8.4 The control group study

Jane Edwards and Willem Levelt[5]

Introduction

We noted in the Field Manual (p. 32) the inescapable fact that participation in the ESF project would provide the longitudinal informants with additional target language experience of a somewhat different type and a somewhat greater amount than that they would have obtained in everyday contact alone. They were involved in a continuing relationship with target language speakers, and knew they were important participants in a large research project. Such factors could be expected to increase their motivation to learn the target language. Furthermore, they received practice in the target language, in the form of performing the task activities in the presence of target language speakers during each of the encounters, circumstances which might have enhanced their awareness of their linguistic productions and particular difficulties than would otherwise have been the case.

It was necessary for this reason to assess whether the 'Longitudinals' differed from individuals who had minimal (or no) contact with the project, in terms of the speed and nature of target language acquisition. In addition to providing an indication of how project results could be generalised to individuals not exposed to such experiences, the present investigation may also shed light on the permeability of particular aspects of the acquisition process to environmental influences, or relevant aspects of the input.

For this purpose, comparable data were also gathered from a second group of individuals who were socio-biographically matched to the Longitudinal informants, but who were observed only three times during the entire project. This second group of individuals, who were called the 'initial learner group' in the Field Manual, provided a control with respect to the amount of intensive contact with ESF project researchers and data gathering tasks. The present chapter reports the results of comparisons made across two different activities, and over time, to assess the strength of effects of contact experience.

Two classes of variables were used: (a) linguistic and (b) non-linguistic, or motivational. The linguistic variables used here include the following:

[5] This section is a. version of the control group study presented to the project's Steering Committee in Cambridge, October 1986. A much extended final version of the study will be published elsewhere.

- richness and diversity of the target language lexicon, measured as the number of TL word types in lemmatised samples selected to approximate as closely as possible a length of 200 words (described below);
- degree of reliance on source language, measured as number of occurrences of SL words within otherwise exclusively TL utterances;
- degree of automaticity or idiomaticity in use of the target language, measured as (a) the number of target language expressives (TLXK) and (b) number of target language formulas or collocational expressions (TLFK);
- level of syntactic complexity in use of the target language, measured as the number of uses of target language conjunctions, that is, conventional target language means for marking explicitly the relevance of a particular clause to that which preceded it. This measure includes both 'semantic connectives', that is, the marking of relations between successive parts of an utterance, and 'pragmatic connectives', that is, the contextualising of an utterance with respect to the preceding contributions by other speakers (Gallagher and Craig 1987). The expectation is that more advanced learners will tend toward more frequent use of conventional target language means for marking such interconnections explicitly, whereas earlier learner varieties will rely more on the operation of co-operative implicatures and discourse rules such as those discussed for temporality by von Stutterheim (1986) and in Volume II:I.3, and for such relational predicates as 'cause', 'justification', and 'solutionhood', discussed by Mann and Thompson (1986).

In addition to the measures of linguistic repertoire, the study included several 'nonlinguistic' measures:

- silent pausing
- vocalised pausing
- self-editing
- unelaborated 'yes'/'no' responses

which were included as indirect measures of the amount of effort invested in task completion by the two informant groups. The amount of effort invested on a task can differ between informants for either of two reasons: (a) if the task is easier for one of them than the other (that is, repertoire limitations), or (b) if one of them is trying harder to perform well than the other (that is, motivational pressure).

Considering motivation first, the Longitudinals would generally be expected to try harder than the Controls since their level of personal commitment to the project was considerably greater - they had, unlike the Controls, committed themselves in advance to a two and a half-year series of encounters. So far as level of difficulty is concerned, in the absence of a directly 'pedagogical' facilitation from project participation, level of difficulty in performing the tasks should start at the same level for the two groups and decrease commensurately over time.

Joint interpretation of these non-linguistic measures with the linguistic measures noted above, allows a determination of whether the production of the two groups differ. Four outcomes are possible:

(i) there are no significant differences between the groups;
(ii) there are significant differences for both sets of measures;
(iii) there are significant differences for the linguistic measures only;
(iv) there are significant differences for the motivational measures only.

The preferred pattern of results would be an observed difference between the two groups on the non-linguistic measures, with Longitudinals scoring higher than Controls, but no difference between the groups on (equally reliable) linguistic variables. This would indicate that the groups differed, but only with respect to their motivation to perform on the tasks, and not in terms of their actual linguistic repertoires.

(iii), and hence (ii) are less desirable outcomes, since the presence of a factor 'participation in the ESF project' (whether facilitative, with the Longitudinals scoring higher than the Controls, or detrimental, vice versa!) intervening in the linguistic development of the Longitudinals would make it more difficult to interpret other, shared determining factors. Outcome (i) would cast some doubt on the validity of the measures selected, as it seems unlikely that project participation had no effect whatsoever on the Longitudinals' TL performances.

The data

The Control group consisted of four informants each from the following six SL-TL pairs: Moroccan-French, Spanish-French, Moroccan-Dutch, Turkish-Dutch, Finnish-Swedish, Spanish-Swedish, selected for reasons of similarity to the Longitudinals on several socio-biographical dimensions: sex, age, source country area, source country

schooling, marital status, length of stay in the target country at the time of the first interview, target language proficiency at that time, amount of contact with target language speakers. Details are given in Appendix B.

Data were gathered from Control informants roughly once every cycle, on a sub-set of the activities used with Longitudinal informants. This sub-set included conversation in all cases. In addition, for target languages French and Swedish, the sub-set included a second task: picture description/comparison. Picture description and conversation may be characterised as opposing poles on a continuum ranging from less open-ended, stimulus-oriented interaction to more open-ended, socially-relevant interaction. These differences will be seen to be important for the analyses below.

The study was limited by the amount and comparability of the data available on the deadline for the start of analysis. The Swedish team had supplied conversation and picture descriptions for eight longitudinal/control pairings from the initial study (Time 1) and from the end of the longitudinal study (Time 3); the Dutch team had comparable conversations from Time 1, Time 3 and from a point mid-way through the longitudinal study (Time 2), for eight longitudinal/control pairings; the Aix-en-Provence team had available conversations from Time 1 and Time 3, for two longitudinal/control pairings. Hence, it was impossible to achieve a fully balanced design involving all informants simultaneously. In place of this, several sub-analyses were performed. To minimise possible artifacts due to individual differences, the informants included in a particular comparison are only those from whom data were available for both of the conditions being compared (Times and Tasks; see below).

This methodological (design) necessity of strictly repeated measures plus the need to include as many data points as possible (for increased power in the statistical tests) from as many different TL-SL pairs as possible (for greater external validity of results) dictated the selection of the following comparisons:

- Both tasks at two points in time: eight pairs from Göteborg; total of sixty-four observations;
- Conversation at two points in time: eight pairs each from Tilburg and Göteborg, two pairs from Aix; total of thirty-six observations;
- Conversation at three points in time: eight pairs from Tilburg; total of forty-eight observations.

Data preparation

In order to maximise comparability of data across tasks, groups, time periods, and SL-TL pairs, a specialised computer programme was used for selecting a sub-set of 200 words from the available data for each task for each informant. The programme scanned the text for learner utterances, and for each utterance, gave half a page of context above and below, while querying the user as to whether that utterance should or should not be included in the sample to be analysed. Excluded from the sample were:

- opening or closing greeting sequences;
- segments of the text which were concerned with task instructions;
- utterances spoken totally in the source language;
- utterances which were totally unintelligible or uninterpretable;
- uninterpretable portions within otherwise fully interpretable utterances.

Everything else was included in the sample. In the event that the number of informant words meeting these conditions fell short of the criterion of 200 words, segments of informant utterances from equivalent other encounters were included in the sample until either the criterion was reached or the supply of appropriate data was exhausted.

A concordance listing (exhaustive listing of items in a file, with one-sentence context for each), and a type-token frequency listing (exhaustive listing of different items in a file and their frequencies of occurrence) were generated separately for each sample file. A modified lemmatisation programme (see 8.2) used the type-token frequency listing as input, and, for each word type, prompted the user for responses concerning:

- language identification of the word type (TL, SL, mixed, or other);
- syntactic category of the word type (noun, conjunction, etc.);
- verification of the frequency (in case of polysemy).

In addition to word types, the programme presented frequences for silent pauses, vocalised pauses, repeated words or sentence fragments, and unelaborated 'yes'/'no' responses to be verified and categorised during the course of lemmatisation.

Results

As noted above, it was not possible to achieve the criterion of 200 language tokens per sample for all the comparison conditions (that is,

conversation and picture description at one to three points in time). Since total sample length in some sense places an upper limit on the possible values of the other measures, it was incorporated into the design as an additional measure, computed as the total number of target and source language tokens, excluding expressives and formulaic utterances.

Given that the potential for silent pauses, vocalised pauses, self-corrections, and unelaborated 'yes'/'no' responses depends in a roughly linear way on the length of the encounter, the raw scores for these variables were adjusted simply by multiplying the raw values by a constant equal to the ratio of 200 to the total lemmatised tokens in a given language sample (excluding expressives and formulaic utterances).

The function relating total sample length to total number of word types is not a linear function and is furthermore known to vary as a function of other variables, such as register, content, and conversational style. For this reason, no clearly sound adjustment of score could be made for the number of target language types, or the number of target language conjunctions, so they were used in their raw forms in the analyses.

Whether total sample length should be grouped with the linguistic or with the motivational variables, is not decidable on the basis of theoretical considerations alone. In fact, this variable tended to correlate more strongly with the study's linguistic variables (.76 with total target language types, .55 with total target language conjunction tokens) than with the motivation or effort variables (—.51 with adjusted vocalised pauses, —.29 with adjusted silent pauses, —.25 with adjusted self-repetitions, —.48 with adjusted unelaborated 'yes'/'no' responses). This is all the more surprising given that total sample length was part of the ratio used in adjusting the raw non-linguistic values, and could therefore have been expected to be actually more closely related (via redundant variance) to the non-linguistic variables. It indeed seems (see 8.2) that increase in text length ties in with progress in acquisition. The correlation between total sample length and total number of source language tokens was only —.11, but this is probably due to the small number of these in the samples, and therefore a highly restricted variance.

The most critical analytic task of these analyses was to determine whether the Control informants could be differentiated from Longitudinal informants on the basis of any multivariate function of any combination of linguistic and motivation variables. For this purpose, the most appropriate type of statistical analysis is discriminant anal-

ysis. Three separate discriminant analyses were performed, one for each of the three comparisons noted in the preceding section.

Discriminant analysis proceeds in three steps:

(a) testing 'whether or not groups of previously classified cases differ significantly on one or more linear combinations' of the available variables (McLaughlin 1980: 176);
(b) determining which of the variables are the best discriminating variables; and
(c) as a sort of double check, using in turn the sub-set suggested by (b) to classify individuals into groups, in order to determine the percentage of cases classified correctly.

While the main discrimination to be made in this study was that between Controls and Longitudinals, two additional discriminations were: (a) between successive time points during the acquisition process and (b) between the two task activities. Given the carry-over of task performance strategies, response biases, and other factors, it is usually easier to discriminate between scores from different individuals under equivalent circumstances than between scores from the same individual under different circumstances. For this reason, the finding of discriminable differences due to time or task (that is, intra-individual differences) in the absence of discriminable differences due to group membership (that is, inter-individual differences) would constitute especially strong evidence against the presence of large differences between Controls and Longitudinals along the constructs being measured. Prior to interpreting the absence of a significant discriminant, however, two additional conditions must be met: (a) the groups for which discrimination failed must be known to have comparable variances on the measures (as may be determined with Box's M test), and (b) the measures employed must be precise enough that the researcher can be confident that the null result is not due simply to error variation.

One final constraint on the analysis concerns the total number of variables used in the discrimination. There is a limit to the allowable number of variables, and this limit is determined by the number of observations. Where the groups are known to be very different, a ratio of ten observations per variable is recommended; otherwise, a ratio of twenty to one (McLaughlin 1980: 188). The danger in entering too many variables is that the discriminant functions become overly affected by chance variation and the classifications are then noticeably less accurate. For the present analysis, in order to give the 'null hypothesis' (of no difference between Control and Longitudinals)

the fairest chance of being rejected, it was decided to include only the most precise measures of each of the dimensions of interest - the linguistic and the motivational - and as few measures of each as possible.

To determine the most reliable sub-set of measures of each dimension, the two groups of measures, linguistic and motivational, were subjected separately to reliability analyses, using first the full data set (Dutch, Swedish, and the French from Aix) and then, in separate analyses, the Dutch and Swedish data. From these analyses, it became apparent that the following measures were much less reliable than the others, and they were therefore excluded from further analyses: expressives, formulas, and source language tokens, self-repetitions, and self-standing 'yes'/'no' responses. Their unreliability is partly due to the relatively low number of occurrences of each of these types of phenomena in the lemmatised samples.

In contrast to these, the following measures were found to be highly reliable, and were included in the discriminant analyses: total number of target language types, total number of target language conjunctions, and total sample length, as the linguistic measures (with Cronbach standardised item alpha coefficients of between .71 and .88 for the three data sets); the adjusted silent pauses and adjusted vocalised pauses as the effort or motivational variables (with Cronbach standardised item alphas of between .77 and .83 for the three data sets). It is important to note that the reliabilities of these measures were equally high for all three data sets and for both measured dimensions, since this suggests that the discriminant analyses based on them had roughly the same potential for precision.

The individual discriminant analyses will be summarised separately for each of the three comparison modes. The first analysis below is presented in greater detail than the others. All other analyses proceeded in the same way.

Both tasks at two points in time: The Swedish data alone were used for this first analysis. These data consisted of four observations each from sixteen informants (eight Controls, eight Longitudinals), for a total of sixty-four observations. These data enabled a test in microcosm (that is, with a small data set) of all three of the types of discriminations to be made here (that is, time, task, and group), and, indirectly, the relative independence of these variables.

1) TIME. The following sub-set of measures was entered 'directly' that is, simultaneously): group, task, sample length, TL types, TL

conjunction tokens, adjusted silent pauses, and adjusted vocalised pauses. The resultant discriminant function was statistically significant, so these same variables were next entered on a step-wise basis. Only three of the entered variables were used in the resulting function: TL types, conjunctions, and adjusted silent pauses, of which the third variable loaded negatively on the factor. This three-variable function was found to have an accuracy rate of 81 per cent in classifying the observations into time groups. This analysis shows a statistically significant difference between scores during the first and third cycles of data collection, with the best discriminators being the main linguistic variables, followed by one of the motivation or effort measures (adjusted silent pauses). The polarity of the factor loadings (and the group means) indicate an increase over time in the number of TL word types and occurrences of conjunctions, and a decrease in the amount of silent pauses.

(2) TASK. The variables entered in this analysis were the same as in (1) except that 'task' was substituted for 'time'. With direct entry, the function was found to be statistically significant, so step-wise entry was used in order to determine the relative importance of the various measures in making the discrimination. Five of the seven variables (that is, all except 'time' and 'group') were entered into the function. The variables were ordered as follows: TL types, TL conjunctions tokens, sample size, adjusted silent pauses, and adjusted vocalised pauses. An examination of factor loadings and sample means indicates that there were more TL types and more vocalised pauses, but fewer silent pauses and conjunctions for conversation than for the picture description task. Of these variables, TL types was clearly the best discriminator. The accuracy rate in classification based on this five-variable function was 78 per cent. These findings suggest that the two tasks differ with respect to both the nature of the linguistic demands (with picture description being associated with a wider range of TL types and more conjunctions than conversation) and the overall level of effort required to perform the task (with picture description being associated with more silent pauses and vocalised pauses than conversation).

(3) GROUP. The variables entered were the same as in (1) except that 'group' was substituted for 'time'. With direct entry, the function discriminating Controls from Longitudinals was found to be statistically significant. From the step-wise analysis the most important variables discriminating the groups were found to be: adjusted silent

pauses, adjusted vocalised pauses, and total sample length, and then, followed after a noticeable drop in the size of the factor loadings, the linguistic variable: TL types. The accuracy of the classification function composed of loadings on these five factors was 73 per cent. This result indicates that the two groups are best discriminated on the basis of the effort or motivational variables, and only to a much lesser degree on the basis of linguistic variables. This finding suggests a motivational rather than repertoire difference between the two groups.

Conversation at two points in time: The data for this analysis came from three of the teams: eight pairs each from Tilburg and Göteborg, two pairs from Aix, for a total of thirty-six observations.

(4) TIME. The same variables were entered as in (1), except for the elimination of the 'task' variable, since only one task is represented in these data. The discriminant function resulting from direct entry, was statistically significant. With step-wise entry, the variables found to contribute most to the discriminant function were, in order: TL types, TL conjunction tokens, and adjusted silent pauses. This function had a classification accuracy of 77 per cent. These results are conceptually identical to those of the Time analysis reported in (1), even though the data base used here only partially overlaps with the other data base, including fewer informants and only one task. Both this and the preceding time analysis indicate a strong developmental influence on both linguistic variables, combined with an influence on one of the effort variables, in a direction suggesting less effort at the later point in time. These results also provide good evidence for the validity of the so-called 'linguistic' variables as measures of repertoire characteristics.

(5) GROUP. The same variables were entered as in (4), except for substituting 'time' for 'group'. The discriminant function resulting from direct entry of the variables was statistically non-significant. The fact that Box's M test for non-equivalence of variance was also non-significant, is evidence that the failure of the discriminant function to reach significance was not due to low power in the test. Though it is not technically necessary to examine the results of a step-wise procedure when the direct procedure yields non-significant functions, when this was examined in the present case, it was found that the discriminant function generated by the step-wise procedure consisted of only two variables: adjusted silent pauses, and sample length, that

is, one effort measure and the ambivalent measure of sample length. Together these two variables had a classification success rate of only 62 per cent, which is the lowest rate of any of the other analyses for the data set, and not so far from a chance rate of 50 per cent accuracy. From these results it seems reasonable to return to the null result of the 'direct' method, and to conclude that there was in fact no evidence of a discriminant function capable of separating the two groups in this analysis, and that this lack of significance is not due to low power but rather to high similarity of the two groups.

Conversation at three points in time: The data for this analysis came from the eight pairs from Tilburg, for a total of forty-eight observations.

(6) TIME. The same variables were used as were entered in (4). The discriminant function resulting from direct entry of the variables was statistically significant. The discriminant function resulting from step-wise entry contained only two variables: TL types, and sample length, which together had a classification accuracy rate of 58 per cent. While these results differ somewhat from those of the other two Time analyses, this difference may be partly attributable to the much reduced number of observations in this data set (that is, forty-eight in comparison with sixty-four). One is reminded of McLaughlin's warning to include at most only one variable for each ten observations. While these results may be based to a larger degree on the effects of random variation, it is nevertheless interesting to note that even under these conditions (that is, of lower power due to fewer total observations), the highest loading discriminant variable here, as in the two prior analyses, was the number of TL types.

(7) GROUP. The same variables were used as were entered in (5). The discriminant function resulting from direct entry of the variables was non-significant, but the presence of a significant result for Box's M test of non-equivalence of variance across the groups, raises doubts concerning the potential power of the test in this particular case. For this reason, the step-wise method was used next. The resulting function, which was significant, contained only three variables, in this order: adjusted silent pauses, adjusted vocalised pauses, and time. The classification accuracy rate of this function was 68 per cent, which is quite good considering the very small number of observations. This result replicates the finding from the Group analysis in (3), that effort or motivational variables serve as the best em-

pirical basis for discriminating between Controls and Longitudinals. The large contribution also of Time in the discriminant function suggested a statistical interaction between Time and the Group variable, as might arise, for example, if the groups differed substantially with respect to the rate at which their value on the effort variables declined over time. In fact, an examination of the means suggested a definite trend of this type in the data. For both measures (that is, silent pauses and vocalised pauses) and for both groups there is a large decrease in values from Time 1 to Time 2, having roughly the same slope and partly overlapping. But the actual values at Time 2 are higher for Longitudinals on both measures than they are for the Controls, and this is true also at Time 3. It would be possible to interpret this interaction as revealing both the effort aspects (Time 1 to Time 2) and the motivation aspects (Time 2 to Time 3) of these measures. That is, from Time 1 to Time 2 the curves for the two groups overlap because they are experiencing similar levels of difficulty with the task. From Time 2 to Time 3, the task is no longer so difficult, but the Longitudinals, being more highly motivated, are trying harder.

Conclusions

Concerning the discrimination of major importance to the generalisability of results from other parts of the ESF project, that is, the discriminations between Control and Longitudinal informants, the three relevant analyses were encouraging. Where the discriminant functions were statistically significant, the variables included in them were effort or motivation variables rather than the 'pure' linguistic measures, TL types or TL conjunction tokens. This suggests that the differences between the groups, are more due to effort or familiarity than to actual differences in repertoire.

Concerning the discriminations based on Time, these analyses provide evidence of a systematic effect on the linguistic variables, together with a milder effect on the effort or motivation variables. This finding serves the joint purpose of (a) showing TL types and TL conjunction tokens to be valid measures of learner repertoire, and (b) actually providing statistical evidence of an acquisition effect.

Concerning the discriminations based on Task, these analyses provide evidence that the demands of the tasks were in fact quite different, which is also important methodologically, in that it ensures that the Control versus Longitudinal comparisons were not non-significant due to a restriction in variance due to stimulus condition.

There always remain two questions in work such as this: (a) were those aspects of learner repertoire included here really the most important ones with reference to questions of generalisability, and (b) were those aspects which were included here measured to an adequately sensitive degree? As was noted above, the findings from the Time discriminations provide some evidence that the measures used were valid measure of some aspects of repertoire, but there are obviously others which could be measured in addition. As regards the question of measurement sensitivity, the number of TL types is certainly a very global measure; the actual degree of overlap of individual types in the repertoires, at least for those categories seen as least sensitive to conversation topic and most enlightening concerning syntactic complexity (that is, conjunctions and adverbs), could also be assessed in future work of this kind.

In conclusion, the relevance of the present results for the generalisability of the ESF findings as a whole can be tersely summarised as follows: the effect of participation in the project and intensive interaction with project researchers is small, and where it does exist, it is of a clearly effort-related or motivational nature and does not show any substantial influence on the structure of the acquisition process.