INTELLIGIBILITY PATTERNS IN SABAH AND THE PROBLEM OF PREDICTION

Paul R. Kroeger

1. INTRODUCTION

The language or dialect boundaries which exist in a given geographical area are not necessarily a barrier to communication among the language groups of the area. What factors enable one speech group to understand the language of a neighbouring group? What factors determine the extent to which a group will be understood by its neighbours?

Intelligibility or comprehension across linguistic boundaries is a very complex phenomenon. Linguistic similarity, social contact, language attitudes, patterns of language use, educational policies and political pressures are some of the factors relevant to explaining intelligibility. Casad (1974), Collier (1977), Simons (1979) and others have been interested in developing models for predicting intelligibility from various linguistic and sociolinguistic measurements. Linguistic surveys, which may be conducted for many reasons, are the source both of the initial data for developing such models and of further data for testing them.

Statistical analysis of intelligibility testing data is a natural approach to the development of empirical models. However, the very nature of a dialect intelligibility survey places constraints on the data collected which call into question the applicability of some statistical procedures.

In this paper I will discuss the results of one particular linguistic survey, using various kinds of statistical measurements which help us interpret the data. But problems were encountered in applying some of the more sophisticated statistical procedures to intelligibility scores. The use of regression analysis with intelligibility data, particularly when the results are generalised beyond a particular sample, seems to be especially prone to error.

In most contexts, I am using the word *intelligibility* simply as a synonym for comprehension. At times it is necessary to distinguish *inherent intelligibility*, defined by Simons (1979) as "the theoretical degree of understanding between dialects whose speakers have had no contact", from learned comprehension due to language contact. At such times, the term *bilingualism* may be used to refer to any degree of learned ability to speak and understand a second language, without specifying any threshold level of competence above which people are said to be "bilingual".

Paul Geraghty, Lois Carrington and S.A. Wurm, eds *FOCAL I: papers from the Fourth International Conference on Austronesian Linguistics*, 309-339. *Pacific Linguistics*, C-93, 1986.
© Paul R. Kroeger

I am assuming the basic model of intelligibility presented in Simons 1979:

total intelligibility = similarity-based intelligibility + contact-based intelligibility

In other words, intelligibility can be broken down into two components: "inherent intelligibility", due to linguistic similarity, and learned intelligibility, due to sociological factors. For modelling purposes, these factors are assumed to be independent and additive.[1]

2. INTELLIGIBILITY IN SABAH

From 1978 to 1981, the Summer Institute of Linguistics carried out a language survey of the state of Sabah, East Malaysia (formerly British North Borneo). The survey is described in detail in King and King 1984.

In the introduction to Part 2 of that volume, Carolyn Miller writes:

> The purpose of the survey was 1) to determine dialect boundaries within defined geographical boundaries comprising the entire state of Sabah, 2) to determine more precisely via lexicostatistics and intelligibility testing the degrees of intelligibility across major and minor dialect boundaries, and 3) to attempt to determine the level of understanding and the extent of the use of the national language in villages across the state.
>
> (Miller 1984)

The first phase of the survey concentrated on collecting wordlists and recording texts; the second phase involved Casad-style intelligibility testing (Casad 1974). Statistical analysis of the results demonstrates a significant correlation between cognate percentages and intelligibility testing (IT) scores, and a smaller but still significant correlation between IT scores and geographical distance.

2.1 Survey design and the nature of the data

A *correlation coefficient* (r) is an indicator of "the degree of association of strength of relationship between two variables" (Kirk 1978). When the relationship between a pair of variables is perfect (i.e. one value is perfectly predictable in terms of the other), r = 1. When the two variables are totally unrelated, r = Ø.

The Pearson product-moment correlation coefficient is the most commonly used index of correlation. It measures the strength of the relationship between two quantitative variables, e.g. IT scores and cognate percentages.

For the full set of data from the Sabah survey (790 cases), IT scores and cognate percentages are related with a Pearson correlation coefficient r = 0.663. Since r is positive, we know that the higher the cognate percentage between two dialect groups, the greater their ability to understand each other is likely to be. The statistical measurement agrees with our intuitive expectation. But in order to interpret the statistics in any meaningful way, we must know quite a bit about the units of data, how they were collected, and what these measurements represent.

Cognate percentages were calculated by computer based on a 327-word subset of the S.I.L. Philippines wordlist. Wordlists were collected from some 325 villages across the state, plus a few more from Sarawak and the Philippines.

Intelligibility testing, using the technique described in Casad 1974, was carried out at 143 of these 325 villages, and at perhaps a dozen other villages where no wordlist had been collected. A group of subjects (ideally 10) was chosen in each village, and the individual scores for each test were averaged to determine the group score (expressed as a percentage).

Constraints of time, energy and attention span forced the survey team to limit the number of tapes tested at any one village to seven. One of these was the hometown tape, one was a national language tape, and five were from other vernacular dialects. The national language score is not relevant to this study, and is not included in the data base; this leaves a total of 790 cases, i.e. 790 pairs of corresponding IT scores and cognate percentages.

Intelligibility testing was used primarily to check the language boundaries presented in Smith 1984, which were based purely on lexicostatistic relationships. This goal determined how test points were chosen and which tapes were tested at each point.

Generally, no testing was done between villages whose wordlists were more than 90% cognate, unless sociological factors made reduced intelligibility plausible (e.g. the Muslim Ida'an with the non-Muslim Begahak). Such cases were relatively rare. Very little testing was done when cognate percentages were below 70%, unless geographic proximity indicated that language learning was a strong possibility; and almost no testing was done in cases where cognate percentages fell below 50%.

In most cases, the test tapes played in each village were recorded within 50km. of that village. Testing at greater distances was done only between related dialects or closely related languages, e.g. 70% cognate or closer. On the other hand, languages much more distantly related were tested if the geographical distance between them was small. These facts tend to weaken the expected negative correlation between distance and intelligibility.

2.2 Intelligibility and lexical similarity

As mentioned above, the correlation coefficient between intelligibility (INT) and lexical similarity (LEX) over the full data set of 790 cases is r = 0.663. The square of this figure, $r^2 = 0.4398$, has a more intuitive interpretation. From the formula for r, we can show that r^2 is equivalent to the percentage of variation in one variable that is explainable by the variation in the related variable. In other words, 44% of the variation in IT scores can be explained by the corresponding variation in cognate percentages.

It is helpful to compare the results for the Sabah data with those from other similar studies. Simons (1979) analysed the correlation between INT and LEX for 10 different surveys in various parts of the world. These 10 represented all the studies Simons could find, prior to 1977, where both INT and LEX had been measured. The results of his analysis are shown in Figure 1-a. The corresponding values for the full data set from Sabah are shown in Figure 1-b. The meaning of the regression equations will be discussed below.

Study	N	Corr	%EV	Regression Equation
Biliau	9	.425	18.1	INT = 0.28 LEX + 66.3
Buang	21	.702	49.3	INT = 0.81 LEX - 12.4
Ethiopia	30	.846	71.6	INT = 1.22 LEX - 30.5
Iroquois	14	.813	66.0	INT = 1.52 LEX - 76.9
Mazatec	19	.807	65.1	INT = 1.77 LEX - 81.5
Polynesia	77	.864	74.6	INT = 1.59 LEX - 67.2
Siouan	25	.805	64.9	INT = 4.39 LEX - 336.0
Trique	15	.765	58.5	INT = 1.41 LEX - 41.3
Uganda	10	.905	81.8	INT = 1.33 LEX - 52.2
Yuman	25	.983	96.6	INT = 2.04 LEX - 106.2
Average	24	.791	64.6	INT = 1.05 LEX - 15.4

Figure 1-a: Ten studies from Simons 1979; INT vs. LEX

Study	N	Corr	%EV	Regression Equation
Sabah	790	.663	43.98	INT = 1.03 LEX - 3.28

Figure 1-b: Full raw data from Sabah survey, INT vs. LEX

Key:

N = number of cases
Corr = correlation coefficient, r
%EV = percentage of explained variation, r^2
Regression Equation = formula for predicting INT from LEX; defines regression line

The Sabah data set is huge in relation to any other published study of this type: 790 cases, as compared with 245 *total* cases for the 10 studies to which Simons had access. In terms of the strength of correlation (shown by r and r^2), the Sabah data is somewhat below the average of the 10 studies.

For the purposes of this study, it was decided to eliminate the hometown IT scores (i.e. subjects' scores on the test tape from their own village) from the data set. Hometown scores were included in Simons' calculations, but they are not really the same kind of measurement as other IT scores. Our model assumes that everyone understands his own dialect perfectly (i.e. 100%). Hometown tests are not tests of intelligibility but of the test itself and the subject's ability to take it.

The hometown scores in the Sabah survey were generally quite high, ranging from 80 to 100 with a mean value of 97.1. This reassures us that, on the average, the technical quality of the tests (e.g. tape quality, stories used, question construction) and the abilities of the subjects were not a major source of testing error.

The 133 hometown scores comprise 16.8%, just over one-sixth, of our data set. The LEX value for a hometown test is always 100%, and the INT values are generally very close to 100%. Therefore, when the data is displayed as a scattergram (as in Figure 5 below), the hometown scores form a large cluster of cases around the point (100,100). The effect of removing this cluster naturally reduces the calculated strength of correlation between LEX and INT, as shown in Figure 2.

Study	N	Corr	%EV	Regression Equation
Full data	790	.663	43.98	INT = 1.03 LEX - 3.28
Exclude HT scores	657	.568	32.28	INT = 1.17 LEX - 13.12

Figure 2: INT vs. LEX, Sabah data

How can we evaluate the strength of the relationship indicated by r = 0.568? By way of analogy, we could view the measurement of cognate percentages as a kind of aptitude test. The degree of linguistic similarity between two dialects represents the innate ability of members of one dialect group to understand speakers of the other. Lexical similarity is an imperfect but useful, and easily measured, index of linguistic similarity. Taking an intelligibility test represents a complex task to which linguistic similarity is obviously relevant. Aptitude (LEX) is one of a number of variables which determine the level of actual performance of that task (INT).

Kirk (1978:108) states that the best scholastic aptitude tests rarely achieve a correlation coefficient higher than r = 0.60 between aptitude test scores and actual academic performance. If our analogy could be extended in detail, the correlation coefficient (excluding hometown scores) r = 0.568 is very respectable for an aptitude test.

However, the actual strength of the relationship between LEX and INT in Sabah is almost certainly higher than the value of r would indicate. The correlation coefficient has been reduced by the nature of the data sample, specifically its range and distribution. These problems relate to the basic design of the survey, and may be inherent in any dialect intelligibility survey situation (see section 4 below).

Kirk (1978) states that "the restriction or *truncation* of the range of [either] variable results in a misleadingly low correlation coefficient". He points out that college aptitude scores do not correlate very highly with grade point averages in college, because the college admissions process truncates the range of data. People whose aptitude scores are low do not get in.

The design of the Sabah survey had a similar effect on the range of LEX values. Because the primary aim of the survey was to establish or verify language and dialect boundaries, very little testing was done between groups that were clearly distinct linguistically. Smith (1984) used 80% cognate as an approximate threshold value below which two speech varieties could be considered distinct languages. Thus the intelligibility testing focused on the cognate range of 60-90%; only three cases below 50% cognate were tested (see Figure 3). The data set was effectively truncated at LEX = 50%.

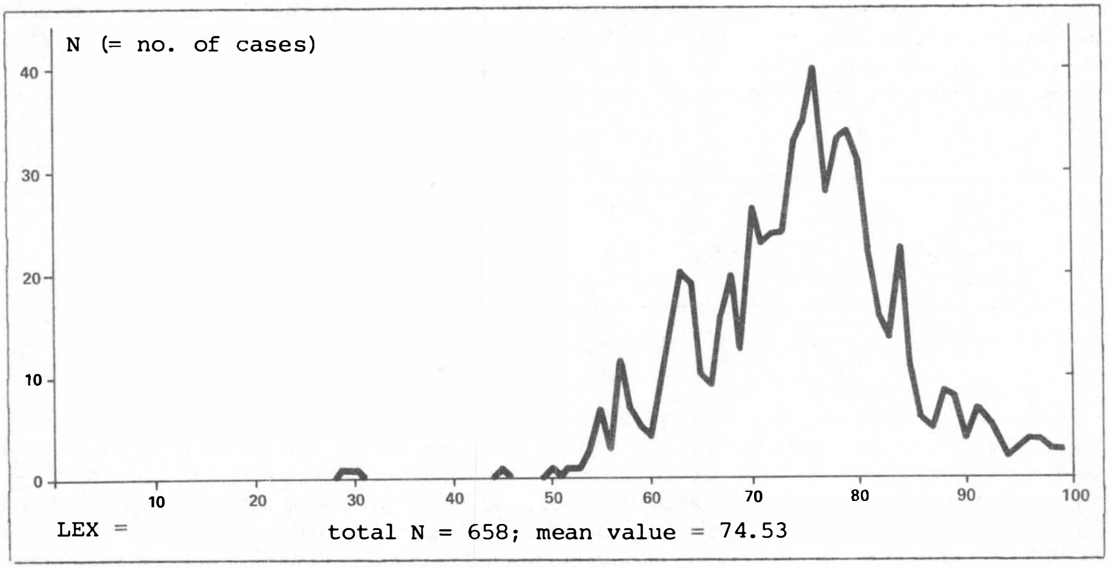


Figure 3: Distribution of LEX values for Sabah intelligibility survey

The second factor contributing to lower values of r was the *skewing* of the data, or uneven distribution of cases, particularly of the INT values. Figure 4 shows the distribution of INT. The average value of INT excluding hometown scores was 73.6%. There are slightly more cases above the mean value than below it, and there are far more occurrences of each value above 73% than of the values below that figure. The distribution of LEX is also skewed somewhat to the right (higher values), but far less so than INT.[2]

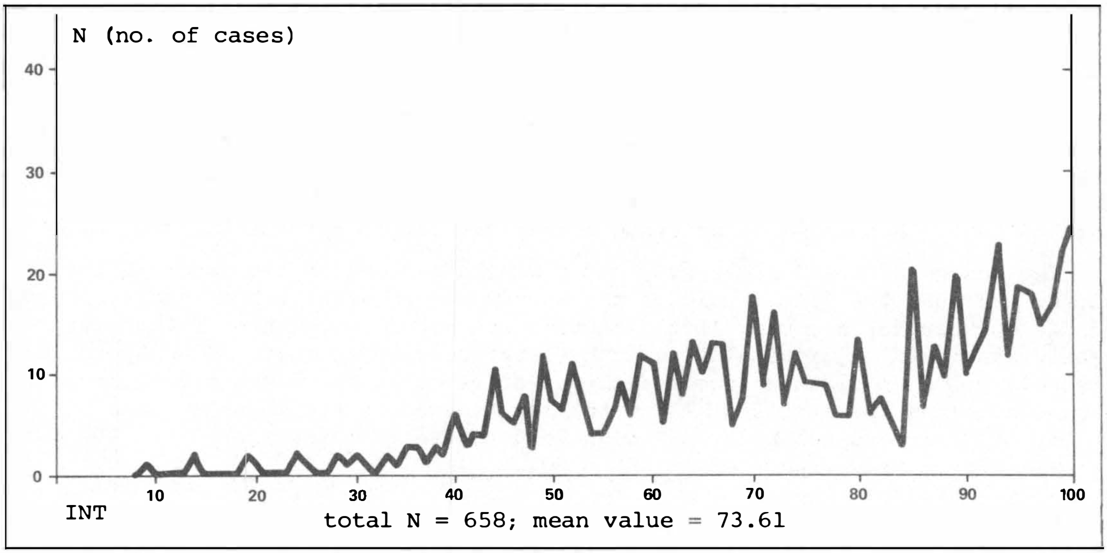


Figure 4: Distribution of INT values

The distribution of IT scores is related to the testing methodology. Following Casad 1974, simple personal experience stories were used, so average IT scores between dialects of the same language were rarely below 80%. On simple first-person narratives of this type, the ability to answer 10 questions out of 10 correctly does not necessarily indicate comprehension equaling that of a native speaker. But since no score higher than 100% is possible, the mean IT scores between related dialects appear as a dense cluster between 80% and 100%.

Another factor at work is the effect of language learning, which tends to inflate IT scores. There is no effect of anything like the same magnitude working to lower scores, so the net result is a higher frequency of high scores.

"If the distributions of [the variables] are markedly skewed, the value of r will be less than if the variables are approximately normally distributed" (Kirk 1978:113). Once again, it seems safe to predict that the actual correlation between LEX and INT in Sabah is greater than the measured value, r = 0.568.

The relationship of LEX and INT can be approximated by a linear equation of the form: INT = aLEX + b. Linear regression analysis is a technique for calculating the parameters (a and b) of this equation. The *regression line* defined by the linear equation is the line of best prediction for INT in terms of LEX. For the data set as a whole, the total difference between actual measured values in INT and the predicted values based on corresponding values of LEX (that is, the total prediction error) is minimised. Figure 5 shows a scattergram of the Sabah data (excluding hometown scores) with the associated regression line (line a), defined by the equation INT = 1.17 LEX - 13.12.

For normally distributed data, the regression line is the line of best fit (the line which passes closest to all the points in the scattergram), and r is a measure of how closely the points cluster around the line. However, we have already noted that the Sabah data is not normally distributed, and visual inspection of the scattergram shows that the regression line does not fit the points very well. This is confirmed by residual analysis, i.e. plotting prediction error against observed values of INT (see Figure 6). The rising trend in Figure 6-a indicates that the prediction error is roughly linearly dependent on the observed value of INT, and thus that the formula for the regression line does not adequately describe the data.

In our regression analysis, we have assumed that LEX was the independent variable and INT the dependent. In other words, we choose (for theoretical reasons) to predict INT from LEX rather than vice versa.

The shape of the regression line is dependent on the choice of dependent variable, and, as Figure 5 shows, the regression line for predicting LEX from INT (line b) is quite different from the first line. By visual inspection, it appears to fit the data much better than the first regression line. Figure 6-b shows the residual analysis for line b, Figure 5. The even, horizontal pattern in Figure 6-b indicates that the prediction error is random with respect to the observed value of INT, and so line b does in fact fit the data better than line a.

Line b is also a more plausible model of the actual relationship between LEX and INT than line a. Line b (INT = 3.57 LEX - 193.2) predicts zero inherent intelligibility between any two languages less than 54% cognate with each other, and full intelligibility between pairs of languages above 82% cognate. Intuitively, we would expect a higher threshold for full intelligibility, e.g. 90-92% cognate,[3] but the basic shape of line b is at least suggestive of the type of model we expect.

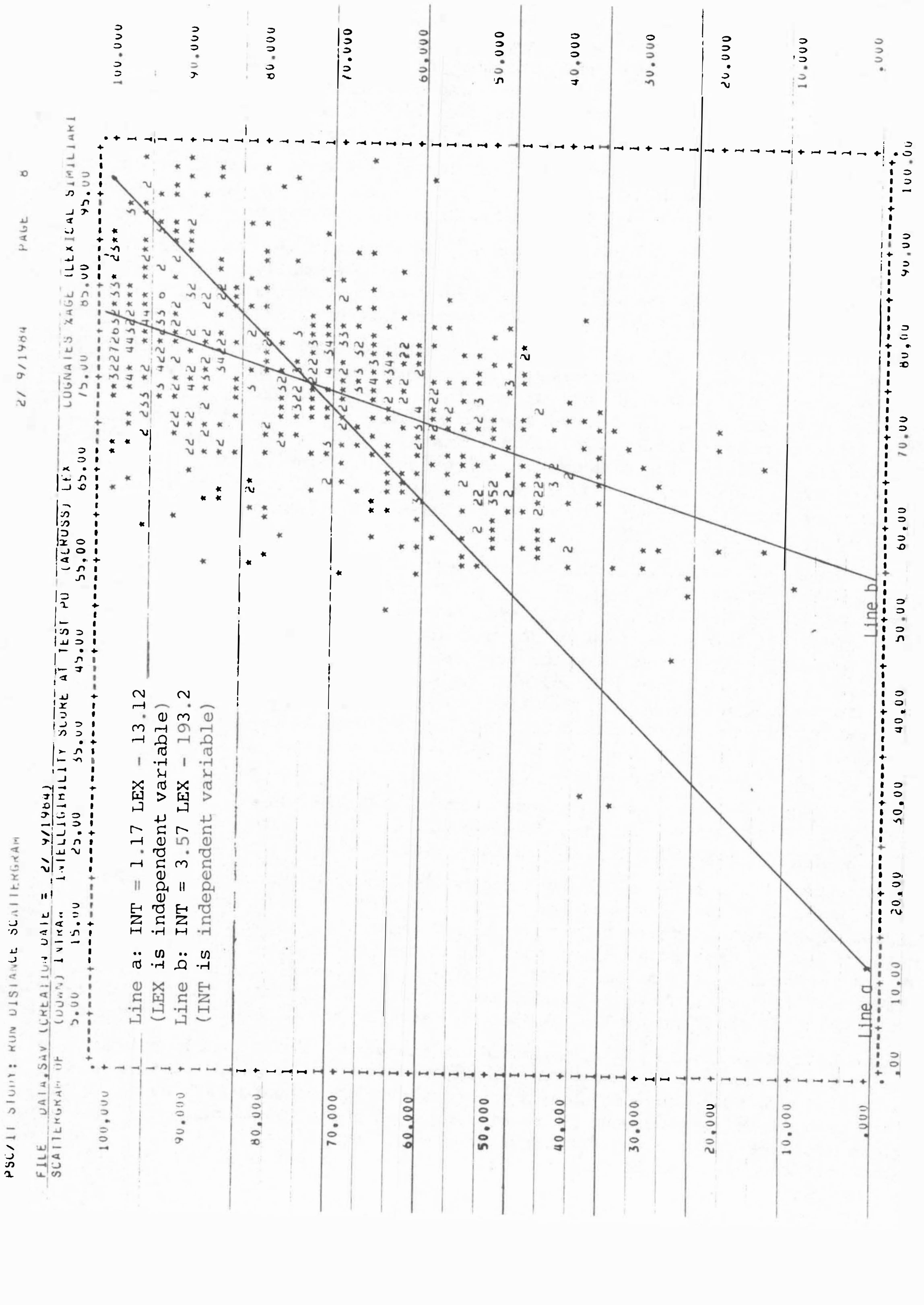


Figure 5: Scattergram of INT vs. LEX with regression lines

```
PSC/II STUDY: RUN SCATTERGRAM                                       6/11/1984        PAGE    2

FILE    DATA.SAV (CREATION DATE =  4/25/1984)
SCATTERGRAM OF    (DOWN) Z1                                        (ACROSS) INTRAW    INTELLIGIBILITY SCORE AT TEST PO
```

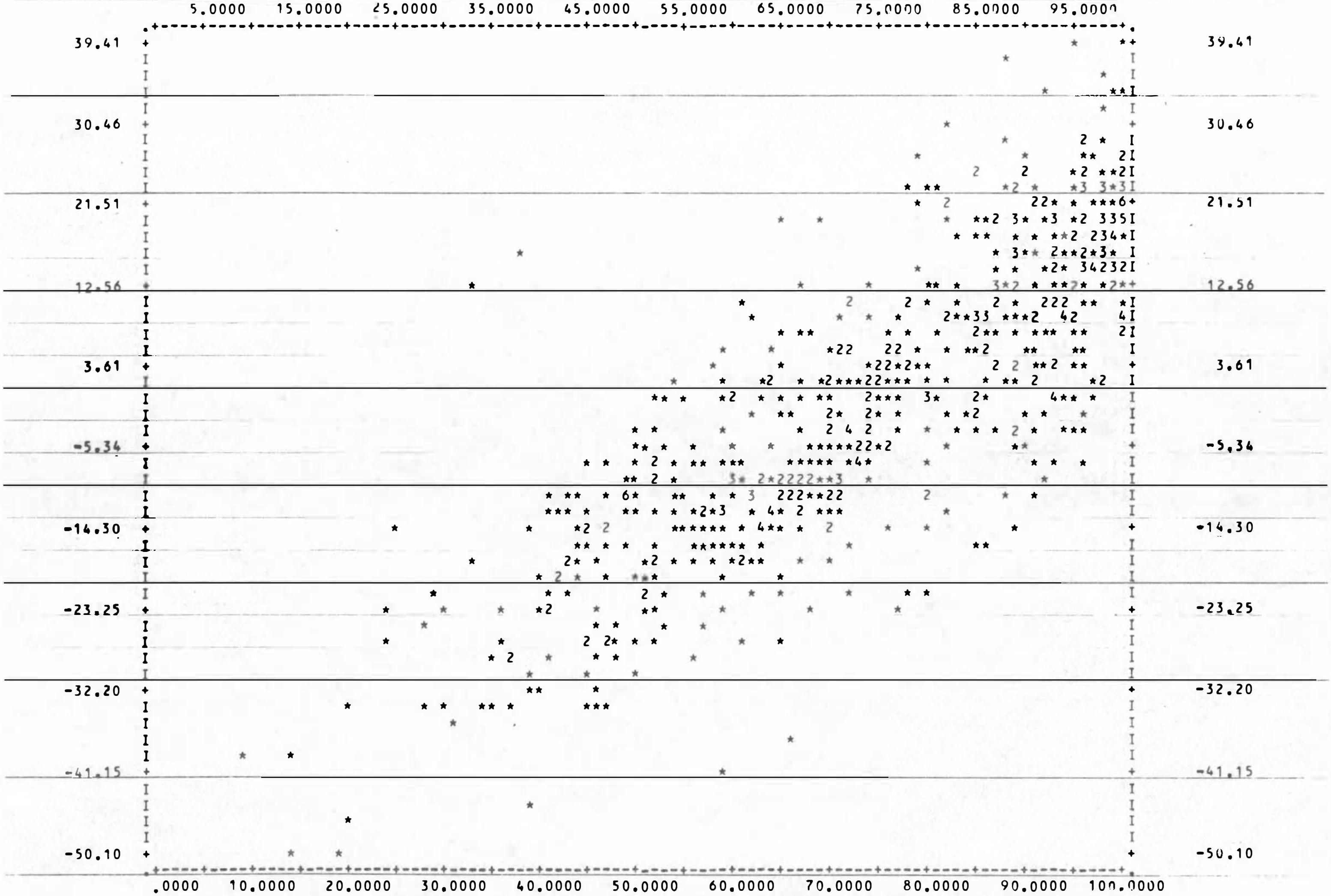


Figure 6-a: Residual analysis with LEX as independent variable. Prediction error vs. measured INT

```
PSC/IT STUDY: RUN SCATTERGRAM                                   6/11/1984        PAGE    5

FILE    DATA.SAV (CREATION DATE =   4/25/1984)
SCATTERGRAM OF    (DOWN) Z2                       (ACROSS) INTRAW   INTELLIGIBILITY SCORE AT TEST PO
```

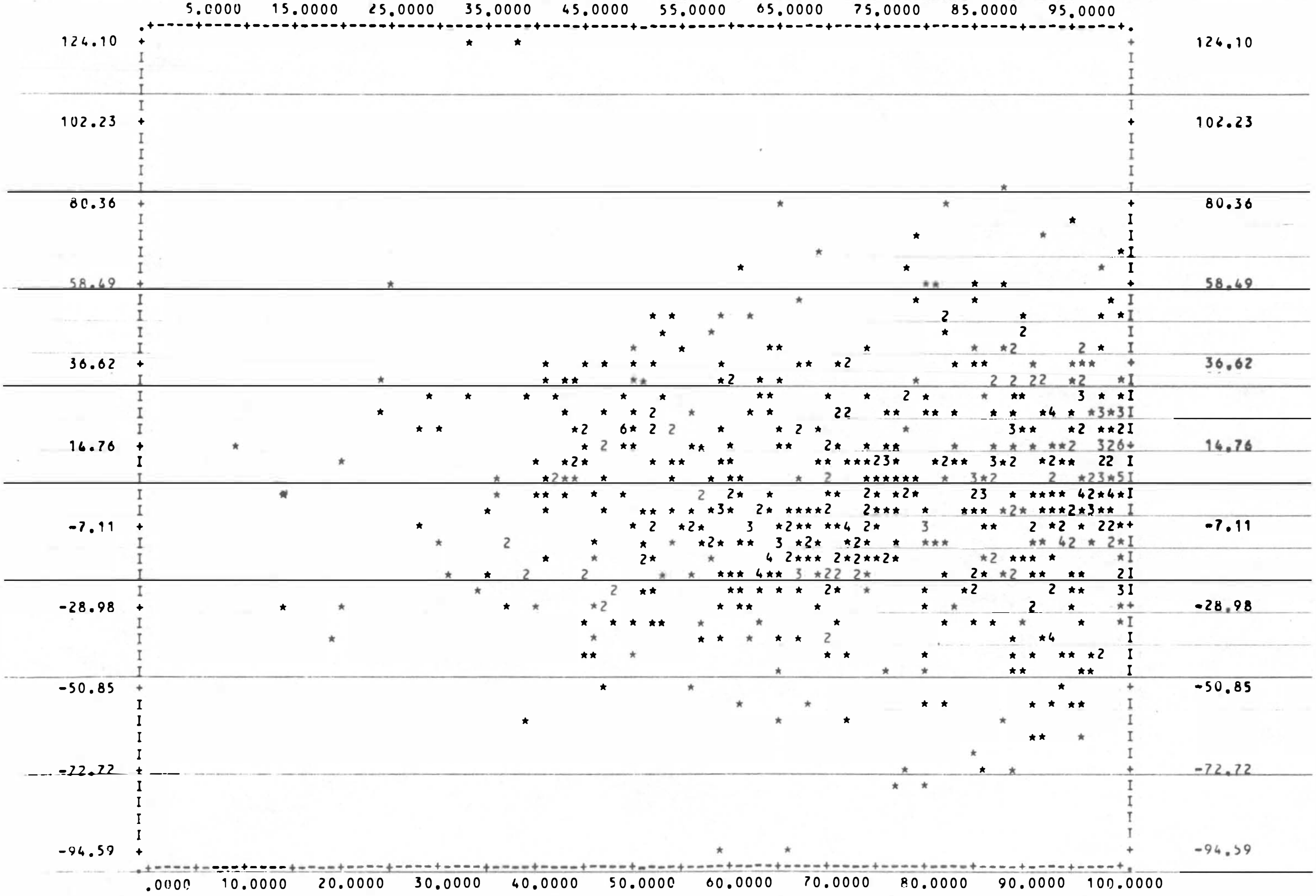


Figure 6-b: Residual analysis with INT as independent variable. Prediction error vs. measured INT

Line a, on the other hand, is much too "flat". It predicts zero intelligibility only below 11% cognate, when experience tells us that this threshold must be closer to 50%.[4]

But no theory of language would treat lexical similarity as dependent on intelligibility, so it is nonsense to say that the second line is the "correct" one. The difference in the two lines may be related to the fact that LEX is more nearly normal in distribution than INT. Both lines tell us something about our particular data set, but probably neither line tells us much about the actual relationship between LEX and INT in Sabah.

2.3 Intelligibility and distance

People who live near each other are more likely to interact than people who live far apart. We expect to find a negative correlation between distance and language contact, and therefore a negative correlation between distance and intelligibility.

2.3.1 Previous studies

Simons (1979:ch.6) shows that relative distance (relative to position in the dialect area) is a better predictor of intelligibility than absolute distance, for the dialects on Santa Cruz Island. Unfortunately, the results of that study are not comparable to ours, because he measured intelligibility on a discrete point scale (3 = full, 2 = partial, 1 = sporadic, Ø = none) rather than as a percentage.

Simons measured distance in travel time, which is clearly more relevant than raw physical distance. However, Walter and Echerd (n.d.) present a very interesting study using raw, straight-line distances measured on a map. For the Cakchiquel dialect system, intelligibility correlates very strongly with the natural logarithm of distance (see Figure 7).

Study	N	Corr	%EV	Equation
Cakchiquel	19	.96	92.16	INT = 332.77 - 93.17 ln(DIST)

Figure 7: Cakchiquel data from Walter and Echerd (n.d.) LEX vs. INT

The correlation coefficient r is an astonishing 0.96, equivalent to 92% explained variation. The authors apologise for the crudeness of the measurement (straight-line measurements on a map), but the results for the Cakchiquel data leave no room for improvement! Indeed, their model (using only distance) seems to be more accurate than intelligibility testing itself.

An interesting feature of that model is that it predicts full intelligibility for any pair of villages less than 12.6km. apart. Walter and Echerd suggest that this distance is the radius of an "interaction zone", defined as "that geolinguistic zone in which a person moves with sufficient freedom and regularity so that he characteristically attains and maintains complete intelligibility of communication with all those (with) whom he comes in contact."

2.3.2 The Sabah data

In the present study, simple straight-line measurements on a map are used as the distance measure. The data set is reduced to include just the three major indigenous language families of Sabah: Dusunic, Murutic and Paitanic (see language map in Figure 8). However, because the bulk of the survey was focused on these groups, this subset includes 700 cases, or 88.6% of the full data set.

Taking the data as a whole, we find the expected negative correlation between distance and intelligibility but the correspondence is weak (see Figure 9).

Study	N	Corr	%EV	Equation
Full data	700	-.444	19.71	INT = 86.08 - 1.02 DIST
Excl. Hometown scores	592	-.316	9.97	INT = 81.03 - 0.73 DIST

Figure 9: Correlation between distance and intelligibility in Sabah (Dusunic, Murutic and Paitanic families only)

The weak correlation between INT and DIST is partly a result of the survey design, as discussed above, and partly due to the mixture of groups from different language families in many areas of the state.

When we take various subsets of the data, more interesting patterns emerge. For the Murutic subset (both Speaker and Hearer belonging to Murutic language groups), distance is a much better predictor of intelligibility than lexical similarity is. Figure 10 shows that 44% of the variation in INT can be explained as a function of the variation in DIST, compared with only 15% for LEX.

Study	N	Corr	%EV	Equation
INT vs. DIST	96	-.664	44.09	INT = 92.43 - 1.68 DIST
INT vs. LEX	96	.396	15.68	INT = 0.96 LEX + 5.32
LEX vs. DIST	96	-.349	12.20	LEX = 102.43 - 2.97 DIST

Figure 10: Murutic subset, distance study (Hometown scores excluded)

For the Dusunic language family (Figure 11), the correlation between INT and DIST is stronger than that shown in Figure 9, line 2, for the data set as a whole, but still not very high. LEX is a slightly better predictor of intelligibility than is DIST, but neither LEX nor DIST alone can account for even 30% of the variation in INT. This effect is probably due to the fact that sociological factors, e.g. relative prestige differences, are more extreme in the Dusunic family than in the other language families of Sabah.

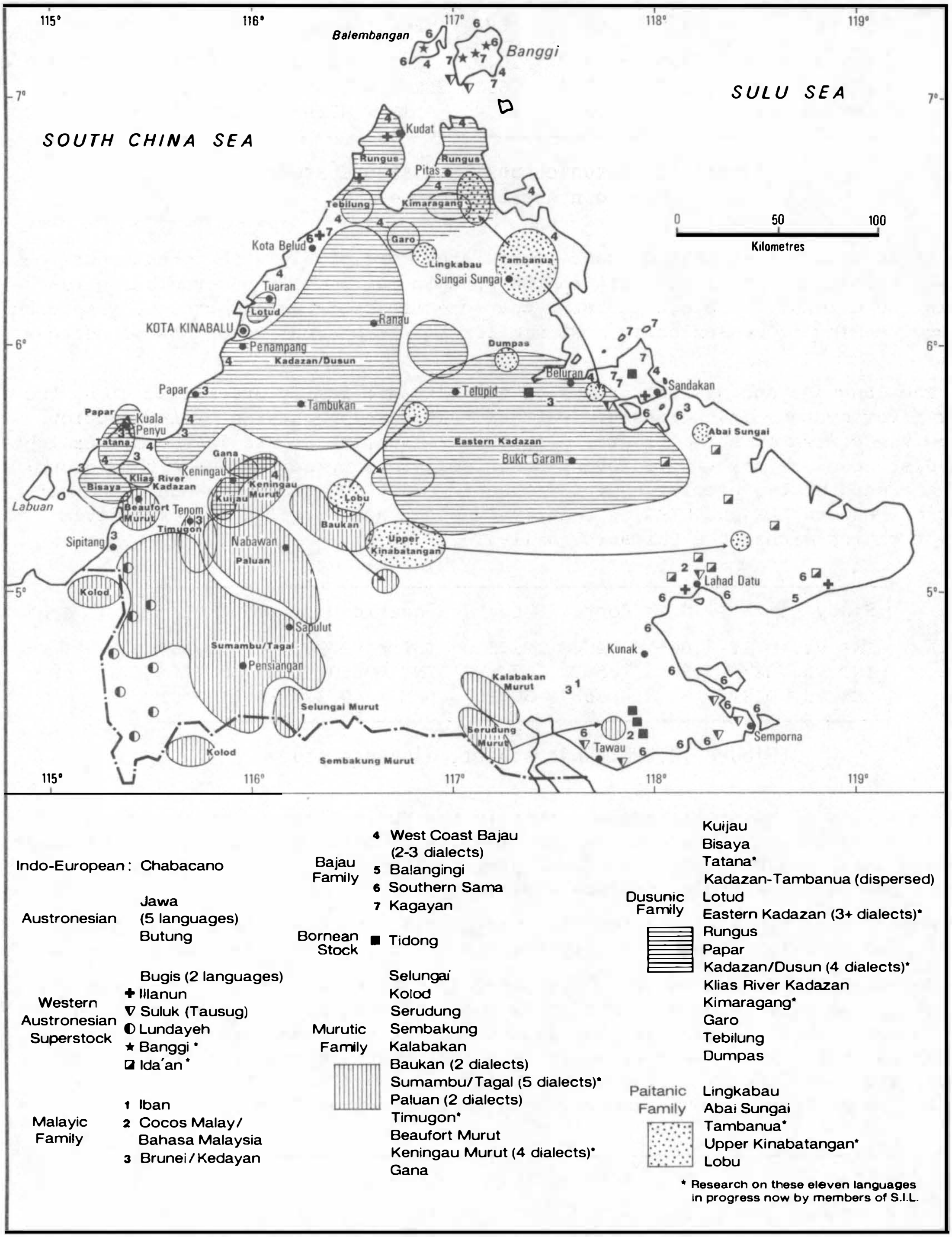


Figure 8: Languages of Sabah

Study	N	Corr	%EV	Equation
INT vs. DIST	325	-.447	20.02	INT = 89.52 - 1.16 DIST
INT vs. LEX	325	.514	26.37	INT = 1.40 LEX - 30.98
LEX vs. DIST	325	-.505	25.49	LEX = 97.69 - 1.89 DIST

Figure 11: Dusunic subset, distance study
(Hometown scores excluded)

It is interesting that distance correlates more highly with cognate percentage than with intelligibility in the Dusunic family. This correlation reflects the extensive dialect chaining characteristic of Dusunic groups, with each village tending to be lexically most similar to its nearest geographical neighbours.

The language and dialect groups of the Paitanic family are spread along the major river systems in the eastern part of the state. Patterns of interaction follow the course of these rivers. For this reason, as Figure 12 shows, straight-line distances are not very relevant to intelligibility scores among this group; LEX is a much better predictor of INT than is DIST. On the other hand, the correlation between LEX and DIST is relatively high, again reflecting extensive dialect chains within the Paitanic family.

Study	N	Corr	%EV	Equation
INT vs. DIST	86	-.464	21.52	INT = 82.98 - 0.60 DIST
INT vs. LEX	86	.582	33.84	INT = 0.94 LEX - 0.74
LEX vs. DIST	86	-.608	36.95	LEX = 99.37 - 1.31 DIST

Figure 12: Paitanic subset, distance study

Rungus is a Dusunic language spoken in the Kudat Division in the northern part of the state. The Rungus are the majority population group in most of their language area, and Rungus is the local prestige language, the church language even in several non-Rungus speaking areas.

When Rungus tapes are tested with non-Rungus subjects, distance is a much better predictor for intelligibility than is lexical similarity (see Figure 13-a).

However, for Rungus subjects listening to non-Rungus tapes, LEX is a better predictor than DIST (Figure 13-b). This suggests that Rungus speakers do not tend to learn other languages; they understand dialects that are linguistically similar to their own. However, other language groups in the area tend to learn Rungus, and those groups living closest to the Rungus learn it best. This is exactly the pattern we would expect for a local prestige language.

Study	N	Corr	%EV	Equation
INT vs. DIST	20	-.707	49.99	INT = 69.61 - 1.22 DIST
INT vs. LEX	20	.498	24.78	INT = 1.27 LEX - 34.98

Figure 13-a: Rungus speaker, non-Rungus hearer

Study	N	Corr	%EV	Equation
INT vs. DIST	31	-.318	10.09	INT = 66.68 - 0.59 DIST
INT vs. LEX	31	.614	37.75	INT = 1.08 LEX - 18.56

Figure 13-b: Non-Rungus speaker, Rungus hearer

Penampang Kadazan (the subdialect of Coastal Kadazan spoken in western Penampang District) is also a high-prestige dialect. Like the Rungus, the Penampang Kadazan dialect group has a high awareness of its identity as a group. It is larger in population than the Rungus, and is the most politically influential of the indigenous language groups in Sabah. Penampang Kadazan is used by the Catholic church in many areas of the state, and is also used in newspapers, magazines, radio, etc.

The pattern of intelligibility for other groups listening to Penampang Kadazan tapes (Figure 14) is even more striking than in the Rungus case. In terms of percentage of explained variation, distance is twice as good a predictor of intelligibility as lexical similarity is. However, no pattern emerges from what little data are available (only 13 cases) for Penampang Kadazan listening to other dialects.

Study	N	Corr	%EV	Equation
INT vs. DIST	44	-.802	64.39	INT = 94.79 - 1.90 DIST
INT vs. LEX	44	.579	33.58	INT = 1.83 LEX - 64.28

Figure 14: Penampang Kadazan speaker, non-Penampang Kadazan hearer

Another interesting case study is Kuijau, a Dusunic language heavily mixed with Murutic vocabulary. The Kuijau are a lower-prestige and somewhat scattered group living among the various Murutic groups of the Keningau District. The IT scores for Kuijau subjects listening to other languages show a surprising *positive* correlation between DIST and INT; in other words, the farther away a group lives, the better the Kuijau understand them (Figure 15).

Study	N	Corr	%EV	Equation
INT vs. DIST	16	.444	19.74	INT = 0.86 DIST + 66.46
INT vs. LEX	16	-.106	1.13	(no significant relationship)

Figure 15: Kuijau hearer, non-Kuijau speaker

This pattern is partly due to dialect geography, partly the result of the survey design. The closest neighbours of the Kuijau are speakers of various Murutic dialects; but the Kuijau are linguistically more similar to the Central Dusun of neighbouring Tambunan District. The Kuijau have contact with Dusun immigrants from Tambunan, and often refer to themselves as Dusun.

Kuijau subjects were tested with geographically close but linguistically distant Murutic tapes; and with geographically distant but linguistically closer -

and highly prestigious - Central Dusun and Penampang Kadazan tapes. Thus the expected relationship between IT scores and distance was reversed.

2.4 Summary

Linguistic similarity is obviously an important factor in predicting how well members of one dialect group will understand speakers of another group. Lexical similarity (cognate percentage) is a useful index of linguistic similarity, and the Sabah data exhibit the expected correlation between cognate percentage and intelligibility. However, simple regression analysis of the data cannot determine the precise mathematical relationship between these two measurements, for reasons that will be discussed further in section 4 below.

Geographical distance is related to language contact, and is found to be a significant factor in situations where bilingualism is a major component of the measured intelligibility. Other social factors affecting intelligibility in particular situations in Sabah will be discussed in the following section.

3. PATTERNS OF BILINGUALISM: TWO CASE STUDIES

As shown in section 2.3 above, there are some subsets of the data for which distance is a better predictor of intelligibility than is lexical similarity. This is particularly true in the case of prestige dialects, such as Rungus and Penampang Kadazan.

The distance between two groups is related to the opportunity for contact between them, thus the component of intelligibility due to social factors rather than the component due to linguistic similarity is in focus here. The relatively high correlation between distance and measured intelligibility in some sets of data suggests that the IT scores reflect not only inherent intelligibility but also a significant amount of learned comprehension or bilingualism.

In studying inherent intelligibility, it is appropriate to use average IT scores, because inherent intelligibility is assumed to be fairly uniform throughout a speech community. However, average IT scores are useless for investigating bilingualism. There is often a wide range of scores among different segments of the population, and that is precisely the phenomenon that we want to investigate. What members of group X understand language Y and to what extent?

To get this information, we must use individual IT scores. This greatly increases the volume of data, and the amount of work involved in processing and analysing the data. Two case studies are discussed here: outsiders' comprehension of oral texts in Rungus and Penampang Kadazan. Even for these two examples, it has not been possible to process all the data available. Out of roughly 200 non-Rungus individuals tested with Rungus tapes, a representative sample of 88 individuals was chosen for this part of the study. Out of roughly 440 non-Penampang Kadazan subjects tested with Penampang Kadazan texts, a sample of 192 individuals was selected. For each sample group, an attempt was made to include at least one village from each dialect group where Rungus or Penampang Kadazan was tested.

3.1 Correlation study

Each subject's score on the Casad-style intelligibility test is taken here as a measurement of his ability to understand Rungus or Penampang Kadazan, as the case may be. Therefore, the INT values for each subject range from 0 to 10, representing the number of questions answered correctly.

Various other measurements were included in this phase of the study, including: LEX, the cognate percentage between the subject's dialect and the test dialect; DIST, the straight-line distance from the subject's village to the village where the test tape was recorded; MALAY, the subject's individual score on the Bahasa Malaysia test, ranging from 0 to 10; AGE, the subject's age in years; SEX, subject's gender; EDUC, amount of schooling in years; TRAVEL, extent of travel outside home language area, on a scale of 1 to 5; BIRTH, relative distance of birthplace from present residence, i.e. from village where the subject was tested, on a scale of 1 to 3.

For this part of the study, the Spearman rank-order coefficient r_s is used as a measure of correlation. Unlike the Pearson product-moment correlation coefficient used above, the rank-order coefficient is a non-parametric measurement which does not assume that the data is normally distributed, and which can be applied to simple ranking scales as well as pure quantitative measurements. As with Pearson's r, the values of r_s range from -1 to 1, with $r_s = Ø$ indicating that no relationship is measurable between the two variables.

Correlation analysis is useful for picking out linear relationships among the variables being studied. Figure 16 shows the correlation coefficients (r_s) and associated measures of significance for pairs of variables which seem most strongly related in the Penampang and Rungus data.

Significance[5] is a measurement of the likelihood that a particular pattern is purely accidental. For example, a significance value of .01 indicates that there is one chance in a hundred that an observed association is purely random; or, 99% certainty that it reflects some real characteristic of the population from which the data was drawn. A significance value of zero indicates perfect confidence, i.e. zero probability that the pattern is due merely to chance. A value of 1 is the worst possible case; it means that there is no room for doubt that the observed pattern is accidental.

In general, the significance value is closer to zero, i.e. better, for stronger correlations (larger values of r or r_s) and for larger data sets. A correlation of $r_s = .25$ may be significant in the Penampang study, with 191 cases, but not in the Rungus study with only 87 cases.

Simons (1977) recommends using a .10 level of significance (i.e. a 90% confidence level) for determining significant differences in lexicostatistics. For the purposes of this study, I would consider any correlation with a significance value below .01 as definitely significant, and any below .10 as being worthy of further investigation.

	Penampang Kadazan		Rungus	
VARIABLES	CORR	SIG	CORR	SIG
INT/LEX	---	---	.3853	.001
INT/DIST	-.4788	.001	-.4719	.001
INT/MALAY	.2786	.001	---	---
INT/EDUC	.2907	.001	---	---
INT/TRAVEL	.2286	.002	(-.1986	.067)
LEX/DIST	---	---	-.3483	.001
MALAY/EDUC	.5178	.001	.7433	.001
AGE/EDUC	-.6097	.001	-.5499	.001
MALAY/AGE	-.3379	.001	-.3959	.001
DIST/EDUC	(-.1392	.055)	(.2053	.056)

Figure 16: Correlation and significance values for Penampang Kadazan and Rungus bilingualism studies

Key:

CORR = Spearman rank-order correlation coefficient
SIG = significance (two-tailed test)
INT = individual IT score
LEX = cognate percentage between subject's dialect and test dialect
DIST = linear distance between test point and village where test tape was recorded
MALAY = individual score on Bahasa Malaysia test
EDUC = years of formal education
TRAVEL = extent of travel outside subject's dialect area
AGE = subject's age in years
( ) = marginally significant relationship

The first two lines of Figure 16 confirm the results of section 2.3, which were based on average INT scores. The correlation between INT and DIST in the Penampang data is not so striking here, and there is now no significant correlation between INT and LEX. This is because using individual IT scores allows differences in age, education, travel, etc. to overshadow the relatively smaller effect of lexical similarity. However, for the Rungus test, the INT vs. LEX and INT vs. DIST correlations were roughly the same as those computed from average IT scores.

We are primarily interested in factors which correlate highly with INT. However, the strongest patterns in the data involve three highly interrelated variables: age, amount of education, and ability in the national language, Bahasa Malaysia (lines 7, 8 and 9).

The high correlation between ED and MALAY tells us that the more schooling a person has, the better he will understand Malay. This pattern is especially striking in the Rungus study (r_s = .7433). The negative correlation between AGE and ED says that the younger people are generally better educated than their elders. This is the strongest single relationship in the Penampang testing sample, r_s = -.6097. And the negative correlation between AGE and MALAY says that, on the average, young people understand Malay better than their elders.

A crucial difference between the two studies shows up in the correlations of INT with ED and MALAY (lines 3 and 4). Penampang Kadazan is a state-wide prestige language, used on the radio, in newspapers, at political rallies, etc. A significant body of Penampang Kadazan literature exists. Among non-Kadazans, the better educated and more upwardly mobile (those who tend to speak Malay better) understand the Penampang dialect better.

The prestige of Rungus, by contrast extends over a fairly limited area. People learn Rungus at weekly markets and in other traditional contexts of social interaction, not through the mass media. Only a very small body of Rungus literature exists, and it is not widely distributed even within the Rungus community. Thus, education and ability to understand Malay are irrelevant to a person's ability to understand Rungus.

A related difference is apparent in the relationship between INT and TRAVEL (line 5). The weak but fairly significant positive correlation in the Kadazan study (r_s = .2286, significance = .002) indicates that people who have travelled further from their native language areas tend to understand Penampang Kadazan better than those who stay at home. The marginally significant *negative* correlation in the Rungus study (r_s = -.1986, significance = .067) suggests that those who stay at home tend to understand Rungus better than those who travel.

This difference could be related to the urbanisation of Sabah, one of the most important population trends in the state today. The Penampang Kadazan dialect area is contiguous to the state capital, Kota Kinabalu. Of the three major towns in Sabah, the capital naturally exerts the strongest attraction on people from outlying districts. For most of the people in the Penampang test sample, when they travel, they travel towards Penampang.

The Rungus area, on the other hand, is one of the least developed areas in the state. The non-Rungus who leave their own areas have little incentive to go north towards the Rungus area, and as they go south towards the capital, they are cut off from contact with the Rungus language.

Finally, there is a marginally significant correlation in both studies between DIST and ED (line 10). This suggests that people living closer to Penampang (and thus to the capital) tend to get more education than those farther in the interior (r_s = -.1392). People who live closer to the Rungus area, i.e. farther north, tend to get *less* education than those who live to the south (r_s = 0.2053).

In addition to the variables listed in Figure 16, data were collected for each subject about his or her spouse's first language. A simple scale was used to rate the degree of difference between the spouse's language and the subject's mother tongue: 1 if both were native speakers of the same dialect, 2 if they spoke different dialects or languages within the same language family (e.g. both Dusunic or both Murutic), 3 if they spoke dialects from different language families (e.g. Kadazan and Bajau, or Murut and Malay).

Somewhat surprisingly, no correlation was found between linguistic diversity in the marriage and ability to understand either Rungus or Kadazan. However, in both studies it appears that people who marry outside their own language group tend to be better educated than those who marry within the group (Rungus study: r_s = .3596, sig = .002; Penampang study: r_s = .2693, sig = .001). There is also a tendency for better educated people in both sample groups to travel more widely than their less-educated counterparts, and a weak tendency for people who marry outside the language group to be more widely travelled than those who marry within the group. Finally, in both studies, subjects who married outside their own language group scored higher on the Malay test than those who married within the group (Rungus study: r_s = .3277, sig = .005; Penampang study, r_s = .2092, sig = .018).

3.2 Tabulation of the data

Correlation analysis can reveal linear trends in the data, but a simple tabulation of the data is helpful in interpreting these trends, and in finding other, non-linear, relationships.

One obvious pattern which correlation analysis could not reveal is the fact that men, on the average, understand Penampang Kadazan better than women.[6] Figure 17 shows the breakdown of scores by sex for both studies; notice that in the Rungus test, there was virtually no difference in scores between the sexes.

	Penampang Kadazan			Rungus		
SEX	N	MEAN	STD.DEV.	N	MEAN	STD.DEV.
Male	111	7.14	1.900	57	5.46	1.864
Female	81	6.54	2.060	31	5.37	2.152
Total	192	6.89	1.965	88	5.43	1.970

Figure 17: Breakdown of intelligibility scores by sex for the Penampang Kadazan and Rungus studies

Key:

N = number of cases
MEAN = average score
STD.DEV. = standard deviation

The difference between men's and women's scores in the Penampang study may be related to the correlation mentioned in section 3.1 between extent of travel and ability to understand Kadazan. In both sample groups, men on the average have travelled more widely outside their home language area than women.[7] In the Rungus study, men are also better educated than women (average 3.79 years for men, 2.68 years for women);[8] but we have already seen that there is no correlation between years of education and ability to understand Rungus. In the Penampang study, the difference in education is not statistically significant (the mean for women being slightly higher than that for men). Thus the difference between men's and women's scores on the Kadazan test are not related to educational differences.

Figure 18 shows a breakdown of scores by occupation. The "agricultural" category includes farmers and fishermen; in these two samples, most people in this category are rice farmers. "Government employee" includes village headmen, native chiefs, teachers, community development officers and civil servants (all those tested were also residents of the villages where data were collected and native speakers of the dialect spoken in their village). "Other" includes small business men, students, unemployed, etc.

	Penampang Kadazan			Rungus		
OCCUPATION	N	MEAN	STD.DEV.	N	MEAN	STD.DEV.
Agriculture	65	7.01	1.882	47	5.55	1.877
Government	23	7.59	2.175	15	4.37	2.254
Other	93	6.69	2.031	26	5.81	1.647

Figure 18: Breakdown of intelligibility scores by occupation for the Penampang Kadazan and Rungus studies

The interesting comparison here is in the scores of government employees. This group did better than either farmers or "others" on the Penampang Kadazan test, but scored lower than either of the other categories on the Rungus test.[9]

Figure 19-a presents a breakdown of INT scores based on how extensively a subject had travelled outside his own language area. Category 1 indicates that the subject had never left the language area; 2 that he/she had travelled only to neighbouring districts; 3 indicates extensive travel within the state, e.g. from west coast to east coast; 4 indicates travel outside the state, generally to Sarawak, West Malaysia or Singapore; 5 indicates that the subject had lived for extended periods of work or study outside his/her own language area (whether in Sabah or elsewhere).

These results confirm the correlation findings (see Figure 16, line 5) showing that the more widely travelled subjects understood Penampang Kadazan better, while those who had travelled less understood Rungus better. In category 5 (which deals with residence rather than travel) is ignored, the trend lines in Figure 19-b are monotonic in both studies, and strictly monotonic for the Penampang study.

	Penampang Kadazan			Rungus		
TRAVEL	N	INT	STD.DEV.	N	INT	STD.DEV.
1	23	6.17	2.278	19	5.68	1.407
2	90	6.67	1.774	46	5.68	1.872
3	50	7.14	2.124	8	4.44	2.468
4	10	7.65	1.718	3	4.17	1.312
5	17	7.65	1.845	10	4.60	2.289

Figure 19-a: Breakdown of INT by travel for the Penampang Kadazan and Rungus studies

Key:

N = number of cases
INT = average of individual IT scores
STD.DEV. = standard deviation
TRAVEL = extent of travel outside subject's home language area:
1 = never left language area
2 = travel only to neighbouring districts
3 = state-wide travel
4 = travel outside Sabah
5 = live for one year or more outside home language area

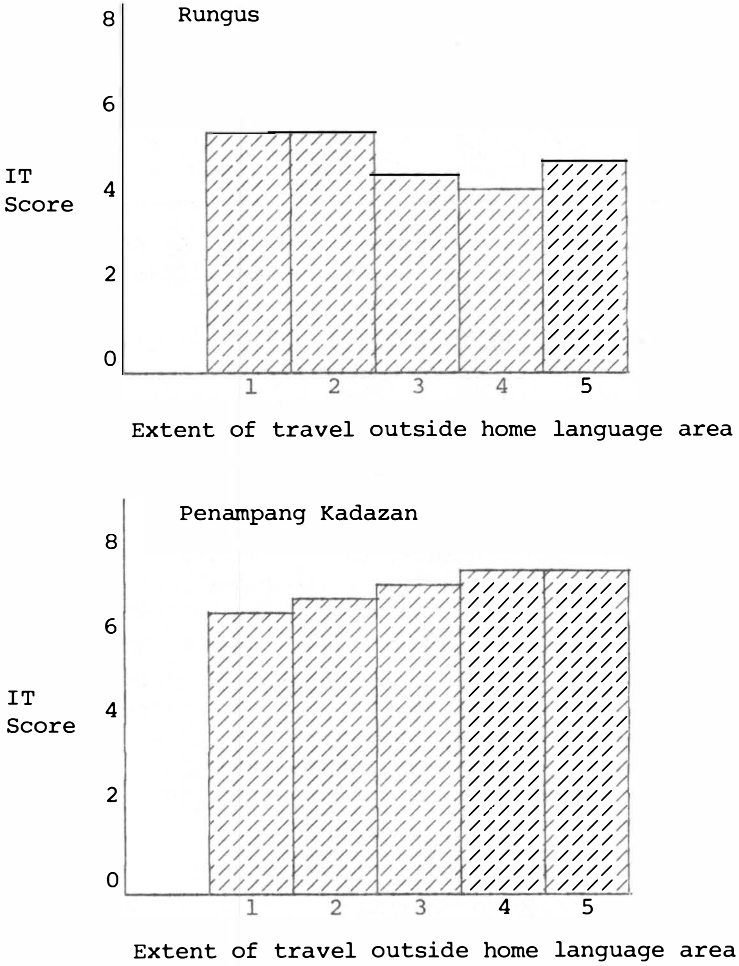


Figure 19-b: IT score vs. travel for Rungus and Penampang Kadazan studies

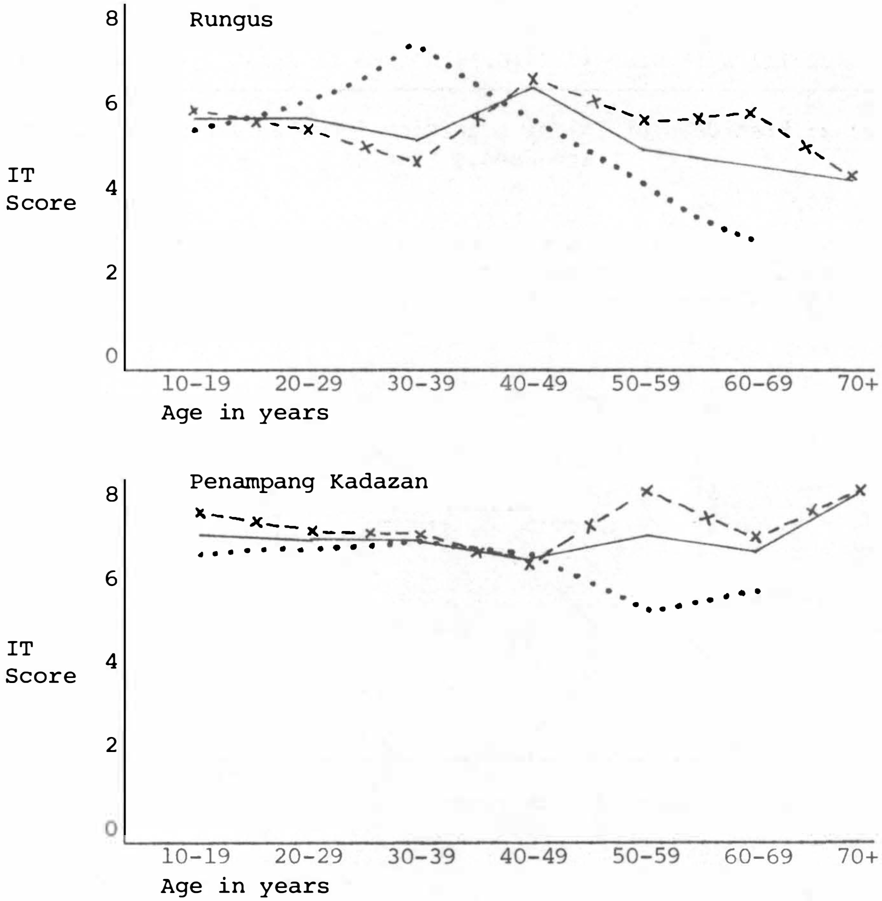
Correlation analysis revealed no linear relationship between AGE and INT for either test. Figure 20-a shows a breakdown of scores by age and sex, and Figure 20-b shows the same information as a graph. There is no clear trend for the sample as a whole in either graph, but in both studies women tend to score highest in middle age. In both studies, the highest mean score for women is in the 30-39 age group, and women less than 30 score higher than women over 50.

	Penampang Kadazan		Rungus	
AGE	N(m/f)	INT(men/women)	N(m/f)	INT(men/women)
10-19	33 (13/20)	6.97 (7.58/6.58)	14 (7.7)	5.61 (5.71/5.50)
20-29	60 (33/27)	6.89 (7.05/6.70)	24(14/10)	5.71 (5.46/6.05)
30-39	33 (20/13)	7.11 (7.15/7.04)	15 (12/3)	5.13 (4.54/7.50)
40-49	32 (22/10)	6.59 (6.57/6.65)	17 (13/4)	6.20 (6.38/5.63)
50-59	17 (10/7)	6.94 (8.10/5.29)	7 (4/3)	4.57 (5.25/3.67)
60-69	14 (10/4)	6.50 (6.85/5.63)	9 (5/4)	4.34 (5.50/2.88)
over 70	3 (3/0)	8.00 (8.0/-)	2 (2/0)	4.25 (4.25/-)

Figure 20-a: Breakdown of INT by age group for the Penampang Kadazan and Rungus studies

Key:

N = number of cases (men/women)
INT = average of individual IT scores (men/women)



(..... = women; --x-- = men; ——— = total)

Figure 20-b: IT score vs. age group for Rungus and Penampang Kadazan studies

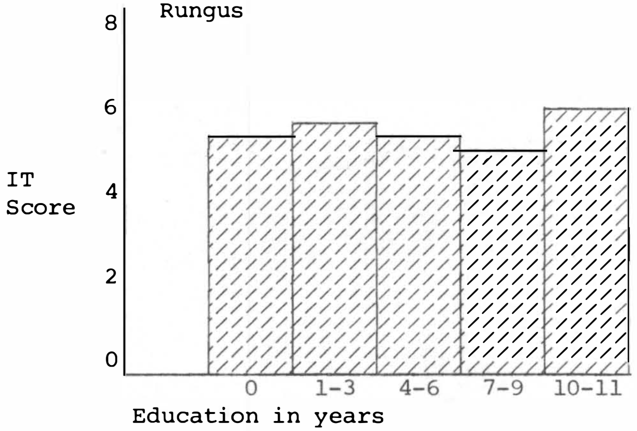
Figure 21-a shows a breakdown of scores by amount of education (in years). The graph in Figure 21-b confirms the positive correlation mentioned in Section 3.1 between INT and ED for the Penampang Kadazan test. No trend is apparent in the Rungus study. It is apparent that the subjects who took the Penampang Kadazan test were, on the average better educated than those who took the Rungus test. A higher percentage of those in the Rungus study had no education (50% vs. 34%), and a higher percentage of those that had been to school never got beyond Primary Six (66% vs. 56%). This is a significant difference between the two sample sets, and makes a rigorous comparison of specific results between the two studies more difficult. However, it should not affect the interpretation of trends within each study.

	Penampang Kadazan		Rungus	
EDUC	N(m/f)	INT(men/women)	N(m/f)	INT(men/women)
0	66 (35/31)	6.27 (6.74/5.73)	44 (26/18)	5.46 (5.75/5.03)
1-3	17 (14/3)	6.74 (6.78/6.50)	3 (2/1)	5.50 (4.75/7.00)
4-6	53 (33/20)	6.92 (7.20/6.48)	26 (18/8)	5.39 (5.33/5.50)
7-9	45 (25/20)	7.63 (7.56/7.73)	9 (5/4)	5.06 (4.10/6.25)
10-11	9 (2/7)	7.39 (8.75/7.00)	6 (6/0)	5.92 (5.92/-)
12-over	2 (2/0)	8.50 (8.50/-)	-- --	-- --
total	192(111/81)	6.89 (7.14/6.54)	88 (57/31)	5.43 (5.46/5.37)

Figure 21-a: Breakdown of INT by education for the Penampang Kadazan and Rungus studies

Key:

N = number of cases (men/women)
EDUC = years of formal education
INT = average of individual IT scores (men/women)



(Figure 21-b continued over)

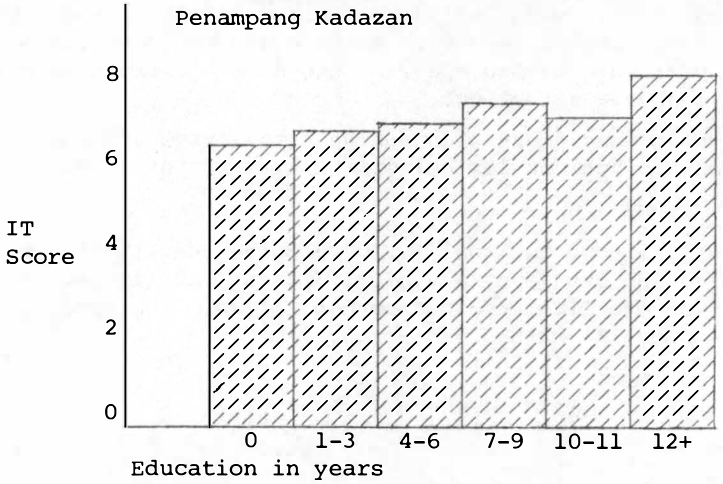


Figure 21-b: IT score vs. education for Rungus and Penampang Kadazan studies

3.3 Summary

We have presented a detailed comparison of two superficially similar language situations, Rungus and Penampang Kadazan. It is clear that a very different set of criteria would be used to predict a person's ability to understand Rungus from that which would be used to predict comprehension of Kadazan. In both cases, geographical distance is the single most important factor, and the strength of the correlation between distance and IT score is remarkably similar in the two studies (see Figure 16 above).

Aside from distance, however, no single factor is found to correlate highly with intelligibility in both of the studies. The striking differences between the two cases give us some insight into the complexity of the problem of developing a general predictive model for intelligibility. A fairly sophisticated model would be needed to account for the Penampang and Rungus situations, to say nothing of the hundred or so other major dialects in Sabah. Such a model would probably have reference to lexical similarity, distance, and other factors not measured in the Sabah survey such as language use, language contact and language attitudes.

4. OBSTACLES TO THE DEVELOPMENT OF A PREDICTIVE MODEL

Simon (1969) defines *secondary analysis* (or "data dredging") as "searching for new relationships in existing data", trying to shed light on new problems with data originally collected for some other purpose. Sections 2 and 3 of this paper present a secondary analysis of the Sabah survey data, a statistical analysis aimed at discovering relationships between IT scores and various other factors.

Developing the kind of predictive model for intelligibility that Simons (1979) proposes will necessarily involve secondary analysis of linguistic and sociolinguistic survey data. No one has ever (to my knowledge) designed and carried out a dialect intelligibility survey solely for the purpose of developing a general model of intelligibility. Such surveys are so expensive and exhausting that they are only carried out when specific pieces of information are needed about a particular language situation.

Simons (1979:ch.5) used regression analysis to derive a formula expressing the general relationship between lexical similarity and intelligibility:[10]

$$\text{INT} = 1.67 \ \ \text{LEX} - 66.7$$

Regression analysis is a powerful tool for developing quantitative models, but when it is applied to the data from a typical dialect intelligibility survey, three sources of error are likely to be present: 1) sampling bias; 2) non-normal distribution of the data; 3) masked variation.

4.1 Sampling error

When a researcher tries to generalise any observed pattern from a particular data sample to an entire population, it is crucial that the sample be fair[11] and representative. The normal way to ensure fairness is to use some form of random sampling; and if the sample size is big enough, a random sample is also very likely to be a good representative of the population as a whole.

Unfortunately, no meaningful language survey could be designed based purely on random sampling. The nature of the information required and the complexity of both the data-gathering and the interpretation stages of the task force us to make careful, principled selections at each step of the survey design.[12]

In a study concerned with estimating a particular measurement (e.g. national language comprehension) for an entire population, the sampling procedure is fairly straightforward - or as straightforward as any social research can be. However, in investigating the general relationship between LEX and INT, the sampling problem is several orders of magnitude more complicated. The researcher must choose a set of subjects from a set of villages representing a set of dialect groups, and for each subject choose a set of measurements (i.e. a set of dialects for which the subject's comprehension will be tested).[13]

In terms of survey design, the goals of the survey determine how choices are made at each level. But in terms of statistical analysis or developing a general model, each "principled decision" becomes a possible source of sampling bias.

On the other hand, Simon (1969:263) points out that random sampling is unnecessary for some kinds of research. Many biological traits, for example, are so homogeneous throughout an entire population that any sample at all is adequate; measurement over even a small, non-random sample can be generalised over the entire population.

We have already mentioned the difference between inherent intelligibility and bilingualism in this regard. If X and Y are related dialects, and group X can be assumed to have had no exposure to dialect Y, then every normal adult native speaker of X should have roughly the same ability to understand Y. A relatively small sample is adequate for measuring inherent intelligibility, but this is emphatically not true for measuring bilingualism.

The relationship between linguistic similarity and intelligibility can be thought of as determined by the innate language faculty common to all humans. In studying this relationship, a small, non-random sample is adequate if the IT scores reflect only inherent intelligibility, or if cases of bilingualism can be reliably excluded, as Simons (1979) attempted to do. Unfortunately, most dialect surveys include mixed intelligibility and bilingualism situations, i.e. related dialects whose members have significant amounts of contact with each others' speech varieties. This seems to be the normal pattern in most areas of Sabah. In such cases, the validity of generalising any observed relationships depends heavily on the reliability of sampling methods used in collecting the data.

4.2 Distribution of the data

Regression analysis assumes that the data are more or less normally distributed. In section 2.2, we showed that the data from the Sabah survey violate this assumption, especially in the distribution of IT scores. The effect of this skewing is to create a regression line that does not appear to fit the data, and to reduce the correlation coefficient, r. The measured correlation is further reduced by the truncated range of LEX values.

Both the skewing of INT values and the truncated range of LEX are characteristic of most Casad-type dialect intelligibility surveys. Intelligibility is rarely tested where it is not expected to exist, and most surveys have focused on the cognate range of 60-90 percent.

Experience in Sabah has shown that it is very difficult to measure low levels of intelligibility accurately. Subjects began to lose interest in a test when they could not follow the story easily, and some people refused to listen to stories they felt they could not understand.

It is probably easier to use some written form of intelligibility testing, rather than tests of oral comprehension, to measure the low end of the INT scale. However, use of written testing materials introduces an obvious sampling bias by selecting only literate subjects. The distribution of INT values could be made closer to normal by varying the level of difficulty of the texts and questions so that only someone approaching native speaker fluency would be expected to score 100 percent, while 50 percent would correspond roughly to the threshold between language and dialect distinctions. However, increasing difficulty of oral tests also reduces the useful range of the test in terms of LEX. Subjects who were willing to listen to the easy test stories used in Sabah, even when the test dialect was only 65 percent cognate with their own, would be less willing to listen to harder stories in that same dialect.

4.3 Averaging

In most survey reports, the average of the raw IT scores for a given village or dialect group is used as the index of that group's ability to understand some other dialect. The individual scores are neither reported nor (in most cases) used in analysing the data.

As mentioned above, the mean IT score is a valid index of inherent intelligibility, though not of bilingualism. Even so, when mean scores are used in correlation or regression analysis, the results are less accurate than they would be if raw (i.e. individual) IT scores were used.

Correlation and regression analysis are based on calculations of the amount of variation within the data set. When mean scores are used, the variation within each test set of 10 subjects from a particular village is masked. If variations within each test set are too extreme, the correlation and regression results based on mean scores will be meaningless. The potential magnitude of the error increases with the size of the data set (total number of cases).

With only 10 subjects per test, and raw scores ranging from 0 to 10, the amount of possible variation within each test set is fairly tightly constrained. However, with a data set as massive as that from the Sabah survey, this is still a potentially significant source of error.

4.4 Summary

Regression analysis is potentially the best tool for developing descriptive and predictive models for intelligibility. The results are valid and generalisable for IT scores which reflect only inherent intelligibility.

For studies involving bilingualism or mixed intelligibility and bilingualism, the validity of the results will depend on the reliability of the sampling method. No language survey can be based on a purely random sampling procedure, but studies of bilingualism in particular language situations will be more reliable than broad-scale analyses of bilingualism in general, because the sampling problems are much more manageable.

In studying the relationship between LEX and INT, the accuracy of analysis will be reduced if the distribution of either variable differs greatly from the normal distribution. Survey design plays a crucial part in shaping the distribution of the data collected, and traditional dialect intelligibility surveys seem especially prone to skewing the distribution of IT scores toward the higher values. New methods of measuring intelligibility need to be developed to reach an adequate range of LEX while producing approximately normal distributions of both LEX and INT, without introducing new sources of sampling bias (e.g. relying on literate subjects).

Future analysis of intelligibility data should work directly with raw (individual) scores, rather than average or aggregate scores. This approach is planned for further research, now in progress, using the Sabah data.

All three of the problems discussed above become more serious as the data set becomes larger. Perhaps this is why the cumulative effect was so noticeable for the massive Sabah survey data set. The most promising way to minimise these problems may be to adopt the approach of Simons (1979), i.e. by comparing the results of many relatively small studies. At any rate, more studies are needed concerning the nature of intelligibility and the various factors which affect it in specific language situations.

Acknowledgements

Among the many people who have helped me in various ways with this project, special thanks are due to: Chiang On Khiong and Alfred Chan, of the Sabah State Computer Service Unit, Lim Suk Lee, Tim Friberg, Sue Harris, Priscilla Jones, David Moody, and Gary Simons.

NOTES

1. It is well known that the same factors which promote language learning, such as language contact and positive language attitudes, also promote lexical borrowing over longer time periods. So linguistic similarity is not, strictly speaking, independent of the sociological component of intelligibility. However, the effects of this dependence are assumed to be very small in comparison with other factors involved.

2. The skewness measurement for INT (excluding hometown scores) was -0.526; for LEX, it was -0.276. The kurtosis value for INT was -0.504; for LEX, 1.145. For normally distributed data, skewness = 0, kurtosis = 1.0.

3. In some areas of the Philippines, loss of intelligibility is reported even in the 95% cognate range, due to different use of grammatical markers and other particles (Chuck Walton, personal communication).

4. Simons (1979) suggests 40%.

5. In this study, significance figures are computed using a two-tail test, because no prior assumption is made about the direction (plus or minus) of the correlation.

6. The z-test is used to confirm that the difference between men's and women's scores is statistically significant. For the Penampang scores, the z value is 2.05, indicating that the difference is significant at a confidence level of .02. The difference in the Rungus scores is not significant.

7. In the Kadazan sample, men averaged 2.62 on the travel scale (N = 109, std. dev. = 1.043) while women averaged 2.37 (N = 81, std.dev. = 1.089); z = 1.59, meaning that the difference has a significance of .06.

   For the Rungus sample, men averaged 2.53 (N = 55, std.dev. = 1.331) and women averaged 1.87 (N = 31, std.dev. = 0.763); z = 2.92, indicating a significance of less than .002.

8. z = 1.38, sig = .09.

9. For the Penampang study, the difference between farmers and government employees was significant only at the 0.13 confidence level. The difference between government employees and all other groups was significant at the .04 level.

   For the Rungus study, the significance of the difference between farmers and government employees was below .04; between government employees and others, below .02.

10. One of the strengths of Simons' study was that the raw data came from surveys using a variety of different methods for testing intelligibility. Thus, while INT would be defined operationally as "average score on an intelligibility test", the testing method is not specified. The implied claim seems to be that the relationship expressed in the formula is independent of the testing method used.

11. "Fair" in this sense means that each individual in the population has an equal chance of being included in the sample.

12. Choosing the sample for the Sabah survey involved several levels of decisions. The first question was, at which villages should data be collected? The basic goal was to get data from at least one village from every dialect

group that was either reported (by local residents) or observed (by survey technicians) to be distinctive; and to get a geographically representative sampling of villages from the larger dialect groups. When possible, linguistically homogeneous villages were chosen; but since this tends to be the norm in Sabah, it was not a major constraint. Other factors considered included reported "purity" of language, reported or observed prestige factors, migration patterns (preference being given to long term residents of an area, rather than recent arrivals from other language areas), accessibility and the results of previous survey work by other scholars. The advice and guidance of local government officials was crucial in these decisions, particularly in the first phase of the survey (collection of wordlists and texts).

For each village where intelligibility testing was done, the second question was: which tapes (i.e. which dialects) should be tested here? As discussed in 2.1 above, lexical similarity and distance were primary considerations in determining which dialects should be tested with each other. A further consideration was the desire for comparability between tests. As much as possible, one good test was used to represent a particular dialect everywhere that dialect was tested (rather than a random choice among the tapes recorded in that dialect). For example, the tape from Kampung Bunduon, Penampang, was used in all the Penampang Kadazan testing discussed in sections 2 and 3.

The third level of sampling was the selection of ten individuals to take the test in each village. The strategy called for a rough quota based on age and sex - some old men, some young men, some old women, some young women (no-one under 15). Within these guidelines, the village headmen were generally responsible for finding the subjects.

Each of the factors listed above is a possible source of sampling bias - although some factors could tend to offset each other, e.g. accessibility and prestige vs. "purity" of language.

13. One possible strategy would be a stratified sampling of tests for each subject in the sample group, based on cognate percentages with the subject's own dialect. For a subject from group X, we would divide all other dialects in the state into five sets: 1) all dialects 80-99 percent cognate with X; 2) 70-79 percent cognate with X; 3) 60-69 percent cognate with X; 4) 50-59 percent cognate with X; 5) below 50 percent cognate. The subject would be tested with one dialect selected at random from each group. Needless to say, a survey of this type would be a logistical nightmare, and the results would be virtually useless for any other purpose, such as determining linguistic boundaries or mapping patterns of communication.

REFERENCES

CASAD, Eugene H.

1974 *Dialect intelligibility testing.* Summer Institute of Linguistics Publications in Linguistics and Related Fields, Number 38. Norman, Oklahoma: Summer Institute of Linguistics of the University of Oklahoma.

COLLIER, Ken J.

1977 Predicting intelligibility: a suggested technique. In Loving and Simons, eds, 1977:253-262.

KING, Julie K. and John Wayne KING, eds

1984 *Languages of Sabah: a survey report. PL*, C-78.

KIRK, Roger E.

1978 *Introductory statistics.* Monterey, CA: Brooks/Cole.

LOVING, Richard and Gary F. SIMONS, eds

1977 *Language variation and survey techniques. Workpapers in Papua New Guinea Languages*, vol.21. Ukarumpa, Papua New Guinea: Summer Institute of Linguistics.

MILLER, Carolyn P.

1984 Introduction. In King and King, eds, 1984:51-57.

SANKOFF, Gillian

1969 Mutual intelligibility, bilingualism and linguistic boundaries. *International days of sociolinguistics*, 839-848. Rome: Istituto Luigi Sturzo.

SIMON, Julian L.

1969 *Basic research methods in social science.* New York: Random House.

SIMONS, Gary F.

1977 Tables of significance for lexicostatistics. In Loving and Simons, eds, 1977:75-106.

1979 *Language variation and limits to communication.* Technical Report No.3. Ithaca, New York: Department of Modern Languages and Linguistics, Cornell University.

SMITH, Kenneth D.

1984 The languages of Sabah: a tentative lexicostatistical classification. In King and King, eds, 1984:1-49.

WALTER, Steve and Steve ECHERD

n.d. Interpreting intelligibility asymmetries in dialect survey data. MS.