to conduct, would be randomly to select telephone numbers from all the private telephone numbers in the country, and dial them. Each time somebody responds, we ask several screening questions, and on the basis of the answers, we can decide whether or not to include the individual in the sample. If the individual is to be included, we ask for an address and directions to get there. We arrange an appointment, and we try to convince the individual to collaborate.

Practical Considerations and Ethics

We have to choose one of these three sampling approaches based on relevant technical considerations. This choice still leaves a number of open issues that cannot be answered within the framework of the textbook sampling methods. There are several cases in which technical failures may reduce and even bias the sample. Hopefully, technical problems will spread randomly over the population, in which case they will only reduce the sample without biasing it.

How to define the population is a critical decision: Who should and should not be included in the population? What are the age limits? If we include people age 18 and older, how do we interview those who serve in the military? If we include only those who are 20 or 21 years and older, we will miss much of the adolescent language. Are we going to include new immigrants who have just arrived in the country? They may know only a few words in Hebrew, or they may have had Hebrew classes for a dozen

years and speak Hebrew almost fluently. We must decide whether or not to include them, and we must also decide whether to include or exclude foreigners, foreign workers and visitors. Some visitors stay in the country for many years; some are foreign students, mostly in yeshivas (religious schools), who may stay there for several years. All these categories of people may know Hebrew well, or they may not know Hebrew at all, but they are part of the population. We must specify explicitly whether or not to include them, and what will be the selection criteria.

Among the technical problems, invasion of privacy is a serious issue. In the representative corpus study, privacy issues may pose major problems on several levels. Invasion of privacy may increase the refusal rate among the sampled population. How to overcome this problem should be explored in one or more pilot studies. Subjects who initially agree to participate, but later drop out of the sample may pose another problem. They may drop out explicitly, or by turning off the tape recorder for prolonged periods.

As we create the sample, we are going to encounter many refusals and rejections. Typically, in ordinary public opinion surveys, we have refusal rates of about 20% to 25%. In addition, between 5% and 20% do not explicitly refuse, but they either do not know the language or do not understand what they need to do to participate, and so on. As a result, the rate of dropout may reach as high as 30% to 50%. On top of that, in the present project we have to consider the possibility that some subjects will still drop out, or will turn off the tape recorder.

The informants, the recording individuals, will probably screen speech situations. For example, most individuals will turn off the tape recorder when they fight with their partner. Similarly, most people will not record conversations in the bedroom, at the bank, or with their lawyer. We are going to miss, or at least under-represent many speech situations, despite the fact that the individuals agreed, in principle, to record everything. Perhaps we should ask the police or the security forces to provide us with samples from their secret recordings.

Some rare words, situations, speech forms, communities or sub-communities, may not be included in the sample, regardless of which of these methods we choose. For instance, to be certain that the speech of Yemenites or the speech of criminals is adequately represented, we must over-sample these sub-populations. It is legitimate, and it can be done, but we have to

plan for it in advance.

Finally, we must be aware that any project that involves such broad recording requires significant preparation. This preparation may include administrative and logistical support, and solving ethical issues to the satisfaction of both the research team and the International Research Boards, e.g., Helsinki Committees and so on. Recording individuals inherently raises ethical and legal issues.