Comparison of Data-Sources in a Sociolinguistic Study
Muriel Hammer and Sylvia K. Polgar, Biometrics Research Unit, New York State Department of Mental Hygiene, and
Kurt Salzinger, Biometrics Research Unit and Polytechnic Institute of Brooklyn

This paper was presented, in part, at the American Anthropological Association, Denver, Colorado, 1965.

*This research was supported by NSF Grant #GS-611, and NIMH, USPHS MH 07477.

The authors would like to express their gratitude for the excellent assistance of Sid Hammer and Lois Lipper, who conducted the interviews; Mary Ford and Verna Schmauder, who worked on the data analyses; George Smith and Victor Bergenn, who administered the Cloze procedure. They wish also to thank Joseph Zubin, Chief of Biometrics Research, for his help and interest.
The general question investigated in the study discussed here is the relationship between actual social connections and speech forms. The research arises from an interest in studying the mutual relations of social and cultural forms, utilizing speech as an aspect of culture that lends itself more readily than most others to such investigation. "Actual social connections" refer to observed or reported social contacts, as distinguished from classification into social categories such as class or ethnic background. This is not to minimize social distinctions like class; however, the meaning of such distinctions and the processes by which they are created and maintained need further exploration. Our interest here is in techniques for studying social networks and the transmission and change of behavioral forms (in this case, speech) within them.

More specifically, this paper will consider the methodological questions raised by the fact that choice of data-sources on social connections has a profound effect on the conclusions one would draw about the characteristics of those social connections. In the work described below, results based on data from direct observation differ from those based on several kinds of informant-report. Since the primary concern here is with this methodological question, relevant theoretical issues will be dealt with only incidentally.

The basic plan of the study was to gather data on actual social connections and to relate these data to a measure of speech concordance. Since speech occurs primarily in direct interaction, the amount of speech concordance between pairs of individuals may be expected to vary with their amount of direct interaction or with the number of individuals linking them. That is, if A and G do not know or speak to each other (see Figure 1), but both of them speak with C, A with C or G with C is a one-step distance, whereas A with G is a two-step distance. A with O, who knows neither A nor C
but only G, is a three-step distance, and so on. Among one-step distances like A:C and G:C, magnitudes vary with amount of interaction. For example, if A speaks with C daily, and G with C only briefly once a month, A:C is closer than G:C, though both are one-step distances. Speech concordance is expected to vary as a function of social distance, whether measured by amount of direct interaction or by steps.

There are at least two logically independent components of speech concordance: predictability or comprehensibility of others' speech, and similarity of speech;\(^1\) but only the first, predictability, will be dealt with here. The technique used to measure predictability is the Cloze procedure (Taylor, 1953), which measures the ability of a subject to restore to its original form a transcription of speech which has had every fifth word deleted from it. This is an overall kind of measure, depending upon, and reflecting, the verbal redundancy of the language. An overall measure is desirable for our present purpose, and this one has been fruitfully used in relation to a number of other variables in the speech situation, such as class background (Lawton, 1964; Robinson, 1965), schizophrenia (Salzinger, Portnoy and Feldman, 1964; in press), drug intake (Salzinger, Pisoni, Feldman & Bacon, 1961), hesitations (Goldman-Eisler, 1958a, 1958b), statistical approximations to English (Salzinger, Portnoy and Feldman, 1962). The Cloze procedure is used here to investigate the degree to which speech predictability varies with social distance; it is also, in effect, one criterion of the utility of the social measures.

The major question of method to be discussed here concerns the social rather than the speech measures: what procedures must one use to get valid information on social distance—whether in terms of direct contact or number
of links -- which procedures should be applicable to large urban populations as well as rural or village groups.

The first research design was a simple symmetrical network, as diagrammed in Figure 1. An initial individual was interviewed. In response to a direct question about people close to him, he mentioned two others. These two were then interviewed, and were asked the same question about closeness. Their responses provided the basis for the choice of the next subjects, and so on. If the connections on the basis of which subjects were selected were the only ones existing among these individuals, one could simply count steps and assume that even though the relationship between A and B might be much closer -- because of amount of interaction, or by chance or distorted report -- than say that between D and H, still, both A:B and D:H would be much closer socially than A:H or O:H.

A small pilot group of 10 subjects was selected in this way. Their subsequent responses to a sociometric form asking informants to rate how well they knew each of the other subjects, showed considerable interconnectedness. In Figure 2, the uppermost dot indicates the initial individual. Subjects selected by the question on closeness are indicated by a solid line. Additional close connections between these individuals, as shown by the sociometric form, are indicated by broken lines. The accuracy and proper interpretation of our informants' statements about the closeness of their social connections therefore became crucial to the research.

To provide the basis for such interpretation, a set of people was studied about whom data could be obtained both from interviews and from observations. This was a group of people at a neighborhood coffee and doughnut shop on the West Side of Manhattan in New York City. The place has a large number of regular
customers who live and work nearby and who come in for breakfast, lunch, dinner, coffee, and just to spend time. Some of the regular customers knew each other before coming there, some have met there, and many of them see each other independently of the doughnut shop. Although there is no clear-cut division, the owners, employees, and customers tend to form two main groups, one centering around the day shift and one centering around the evening shift.

This group was chosen for study for several reasons. On the practical side, it was important that an active participant, already in the group, was available to conduct the interviews. He was able to elicit a high level of cooperation from subjects, and to conduct intensive interviews with them. Methodologically, the publicly accessible setting made it possible to collect information on people's social ties by systematic observation as well as interview. For theoretical reasons, the structure of the group, with its two sub-groups and its peripheral members, was of special interest. The composition of the group is varied in age, sex, and social background, both in the larger group and in the sample selected for intensive study, although they were not selected on that basis. The people who appeared to be most involved in the shop were selected first, and then several peripheral individuals were added. To summarize the characteristics of the sample of 20 selected for interviewing: they have an age range of 16 to 55; about half have been to college and half have not; about half were born in New York City and half were not.

The sources of data were as follows:

Two sets of interviews were conducted on each of the 20 main subjects by a participant in the group. The interview forms were standard, and all interviews were tape-recorded.

Two sociometric forms were used. One form was given at the end of
the second interview and asked for ratings based on amount of time the informant spent with each other subject. The same form was also used at that time to ask informants to fill in ratings they thought other subjects might give each other. Thus, a subject filled in a form for each of those he spent much time with, as he thought each of them should rate all other subjects (here called a "proxy" rating). The second sociometric form was given at the time the Gloze procedure was administered, and asked subjects to rate others at the doughnut shop in terms of how well they knew each of them. Ratings of 3, 2, 1, and 0, were used on both forms, with "3" as the closest rating.

There were 2 formal observation periods, each lasting about 2 weeks. The first simply noted presence and location of subjects (basically seating arrangement at the counter). The second noted interactions between subjects, particularly concentrating on speech interactions. 2

There have also been one to two years of informal observation and participation by the interviewer, the observer, and another member of the staff.

Finally, the speech samples were elicited during the first interview in standard fashion. The informant was asked to give his opinions about living in the city and he was allowed to speak without interruption for up to 10 minutes. Typescripts were prepared from these monologs, and the second hundred words of each monolog, 3 with every fifth word deleted, was later presented to every subject for him to restore the deleted words. Other passages, from newspapers and from an unrelated speaker, were presented first for practice. A number of interesting measures can be derived from this procedure, but so far only the simplest has been used.
This is the proportion of blanks filled in with exactly the same word as in the original passage. These same passages, and an additional set from the same subjects, were also presented to a group of 17 college students, unrelated to the doughnut shop, to obtain a measure of relative passage difficulty unrelated to degree of social contact. This measure was used as an adjustment in analysing the effects of social contact on the doughnut shop subjects' performance on each other's passages.  

The analyses were directed to two questions: first, are the subjects who are most central socially in the group those whose passages are most predictable to other members of the group; and second, do individuals predict better the speech of the individuals they know better?

Centrality in the group was defined by several different measures. The observed centrality measure was based on a rank ordering of the total number of times each subject spoke with another subject during the second observational period of two weeks. The rank order correlation between this measure of centrality and the predictability of the subjects' passages by the other subjects in the group was positive and high ( +.66, p < .01). That is, the more an individual was observed speaking with the group, the more predictable his speech was to the group (after adjustment for absolute level of predictability on the basis of the performance of the unrelated population).

The measure of predictability was then correlated with a number of centrality measures based on report rather than observation. There were several possible reported centrality measures. The observed centrality measure was by its nature quantitative and required very few arbitrary decisions for computation. The reported data used consisted of ratings of closeness; and there is no mathematical justification for deciding, for example, that a rating of "3" is 1 1/2 times the rating of "2" which is twice the size of "1".
Therefore, a number of reported measures were tested, deriving from three main sources: each subject's rating of how much time he spent with others was the first self-rating; each subject's rating of how well he knew every other subject was the second self-rating; the mode of the subjects' ratings of relationships between other pairs of subjects, provided a "proxy-rating". For each source, three bases were used to derive a rank ordering: 1) a weighted score of the sums of the "3", "2", "1", and "0" ratings; 2) a score based on the number of "3"'s and "2"'s -- that is, the two higher ratings -- assigned each subject; and 3) a score based on just the number of "3"'s -- that is, the highest rating -- assigned each subject. Self and proxy ratings were correlated; and each of these nine reported centrality measures was tested for correlation with both the observed centrality measure and the measure of predictability of speech.

As shown in Table 1, the correlations between self and proxy ratings are generally quite high, ranging from +.72 to +.92. Thus, these procedures do seem to be measuring approximately the same variable. Table 2 indicates that all reported measures also have moderately high significant correlations with the observed centrality measure. Nevertheless, despite these high correlations and the high correlation between observed centrality and speech predictability, the reported measures are generally not significantly related to speech predictability. The only significant correlations with speech predictability, among these reported measures, are obtained by scoring the first sociometric form, using only the highest ratings -- the "3's" --; and this is true of both self-rating and proxy rating, though slightly better for the proxy rating.

A possible rationale for this would be that one can more readily distinguish an extreme from all other points on a scale than to distinguish among them; that one can do this slightly more objectively for others than for oneself;
and that time spent together is more relevant for our research than the perception of knowing a person well. However, we obviously cannot yet say whether this is in fact the more appropriate type of measure to use. A number of reported measures were tested, and the fact that two of the nine correlations are statistically significant must be interpreted with caution.

In summary of the findings on centrality, the speech of the more central individuals in the group is more predictable to the other members of the group. This relationship is clear when centrality is measured by direct observation; it is inconsistent based on reported measures.

Another set of analyses dealt with individual-to-individual speech predictability -- as opposed to the previous group measures of centrality. In using an observational measure of connection between individuals, only those 16 individuals were included who were observed to have at least one person speak to them five or more times. Again the results are statistically significant: individuals performed better on the passages of the people observed to have spoken to them.

A number of reported measures of connection between individuals were also used. On one of them, the second self-rating, significantly more individuals did better on the passages of those they rated as well-known to them than on those of the individuals they gave more distant ratings. However, no other reported measure showed statistically significant relationships to speech predictability on an individual-to-individual basis. These other reported measures included: the individual's rating of those he spends most time with (that is, the first self-rating); his statements on that at the interview; his report of the people he is closest to in general; a social distance measure combining these sources of data; and proxy ratings.
Thus, as with the centrality measures, the expected relationship is found between individual social distances and speech predictability when observed measures of social distance are used, and inconsistent results when reported social measures are used.

In addition to centrality within the group, and connections between individuals, there are sub-groupings even within this small set of people. The observational record allows analysis of patterns of speech interaction -- unavailable by report -- which relate to speech predictability in several ways which appear to coincide and form subgroups; and these seem to be the same subgrouping suggested by informal report and observation. It would take us too far from the main point to spell this out here, but the main characteristics of these subgroups may be summarized as follows: The groups may be roughly classed as "Day" and "Evening" groups, each of them having one of the owners as a central figure. The Day group has more of the college graduates than the Evening group. The Evening group subjects tend to have balanced interaction patterns -- that is, each of these subjects tends to talk to others about as much as they talk to him; the Day group subjects tend to have unbalanced interaction patterns. In addition, the Evening group subjects tend to give speech passages which are initially more predictable (to an unrelated group of subjects) than the passages of the Day subjects, but which decrease in predictability as they continue talking without interruption; whereas predictability for the passages of the Day subjects is initially lower but tends to increase. These groups appear to have differing dominant modes of speech usage, possibly related to Bernstein's (1962a, 1962b) "restricted" and "elaborated" codes.
More directly relevant to the present problem however, is the fact that serious discrepancies between observed and reported social ties are apparent even without the speech predictability criterion. Analyses of the structure of the group based on observation, on two self-ratings, and on proxy-rating, lead to differing expectations. There are many criteria one could use to analyse the group's connections. In illustration, only mutual connections will be considered, based on highest sociometric ratings and on a minimum of 5 units of interaction for the observational material. In Figure 3, only the 3 individuals who appear most crucial in any of the diagrams are labeled. \( O_1 \) and \( O_2 \) are the two owners, and C is one particular customer. It should be noted that the number of connections is approximately the same in all 4 cases: 30, 23, 27, and 26 respectively. The largest difference in this is between the two self-reports, which give 30 and 23 connections. (The lengths of the lines in these diagrams do not represent differences in social distance. The lines signify presence of a connection, and do not vary with magnitude.) (See Harary, Norman & Cartwright, 1965.)

Based on the second self-rating, C is the most central individual, whose absence should split the group into two sub-groups. All three other sources, however -- the first self-rating, observation and the proxy-rating, -- indicate that C is actually a rather peripheral member. (In fact, his subsequent absence from the group seems to have had no notable effect upon it). Based on this second self-rating, also, \( O_1 \) has no effect on the group, and \( O_2 \) very little; deletion of either or both of them would affect no other connections. Based on the first self-rating, deletion of \( O_1 \) and \( O_2 \) would split the group neatly into two. Observation and proxy-rating indicate that deletion of \( O_1 \) and \( O_2 \) would tend to scatter the group. Comparing proxy-ratings and observation, \( O_1 \) seems somewhat more crucial in both cases, and the major differences are, first,
that two instead of five individuals are given no connection at all; and that
the inter-connectedness appears to be considerably higher by observation than
by proxy-rating.

We cannot yet say which of these characterizations is most valid, but they
are certainly different. The observational data does so far correspond better
to the independent speech measures, and we suspect might also turn out to be
a better indicator for social-structural and predictive analysis.

If we reject studies dependent on reports, we reject enormous amounts of
social information that we should have. If we accept them, we are thereby
accepting an unknown amount of built-in distortion. Two approaches to a
resolution of the problem seem possible. The first is the one discussed here,
involving the testing of the kind and amount of difference between data-sources
on a small group about whom different kinds of information can be obtained,
preparatory to working in situations where this is not possible. A major
problem in this approach is the determination of valid ways of extending the
results true of the group studied to a differently structured group. The
analyses so far suggest that reports about others may be a slightly better
source of information than self-report; and that judgments at an extreme
may be more useful data than finer discriminations made by subjects. The
second approach would involve the principle of tracers -- the model for this
is the injection into a population of some phenomenon whose path can be followed
by tapping indicators of its presence at various points. Rumor transmission
is one obvious social example of this. If one could establish with sufficient
precision the relationship under investigation here, between social connection
and aspects of speech concordance, it would constitute such a tool.
In conclusion, let us summarize by saying that the preliminary findings indicate a direct relationship between closeness of social contact and degree of predictability or comprehensibility of speech. To extend the investigation to larger groups, one must resolve the problems involved in having to rely on informants' reports about social contact, which are as yet less satisfactory than observations of social contact.
Footnotes

1. Some of the measures of speech similarity which will be used are described in Hammer and Salzinger, 1964.

2. The doughnut shop is open 6 days a week, Monday through Saturday, from 7 A.M. to midnight. Our observations were scheduled to sample each hour of the day and each day of the week. The first set of observations consisted of three half-hour periods daily (except Sunday) for two weeks in February, 1965. The second set of observations also consisted of three daily periods, but the length of these varied in accordance with the amount of activity in the shop at the time; at busy times, observation continued as long as 1 1/2 hours. These observations were done during March, 1965. Monday through Thursday were covered twice each, Friday and Saturday once each.

   The observational procedures are described in more detail in Salzinger, Hammer, Portnoy, and Polgar, 1965.

3. The first hundred words were not used, since past work indicates that they are more dependent on the interviewer's prior speech than are later passages from the monolog. To allow for exclusion of blanks which cannot be scored (because certain words are masked to protect the confidentiality of the subjects' interviews), 119 word passages are used. Thus the passages were typically from word 101 to word 219. For a few subjects, this was shifted by as many words as necessary to avoid special problems such as a long interruption in that passage.

Bernstein, Basil
1962a Social class, linguistic codes and grammatical elements. Language and Speech 5: 221-240.

Goldman-Eisler, Frieda
1958b The predictability of words in context and the length of pauses in speech. Language and Speech 1: 226-231.

Hammer, Muriel and Kurt Salzinger.

Harary, Frank, Robert Z. Norman & Dorwin Cartwright

Lawton, Denis
1964 Social class language differences in group discussions. Language and Speech 7 (3): 182-204.

Robinson, W. P.

Salzinger, Kurt, Muriel Hammer, Stephanie Portnoy and Sylvia Polgar
1965 Verbal behavior and social distance. Presented at the American Psychological Association, Chicago, III. Mimeograph.

Salzinger, Kurt, Stephanie Pisoni, Richard S. Feldman, and Pauline Bacon

Salzinger, Kurt, Stephanie Portnoy and Richard S. Feldman.

Taylor, Wilson L.
Fig. 1. Diagram of research plan: Lettered points indicate subjects and lines the connections by which they would be chosen.

Fig. 2. Diagram of trial network: Points indicate subjects; solid lines indicate the connections on the basis of which subjects were chosen; broken lines indicate additional close connections subsequently reported by subjects.
Table 1

Rank order correlations between "self-ratings" and proxy-ratings* (N = 20)

<table>
<thead>
<tr>
<th></th>
<th>1st Sociometric form</th>
<th>2nd Sociometric form</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( r_s )</td>
<td>( P )</td>
</tr>
<tr>
<td>Weighted score (sums of &quot;3&quot;, &quot;2&quot;, &quot;1&quot; and &quot;0&quot;):</td>
<td>+.92</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Number of &quot;3&quot; and &quot;2&quot; ratings:</td>
<td>+.75</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Number of &quot;3&quot; ratings:</td>
<td>+.88</td>
<td>&lt;.01</td>
</tr>
</tbody>
</table>

Table 2

Rank order correlations of "reported centrality measures" with (1) observed centrality measure and with (2) speech predictability measure (N = 20) (The rank order correlation between observed centrality and speech predictability is +.66, \( P < .01 \))

<table>
<thead>
<tr>
<th></th>
<th>Observed centrality measure</th>
<th>Speech predictability measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-ratings (1st Sociometric form)</td>
<td>( r_s )</td>
<td>( P )</td>
</tr>
<tr>
<td>Weighted</td>
<td>+.60</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>&quot;3&quot; + &quot;2&quot;</td>
<td>+.50</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>&quot;3&quot;</td>
<td>+.58</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Self-ratings (2nd Sociometric form):</td>
<td>( r_s )</td>
<td>( P )</td>
</tr>
<tr>
<td>Weighted</td>
<td>+.72</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>&quot;3&quot; + &quot;2&quot;</td>
<td>+.61</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>&quot;3&quot;</td>
<td>+.71</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Proxy-ratings*:</td>
<td>( r_s )</td>
<td>( P )</td>
</tr>
<tr>
<td>Weighted</td>
<td>+.58</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>&quot;3&quot; + &quot;2&quot;</td>
<td>+.55</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>&quot;3&quot;</td>
<td>+.66</td>
<td>&lt;.01</td>
</tr>
</tbody>
</table>

*Proxy ratings were given only on the first sociometric form.