Chapter V

The experiment

5.1 Introduction

In the previous chapter, a number of studies on the manipulation of the +/- Planning Time and +/- Here-and-Now variables were reviewed and the questions that this experiment aims at answering were advanced.

This chapter will describe the experiment which was designed to answer those questions by providing information about the participants in the study; elicitation tasks and procedures; the measures of fluency, structural and lexical complexity, and accuracy; as well as information about the transcription and coding of the narratives elicited from participants in the study. Thus, this chapter will provide an answer to the following questions:

i) What was the design of the experiment?

ii) Who were the participants in the experiment?

iii) What kind of tasks was used to elicit spoken data?

iv) What measures were used to calculate the fluency, complexity, and accuracy of learners' production?

v) What statistical instruments were employed to calculate the results?

5.2 Experimental design

A repeated-measures design was used in which the within-subjects factor was Task Complexity. As far as the independent variable is concerned, four levels of Task Complexity were analyzed:

Condition 1: Planned Here-and-Now

Condition 2: Unplanned Here-and-Now

Condition 3: Planned There-and-Then

Condition 4: Unplanned There-and-Then

Repeated-measures analyses of variance (ANOVA) of the 10 dependent variables were carried out. Theses include: Unpruned Speech Rate A and Pruned Speech Rate B (fluency), Percentage of Lexical Words, Ratio of Lexical to Function Words, Guiraud's Index of Lexical Richness (lexical complexity) and S-Nodes per T-units (structural complexity), and Percentage of Error-free T-units, TLU of articles, Percentage of Self-repairs, and Ratio of Repaired to Unrepaired Errors (accuracy).

Given that in this experiment learners were asked to narrate four stories, it was thought that practice or carryover effects might take place from one story to another. Practice effects are thought to make subjects improve during testing or perform poorly because of boredom or fatigue. The design used for data collection assumed that stories were similar to one another and that what made a difference in performance was the condition under which each story was performed. In other

words, if a specific condition were to have an effect on performance, it should have happened regardless of the story type. Besides story type, the sequence of conditions under which the tasks were performed was also thought to potentially affect performance. In order to counterbalance such effects, the following measures were taken:

- Subjects were only given 2 stories in each session; stories 1 and 2 in session 1, and stories 3 and 4 in session 2. Sessions were 2 days apart.
- All students narrated the four stories in the same order, but a Latin square design[1] was used to counterbalance the effects of sequence of condition presentation.

	Story 1	**Story 2**	**Story 3**	**Story 4**
Group A	+ planning time There-and-Then	- planning time There-and-Then	+ planning time Here-and-Now	- planning time Here-and-Now
Group B	- planning time There-and-Then	+ planning time Here-and-Now	- planning time Here-and-Now	+ planning time There-and-Then
Group C	+ planning time Here-and-Now	- planning time Here-and-Now	+ planning time There-and-Then	- planning time There-and-Then
Group D	- planning time Here-and-Now	+ planning time There-and-Then	- planning time There-and-Then	+ planning time Here-and-Now

Figure 15. Latin square design.

[1] In a Latin square design, treatments are assigned at random within rows and columns, with each treatment once per row and once per column, in order to control for variation in two directions. In this case the two directions are condition and sequence (Steel & Torrie, 1980).

Even though such measures may reduce potential carryover effects, it was thought important to test each dependent measure for story type by means of a repeated measures ANOVA. The results and level of significance of comparing stories are presented in Table 2 in Chapter 6, Section 6.2.

5.3 Participants

48 volunteers participated in the study among lower-intermediate, first- and second-year university students who had been placed in the same level of English class by an internal placement test of the Blanquerna Communication Studies program at Universitat Ramon Llull, in Barcelona, Spain. However, since the reliability of such a test has never been statistically tested, homogeneity regarding proficiency levels was also controlled for by means of a C-test[2], as shown by Table 18 (on facing page). Students in the Blanquerna Communication Studies program have usually received instruction in English for approximately the same number of years. Learners' ages range between 18 and 22. Appendix A provides detailed information about the students' codes[3], gender, ages, and years of instruction.

[2] C-tests have been described as having high reliability and validity indices and have often been adopted as instruments for testing learner proficiency. See Klein-Braley, 1997, Jafarpur, A. (1999), and Babaii, E. and Ansary, H. (2001). See Appendix J.
[3] Real names were not included to preserve anonymity.

Table 18

C-test scores: Means, standard deviations, and level of significance of comparison among groups of learners

	Group A		Group B		Group C		Group D		*F*	*p-level*
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*		
C-test n = 48	71.00	8.33	69.09	5.73	68.00	9.19	71.67	8.41	.519	n.s

5.4 Piloting, elicitation tasks and procedures

Piloting of the comic strips was carried out with a small number of students before data collection in order to achieve comparability of the stories and to make sure no significant differences existed regarding the perceived difficulty of plots, their interest, or their length. Eight comic strips were originally selected and, immediately after narrating them, learners were interviewed on the clarity of the comic strip, as well as on their perception of difficulty in terms of narrating the plot and the difficulty of the vocabulary. Once the difficulties of the eight wordless comic strips were assessed, only four stories which were considered to be similar were selected and used for data collection for each group.

The strips used in this research were created by Argentine artist Quino (Salvador, 2001). The stories (See Appendix L) were thought to be especially useful for data collection because they were all wordless comic strips, they all contained a

small number of characters[4] who were involved in the action, and they had a clear climax and resolution. All plots worked in a similar way: in the first vignettes a number of expectations were generated which were reversed towards the end of the story, with the aim of achieving a humorous effect. Finally, despite the fact that all the social events represented by the comic strips were exaggerated and distorted by Quino's sense of humor, they were all thought to refer to situations all students could be familiar with within their cultural parameters. This idea was confirmed by the interviews in the pilot study.

Data collection took place in two different 30-minute sessions in the researcher's office with each learner, and it was conducted by the researcher himself. Learners sat facing the researcher and, after a few minutes of small talk to achieve rapport and collect personal information, they were asked to narrate the stories under the different conditions[5]. The researcher provided learners with one or two words (e.g. 'comb' or 'checkout girl') which had been found to be problematic during piloting. The steps taken during the two sessions are described below:

Session 1:	Brief explanation of research and small talk
	Personal information retrieval
	Story 1
	Story 2
Session 2:	Small talk for rapport
	Story 3
	Story 4
	Affective variables questionnaire
	Protocol analysis

[4] Story 1 contained 1 man, 1 woman, and 1 secondary character; Story 2 contained 2 men; Story 3 contained 1 man, 1 woman, and a secondary character; Story 4 contained 1 man and 1 woman.
[5] See Appendix B for a description of the specific instructions given to subjects.

Following several studies (Foster & Skehan, 1996; Skehan & Foster, 1997; Mehnert, 1998; Ortega, 1999), operationalization of planning time was 10 minutes for planned narratives and 50 seconds for unplanned ones (enough to understand the story). When planning time was available, subjects were encouraged to take notes on what to say and how to say it as they planned, but were told they would not be allowed to keep their notes during task performance.

Regarding the Here-and-Now/There-and-Then distinction, this research followed Robinson's (1995a) operationalization. For Here-and-Now, learners were asked to narrate the story in the present while they looked at the strips. For There-and-Then, learners were asked to narrate the story in the past tense, and they were not allowed to look at the pictures as they performed the task.

5.5 Measures

Fluency:	Rate A (syllables per minute in unpruned speech)
	Rate B (syllables per minute in pruned speech)
StructuralComplexity:	S-Nodes per T-unit
Lexical Complexity:	Percentage of Lexical Words
	Ratio of Lexical Words to Function Words
	Guiraud's Index of Lexical Richness
Accuracy:	Error-free T-units
	TLU of Articles
	Percentage of Self-repairs
	Repaired to Unrepaired Errors

The basic criterion when selecting measures for the three dimensions of production in this study was their degree of comparability to both planning and Here-and-Now studies. In so far as possible, measures that were used on both types of studies were chosen.

5.5.1 Fluency Measures

There has been a wide variety of approaches to measuring fluency. Planning studies have reported significant differences in favor of planning when calculating the number of replacements, repetitions, and hesitations (Foster & Skehan, 1996); the number of pauses and total silence (Foster & Skehan, 1996; Mehnert, 1998); and unpruned and pruned speech rates (Mehnert, 1998; Ortega, 1999; Yuan & Ellis, 2003). Studies that have manipulated task complexity along the Here-and-Now variable have used measures, albeit with no significant differences, such as the total number of 2-second pauses (Robinson, 1995a) and the number of words per pausal unit (Robinson, 1995a; Rahimpour, 1997). In this study, the rate of both unpruned and pruned speech was chosen to code and measure each narrative[6]. The main advantage of this kind of measure is that it in fact includes both the amount of speech and the length of pauses, since it takes into account the number of syllables and the total number of seconds in the narrative (Griffiths, 1991). The difference between Speech Rate A and Speech Rate B is that in the latter, repetitions,

[6] The researcher's original intention was to calculate the number of 2-second pauses, but during transcription it was discovered that a number of narratives contained no or very few 2-second pauses, which distorted the calculations considerably.

reformulations, false starts, and asides in the L1 are eliminated from the calculation (Lennon, 1990). The formula used for the calculation of Rate A and Rate B was:

$$\frac{\text{Number of syllables}}{\text{Total number of seconds}} \times 60$$

5.5.2 Lexical and Structural Complexity Measures

Two basic measures were used in both planning and Here-and-Now studies: the type/token ratio (Ortega, 1999) and the percentage of lexical words (Robinson, 1995a; Rahimpur, 1997). In the present study the percentage of lexical words is calculated, but not the type/token ratio. The type/token ratio has been shown to be extremely sensitive to differences in text length, since the higher the number of tokens, the lower the ratio (Vermeer's, 2000). Several alternatives[7] to the type/token ratio have been advanced by a number of researchers in order to correct the negative correlation existing between type/token results and the number of tokens. After carefully considering Vermeer's (2000) analyses of the different variations of the type/token measure, this study uses the Guiraud's Index of Lexical Richness. The advantage of such a measure is that by including the square root of the tokens it compensates for differences in text length. In addition to the percentage of lexical words, the ratio of lexical to function words is also used. The calculation uses the same raw data needed for the calculation of the percentage of lexical words, but it

[7] For example, following Richards and Malvern (2002), Yuan and Ellis (2003) use the Mean Segmental Type-Token Ratio (MSTTR), which was described in Section 4.2.1 (footnote No. 6). See Vermeer (2000) for a thorough empirical analysis of an array of lexical complexity measures.

establishes a ratio between the two categories of words, that is, lexical and function words. The following are the formulas used for each measure:

Percentage of lexical words:

$$\frac{\text{Number of lexical words}}{\text{Total number of words}} \times 100$$

Ratio of lexical to function words:

$$\frac{\text{Number of lexical words}}{\text{Number of function words}} \times 100$$

Guiraud's Index:

$$\frac{\text{Types}}{\sqrt{\text{Tokens}}}$$

Three basic units of analysis have been used by production researchers for measuring syntactic complexity in oral production: the C-unit, the T-unit, and the utterance[8]. In planning studies, researchers have measured the number of clauses per C-unit (Foster & Skehan, 1996: Skehan & Foster, 1997: Mehnert, 1998); the number of words per C-unit (Mehnert, 1998); the number of subordinate clauses per T-unit (Mehnert, 1998); the number of S-Nodes per T-unit (Mehnert, 1998); the number of words per utterance (Ortega, 1999). Studies of the Here-and-Now variable have measured the number of multipropositional utterances (Robinson,

[8] Other alternatives have been suggested more recently. Foster, P., Tonkyn, A. and Wigglesworth, G. (2000) suggest the AS-Unit (the Analysis of Speech Unit) which corrects some of the shortcomings of T-Units, C-Units, and utterances. Their proposal, however, has not met with widespread acceptance so far.

1995a), and the number of sentence nodes per T-unit (Robinson, 1995a; Rahimpour, 1997). In this study the T-unit is preferred over the C-unit, since it deals with one-way, monologic narratives which are expected to trigger no elliptical answers. In this study, then, syntactic complexity is measured by counting the number of S-Nodes (a term which is interchangeable with 'clause') and dividing it by the total number of T-units. The formula used to calculate syntactic complexity is:

$$\frac{\text{Number of S-Nodes}}{\text{Number of T-Units}}$$

5.5.3 Accuracy Measures

Finally, in both planning studies and Here-and-Now studies, accuracy has been mainly measured by calculating the percentage of error-free units and target-like use of articles; error-free clauses (Foster & Skehan, 1996; Skehan & Foster, 1997; Mehnert, 1998); error-free T-units (Robinson, 1995; Rahimpour, 1997; Ortega, 1999); target-like use of articles (Robinson, 1995a; Rahimpour, 1997). Two measurements, target-like use of articles and the percentage of error-free units, are used in this study. The following are the formulas for the calculation of these two measurements:

Percentage of error-free T-units:

$$\frac{\text{Number of error-free T-units}}{\text{Number of T-units}} \times 100$$

TLU of articles:

$$\frac{\text{No. of accurately supplied articles}}{\text{No. of obligatory contexts + No. of inappropriate suppliance}} \times 100$$

In addition to these two well-established measurements, this research operationalizes two more measurements: the percentage of self-repairs and the ratio of repaired to unrepaired errors. There are a number of arguments that can be advanced in order to justify their use: firstly, self-repairs, whether other-initiated or self-initiated (Schegloff et al., 1977), denote students' awareness of form and can be interpreted as learners' attempts at being accurate (Kormos, 1999). Lyster and Ranta (1997:57), for example, suggest repairs generated by learners as a result of corrective feedback lead them both to automatize the retrieval of target language knowledge they already have and to revise their hypotheses about the target language. Swain (1998:66) has also hypothesized that noticing a hole in their own interlanguage may lead learners to notice the gap by directing their attention to relevant input. All these functions of self-repairs have been said to potentially lead to acquisition, and they have been pointed out in order to defend the benefits of certain types of corrective feedback. Self-initiated repairs, on the other hand, serve the same

purposes as other-initiated repairs only that they are not the result of corrective feedback but rather are spontaneously generated by learners themselves or, in Levelt's terms (1989), they are the result of the speakers' monitoring of their own speech. Secondly, while the percentage of error-free units and the TLU of articles provide information about the accuracy of the 'finished' product of learner's performance, the two new measures present accuracy 'in process' as learners try to correct and improve their own speech. Thirdly, at least one study has reported a higher proportion of self-repairs under certain planning conditions. Yuan & Ellis's (2003:17), when analyzing the effect of on-line planning on learners' production, have reported a higher frequency of reformulations and self-corrections when they are given sufficient time 'during' (as opposed to 'before') performance. About the subjects in their study, they suggest that on-line planners "engaged more fully in searching their linguistic repertoires and in monitoring their speech production". In this research, the percentage of self-repairs is calculated by taking the number of self-repairs[9] and dividing it by the total number of errors[10] and multiplying the results by 100; the ratio of repaired to unrepaired errors, on the other hand, is calculated by dividing the number of repaired errors by the number of unrepaired errors and multiplying the results by 100.

[9] Because no protocol analysis was used to detect neither different repairs nor appropriateness repairs (Kormos, 1999; See Section 1.4.3.3), self-repairs exclusively refers to error-repairs in this study. Also, phonological self-repairs were not considered either, even if they are relatively easy to detect. The reason is to be found in the difficulty to reach an agreement between raters as to what constitutes a phonological error.

[10] The definition of error is adopted from Lennon (1991:182): "a linguistic form or combination of forms, which in the same context and under similar conditions of production would, in all likelihood, not be produced by the speakers' native speaker counterparts.

The following are the formulas used for each measurement:

Percentage of self-repairs:

$$\frac{\text{No. of self-repairs}}{\text{No. of errors}} \times 100$$

Ratio of repaired to unrepaired errors:

$$\frac{\text{No. of repaired errors}}{\text{No. of unrepaired errors}} \times 100$$

5.6 Statistical instruments and analyses

A Microsoft Excel spreadsheet was used to enter data and to design graphics. All statistical analyses were carried out using statistical package SPSS 11.5.1 for Windows.

Four different kinds of statistical analyses are used in this dissertation: descriptive statistics, which provide information about means, standard deviations, skewness, and kurtosis; repeated-measures analyses of variance (ANOVA) are used for the comparison of stories and conditions; *post hoc* Scheffe's comparisons to identify the exact location of differences; two-tailed T-tests are used for comparisons between pairs of groups; and Spearman correlations are used to check any correlations existing between production measures and affective variables.

5.7 Transcription and coding

The CA mode of CHILDES (MacWhinney, 1995) was used for the transcription of the narratives. This software allows for the automatic calculation of items (e.g. words or tags) in a text.

Both intrarater and interrater measures were used in the transcription and coding of the narratives. The transcription of the narratives was carried out by the researcher and a research assistant. Intrarater reliability reached 97%, and interrater[11] agreement out of a randomly selected sample of 10% percent of the data reached 93.7%. The following are the percentages of interrater reliability reached for the coding of stories for the three dimensions of production:

Fluency:	Rate A = 98%
	Rate B = 98%
Complexity:	Lexical Complexity = 95%
	Structural Complexity = 90%
Accuracy:	Error-free T-units = 90%
	TLU of Articles = 92%
	% Self-repairs = 86%
	Ratio of repaired to unrepaired errors = 84%

[11] Interrater reliability was calculated by means of percentage agreement.