ON THE DEVELOPMENT AND EVALUATION OF A C-TEST FOR FRENCH

Appeared in:

Fremdsprachen und Hochschule (FuH) - AKS Rundbrief 13/14, 1985. Bochum: Clearingstelle des Arbeitskreises der Sprachenzentren, Sprachlehrinstitute und Fremdspracheninstitute (AKS) an der Ruhr-Universität Bochum, 101-120

ON THE DEVELOPMENT AND EVALUATION OF A C-TEST FOR FRENCH*

Rüdiger Grotjahn/Brigitte Stemmer (University of Bochum)

1 Introduction

For a relatively long period many language testers have been quite enthusiastic about the cloze test as an instrument for measuring global language proficiency both in first and second languages. Recently, however, on the basis of strong empirical evidence, serious doubts have been expressed about the reliability and validity of cloze tests (cf. Alderson 1979; Klein-Braley 1981). According to Raatz/Klein-Braley (1982) there are, among others, the following problems in respect to the cloze principle:

- The test may be unfair because of text specificity.
- The factors 'text, deletion rate, and starting point' affect the reliability and validity coefficients.
- Tests with high deletion rates have to be extremely long in order to ensure a sufficient number of items.
- The difficulty of the test is affected by the ratio of the number of structure words and content words which have been deleted.

In light of this and other criticism, Raatz/Klein-Braley (1982) have proposed the C-principle as a modification of the classical cloze procedure. While the C-Test is based on the same theoretical assumptions which underlie the classical cloze test, namely the concept of an internalized pragmatic expectancy grammar as outlined by Oller (e.g. 1976), the format chosen for the C-Test is intended to be above the criticisms directed towards the classical cloze procedure. In order to achieve more fairness with regard to text content, several short texts with different topics are chosen rather than one long text. To minimize problems connected with the choice of deletion rate and starting point and to assure a sufficient number of items, the second half of every second word is deleted, one-letter-words not being taken into account. If an n-letter-word has an uneven number of letters, the number of letters deleted is (n+1)/2. One or two lines of each text are left undistorted as a lead-in.

2 Development of a French C-Test

Since its introduction, the C-Test has become more and more popu-

---

* This is a revised version of a paper presented at the Seventh World Congress of Applied Linguistics, Brussels, August 1984. We would like to thank Andrew D. Cohen, Claus Faerch, Gabriele Kasper, Christine Klein-Braley, Wolfhart Matthäus and Ulrich Raatz, who all contributed to it in some way.

lar as a measure of general language proficiency. Up to now, it has been applied mainly to English and German, although there has already been some research done on such languages as Spanish, Finnish, Turkish, and Hebrew (cf. Klein-Braley/Raatz 1984; Süßmilch 1984). In the present article we want to report on a project we began at the University of Bochum in the Federal Republic of Germany in 1982. The project centres around the development of a C-Test for French intended to be used in combination with or possibly even as an alternative to our relatively uneconomical "Bochum Diagnostic Test for French" (henceforth BDF) designed for students starting French at the university level. As a rule, the students taking the test had one to seven years of high school French instruction and were therefore extremely heterogeneous with regard to their French proficiency. According to their total score on the BDF, the students are placed in at least two different proficiency levels of our obligatory remedial French language courses. Some students with a very high score are exempted from the course. Furthermore, the diagnostic information of the different subtests of the BDF is used to improve instruction on each proficiency level (for more details see Grotjahn 1984).

In developing the C-Test we proceeded as follows: first, we looked for texts which, on the basis of our teaching experience, we thought to be suitable. We then prepared a preliminary C-Test consisting of 7 different texts with a total of 200 items. An example of a French C-Test, together with the written test instructions, is given in Figure 1. The German test instructions have been translated into English and only the first text is given.

Figure 1

Test Instructions:

This test consists of 6 different short texts. In these texts, part of every second word is missing. Please try to fill in the missing part. (Words with a hyphen or an apostrophe as e.g. celui-ci, faut-il, l'un, qu'il are considered as one word.)

Example:
"Il est difficile d'imaginer un monde sans publicité. La publicité est une chose qui, aujourd'hui, fait partie de notre vie quotidienne. Il est difficile de concevoir des journaux sans publicité, des murs sans affiches publicitaires.

Text 1:
"Mes filles aiment beaucoup la publicité télévisée. Elles sont très jeunes et c'est vraiment ce qu'elles préfèrent. Je dirais même que cela les passionne. Je pense que c'est parce qu'elles retrouvent des scènes de la vie quotidienne qui, pour elles, représentent quelque chose. Les petits films publicitaires les intéressent parce qu'elle voit une maman faire sa vaisselle ou pas, son aspirateur, une famille à table, un enfant manger son fromage ou sa crème glacée."

It should be mentioned that our instructions differ somewhat from those used in the English C-Test developed by Klein-Braley and Raatz. In their test instructions the student is told that the second half of every second word is missing. In an English C-Test used at the University of Bochum, the students are even told how the gaps of the C-Test are constructed in the case of words with an uneven number of letters. As we know from some students having taken this test, in some cases they found the correct solution only by counting the number of letters of each word under consideration as a blank filler. Since we feel that the ability to count the number of letters of a word correctly is surely not a component of the language processing competence intended to be measured by the C-Test (although it might be an important component of crossword puzzle solving competence), we only tell the students that part of every second word is missing.

Since the cloze test has been criticized for being often too difficult even for competent native speakers of French, we gave our preliminary C-Test to 16 native speakers of French at the University of Paris. We hoped at the same time to obtain information about items with more than one acceptable solution. Fortunately, the C-Test turned out to have an acceptably low level of difficulty for native speakers of French. For each text, at least 92% of the solutions were exactly correct; a maximum of 7 % were acceptable, and a mean of 98% were both exact and acceptable. It should be mentioned that in this investigation and also in the investigations with non-native speakers reported below, we ignored clear-cut instances of spelling errors when classifying solutions as exact or acceptable. Consequently, 'exact' means that the testee provided the original word, and 'acceptable' that he provided an adequate variant of the original word although there may have been spelling errors in both cases.

If we compare our results with those obtained by Klein-Braley and Raatz for English, their C-Test versions seem to yield a higher percentage of exact solutions than ours. Unfortunately, we do not yet know whether this is due to the languages involved or to text-specific characteristics or even to the differences in the test instructions mentioned above.

Subsequent to this pilot study, different versions of the C-Test, consisting of 5 or 6 of the total of the 7 texts pretested with native speakers, were given to more than 400 University students (mainly from the University of Bochum) and also to some high school students. In these investigations our C-Test turned out to be a very practical and economical testing device, which was at the same time highly reliable. Furthermore, it could be shown that the C-Test scores correlate strongly with the amount of French language instruction, a fact which supports the construct validity of the French C-Test.

3 Statistical Analyses

We would now like to present at least some of our data in more detail. We will restrict ourselves to the C-Test version administered together with the BDF to 115 students of French in 1983 and 1984. The

C-Test consisted of 6 texts, each with 22 items. We used both the exact and the acceptable scoring methods (exact and acceptable defined as above). Since the methods yielded essentially the same results, only the data for the exact scoring method are presented.

Table 1

C-Test 1983/84: Some Test Statistics (N=115)

Statistic	TEXT 1	TEXT 2	TEXT 3	TEXT 4	TEXT 5	TEXT 6	TEXTTOT
n	22	22	22	22	22	22	132
$\bar{p}$	.64	.62	.53	.48	.68	.61	.59
s	.21	.24	.27	.25	.22	.21	.24
$\bar{r}$	.23	.18	.14	.26	.16	.29	.17

n Number of Items
$\bar{p}$ Mean Item Difficulties
s Standard Deviation of Mean Item Difficulties
$\bar{r}$ Mean Inter-Item Correlations

As is shown in Table 1, the C-Test is somewhat too easy for our students. This fact reflects our problems in finding sufficiently difficult texts for more advanced students, something also encountered with English C-Tests.

Reliability coefficients for different versions of the C-Test are given in Table 2.

Table 2

C-Test and BDF 1983/84: Reliabilities (N=115)

Test Version	L_1	L_2	L_3	L_4	L_5	L_6
TEXT1,...,TEXT4, TEXT6	.72	.91	.90	.91	.89	.89
TEXT1, TEXT3, TEXT4, TEXT6	.67	.89	.89	.88	.88	.86
TEXT1, TEXT2, TEXT4, TEXT6	.67	.90	.90	.90	.88	.87
TEXT1, TEXT4, TEXT6	.58	.88	.88	.78	.87	.83
BDF: Item Scores	.93	.94	.93	.85	.93	-
BDF: Subtest Scores	.71	.85	.81	.85	.84	.85

To estimate the reliability, we used the so-called Lambda coefficients L_1 to L_6 derived under different sets of assumptions by Guttman (1945). According to one of these assumptions, the items must be experimentally independent from each other. However, as indicated by our qualitative analysis of the problem-solving behaviour of the students taking the test discussed in Section 4 of this article, this assumption is not met in the case of a C-Test. We therefore considered each text to be a 'superitem' and calculated the Guttman coefficients on

the basis of superitems with the help of the reliability procedure provided by SPSS (cf. Hull 1981).

The Guttman coefficients constitute lower bounds for the population reliability. Applied to samples, they can naturally also overestimate this reliability. Some of the best-known reliability coefficients are special cases of the Guttman coefficients, as for instance the Spearman-Brown split-half coefficient, which is a special case of L_4, or Cronbach's Alpha, which numerically corresponds to L_3. Although L_2 can be regarded to be superior to L_3 (and hence also to Cronbach's Alpha), we will focus on L_3 in discussing the reliability, because L_3 (i.e. Cronbach's Alpha) is used in most reliability studies.

As can be seen from Table 2, the French C-Test is highly reliable. Even if we shorten the test to only three texts with a total of only 66 items, we get L_3=.88, an impressive value for a test which can be administered in about 15 minutes and scored in less than 5 minutes per testee.

Since the BDF has not yet been submitted to a thorough statistical item analysis, previous to the computation of reliability coefficients for the BDF a preliminary item analysis was carried out. The analysis yielded a test version consisting of 8 subtests and 116 items, whereas the original version consisted of 9 subtests and 143 items (for more details see Grotjahn 1984). All analyses presented in this article are based on the revised version.

Since a relatively large portion of the items of the BDF are cloze items and thus not experimentally independent, the reliability coefficients have been computed both on the basis of item and subtest scores. In interpreting the value of L_3 (Cronbach's Alpha) based on the subtest scores, it should be taken into account that this value underestimates the reliability for several reasons. Firstly, the subtests have very different lengths, a bias which can, however, be corrected with the help of a generalization of Cronbach's Alpha, namely the Beta coefficient proposed by Raju (1977). Secondly, as will be shown later, not all subtests of the BDF measure on the same dimension. In this case, L_2 is a better estimate than L_3. We therefore assume that a more realistic estimate of the reliability of the BDF will be somewhere below L_3=.93 but will be higher than L_2=.85. Hence the BDF can be regarded as reasonably reliable.(1)

Up to now, the factorial validity of the C-Test has been investigated in only a few studies (cf. Raatz/Klein-Braley 1983; Raatz 1984a; Grotjahn/Stemmer 1984). Since in Grotjahn/Stemmer (1984) the data base was very small, the results of this study cannot be considered very reliable. Therefore, in order to analyze in more detail to what extent the C-Test and the BDF provide similar information, we carried out several factor analyses, all with communalities estimated by iteration. First, to get an impression of the dimensionality of the C-Test, we calculated the intercorrelations between the 6 C-Test texts and also the part-whole-corrected correlations between the total C-Test score and the scores on the 6 texts. The data are presented in Table 3.

Table 3

C-Test 1983/84: Pearson Correlation Coefficients (N=115)

	TEXT 2	TEXT 3	TEXT 4	TEXT 5	TEXT 6	TEXTTOT
TEXT 1	.65	.65	.70	.66	.71	.80
TEXT 2		.56	.68	.68	.69	.77
TEXT 3			.61	.65	.65	.72
TEXT 4				.57	.71	.78
TEXT 5					.64	.75
TEXT 6						.81

Since the correlations are uniformly high, the texts appear to measure on the same dimension. This impression has been confirmed by the factor analysis which we carried out, excluding Text 5 as far too easy for our students. The analysis yielded only one factor which accounted for 73% of the variance. The loadings of the texts on this factor ranged from .74 to .85. Hence, according to this analysis, our C-Test has to be considered as undimensional.

Table 4 presents the intercorrelations between the subtests of the BDF and the C-Test (Text 5 excluded).

Table 4

BDF and C-Test 1983/84: Pearson Correlation Coefficients (N=115)

	2	3	4	5	6	7	8	9	C-TEST
1 PRONOUNS	.33	.08	.47	.55	.62	.59	.46	.69	.63
2 LISTENING		.04	.10	.07	.17	.24	.20	.24	.29
3 QUESTIONS			-.15	.01	-.08	.06	.11	-.01	.00
4 TENSES				.55	.56	.52	.38	.59	.48
5 CONJUNCTIONS					.63	.53	.52	.70	.57
6 SUBJUNCTIVE						.55	.48	.71	.62
7 VERBS							.61	.71	.63
8 ADJECTIVES								.60	.58
9 TOTAL									.74

As can be seen from this Table, the correlation between the C-Test and the BDF is .74. Hence the C-Test has a high convergent validity with regard to the BDF. This result is in line with other studies, where high correlations between C-Tests and school grades, teacher ratings of pupils' language competence, and other language tests have been reported (cf. the review in Klein-Braley/Raatz 1984).

If we now examine the other correlations in Table 4, it is immediately obvious that the subtests 'Listening' and 'Questions' behave quite exceptionally: all correlations with these two subtests are relatively low and in some cases even negative. Thus at least these two subtests seem to measure on a different dimension.

The results of a factor analysis of the data in Table 4 are reported in Table 5.

Table 5

C-Test and BDF 1983/84: Factor Matrices (N=115)

	Loadings				
	Varimax Rotation		Oblique Rotation		
Variable	F1	F2	F1	F2	h^2
PRONOUNS	.73	.25	.75	.13	.60
LISTENING	.23	.30	.25	.26	.14
QUESTIONS	-.05	.35	-.03	.36	.12
TENSES	.72	-.24	.70	-.35	.57
CONJUNCTIONS	.75	-.03	.75	-.15	.65
SUBJUNCTIVE	.80	-.05	.80	-.18	.64
VERBS	.75	.26	.77	.13	.62
ADJECTIVES	.64	.30	.66	.19	.49
C-TEST	.77	.24	.79	.11	.66
Initial Eigenvalues:	4.34	1.15			
Variance Explained:	62%				

Factor Pattern Correlations:

Factors	F1	F2
F1	1.00	.10
F2	.10	1.00

In this factor analysis we used both varimax and oblique rotation. The analysis yielded two factors, but only one with a substantial eigenvalue. If we discard all loadings below .30 and consider only the varimax solution, which differs only slightly from that obtained by oblique rotation, the first factor appears to relate to the C-Test and the subtests 'Pronouns', 'Tenses', 'Conjunctions', 'Subjunctive', 'Verbs', and 'Adjectives', whereas the second factor appears to refer to 'Questions' and also to some extent to 'Listening'.

Unfortunately, the subtests 'Listening' and 'Questions' consist of only 3 and 5 items respectively, and have the lowest reliabilities of all the subtests, namely L_3=.22 (Listening) and L_3=.55 (Questions). The loadings of these two subtests and also their correlations with the other test parts, including the C-Test, should therefore be interpreted only with great caution.

There are, however, some theoretical arguments in favour of the interpretation that these two subtests measure abilities somewhat

different from those measured by the remaining test parts, including the C-Test. The subtest 'Questions' consists of various semantically equivalent questions which are to be evaluated with regard to social appropriateness in situations described in the test. Normally, such information is rarely provided during high school French instruction and only students with relatively intensive contacts with native speakers of French presumably can solve these items. Similarly, in the subtest 'Listening', the student is confronted with a task not very familiar to him or her. This subtest consists of multiple choice questions about the content of an interview disturbed by background noise; normally, listening exercises with background noise are not very often used during high school French instruction.

However, there is also another possible explanation why precisely these two factors emerged, namely differences in the test method. A large portion of the items of the subtests which load primarily on the first factor are some sort of cloze items, whereas the subtests 'Listening' and 'Questions' consist of classical multiple choice items.

As has been pointed out again and again in the relevant literature on factor analysis, the degree of homogeneity of the subject sample is very important (see for example the discussion in Woods 1983). Unfortunately, in most factor analytic studies of language tests (including principal component analysis) the problem of subject group homogeneity has not been taken into account: If the subjects in a language test range from the lowest to the highest proficiency level, as is the case in our data, a general language proficiency factor of overwhelming importance will emerge almost by necessity (see also Hughes/Woods 1983). This has been one reason for the erroneous opinion (held for a long time e.g. by Oller) that language competence should be considered to be unidimensional and indivisible. Thus, in order to reduce the heterogeneity of our sample, we divided it into two different proficiency groups using the median of the C-Test, i.e. 65, as a cut-off point.

Tables 6 and 7 show the results of the factor analyses for Group I, who scored equal to or below 65, and for Group II, who scored above 65.

As could be expected, in both groups the importance of the first factor has been reduced. Furthermore, we now have a third factor. In the first group, according to the obliquely rotated matrix, this factor correlates with the third factor by .52. These two factors measure at least partially on the same dimension. A similar result holds for the second group. We now consider the variables 'Listening' and 'Questions'. Compared to the unselected group, the contribution of 'Listening' has been reduced in both groups, whereas 'Questions' now shows especially in the second group a substantial loading on the second factor.

With regard to the question of what the C-Test actually measures, the results of the factor analyses are however difficult to interpret. We think that in spite of the fact that the correlation between the C-Test and the BDF is .74 and thus not very far from the "magic limit"

Table 6

C-Test and BDF 1983/84:
Factor Matrices for Group I (Low Proficiency, N=56)

	Loadings						
	Varimax Rotation			Oblique Rotation			
Variable	F1	F2	F3	F1	F2	F3	h^2
PRONOUNS	.19	-.11	.83	-.06	.09	.89	.73
LISTENING	.02	-.17	-.01	.04	.18	.00	.03
QUESTIONS	.04	-.48	.08	.04	.48	.12	.24
TENSES	.38	.39	.38	.28	-.39	.28	.44
CONJUNCTIONS	.44	.47	.31	.37	-.47	.18	.51
SUBJUNCTIVE	.40	.50	.54	.24	-.51	.45	.70
VERBS	.53	-.06	.36	.47	.07	.26	.41
ADJECTIVES	.72	-.26	.24	.73	.28	.09	.64
C-TEST	.65	.08	.06	.70	-.07	-.12	.43

Initial Eigenvalues: 3.23 1.42 1.01

Variance Explained: 63%

Factor Pattern Correlations:

Factors	F1	F2	F3
F1	1.00	-.11	.52
F2	-.11	1.00	-.09
F3	.52	-.09	1.00

of .80, where tests are often considered to be interchangeable (as e.g. by Oller/Streiff 1975), the information provided by both tests is not sufficiently similar to justify the use of the C-Test instead of the BDF. We rather feel that for the grouping of students into different levels of global language proficiency, the C-Test is an adequate and economical device, but we also feel that, in order to be able to optimize instruction, we need the kind of information provided by the BDF as well.

Before we report on our qualitative analysis, a final word of caution with regard to factor analysis may be in order. It should be stressed that the kind of factor analysis used in this article should be considered only as an instrument for reducing the information contained in a correlation matrix, but not as an adequate tool for testing the dimensionality of a test (see e.g. Mottaua 1979, Raatz 1982 and Raatz 1984b for some arguments against factor analysis). Fortunately, Moosbrugger and Müller (cf. Moosbrugger/Müller 1982; Moosbrugger 1982) have recently proposed a new model, the Classical

Table 7

C-Test and BDF 1983/84
Factor Matrices for Group II (High Proficiency, N=59)

	Loadings						
	Varimax Rotation			Oblique Rotation			
Variable	F1	F2	F3	F1	F2	F3	h^2
PRONOUNS	.36	.02	.77	.21	-.02	.74	.72
LISTENING	.05	.07	.35	-.03	.06	.37	.13
QUESTIONS	.06	.73	.13	.07	.72	.12	.55
TENSES	.55	-.28	.20	.54	-.30	.08	.43
CONJUNCTIONS	.74	.16	.18	.76	.13	.02	.61
SUBJUNCTIVE	.64	-.05	.31	.62	-.08	.18	.51
VERBS	.66	.05	.21	.66	.02	.06	.48
ADJECTIVES	.71	.14	-.02	.78	.12	-.20	.53
C-TEST	.63	-.03	.37	.59	-.06	.24	.54

Initial Eigenvalues: 3.76 1.17 1.06

Variance Explained: 67%

Factor Pattern Correlations:

Factors	F1	F2	F3
F1	1.00	-.01	.42
F2	-.01	1.00	.03
F3	.42	.03	1.00

Latent Additive Test model (CLA model). This model has the same characteristics as the Rasch model. However, while the Rasch model is suitable only for binary items, the CLA model can be applied to tests whose items measure at least on an ordinal scale, which is the case for the C-Test superitems. The CLA model has been applied to a German C-Test by Raatz (1984b). The fit of the model has been tested statistically and it has been demonstrated that the parts of the test are homogeneous and that the total score is unidimensional and interval-scaled. We plan to do the same kind of analysis for different versions of the French C-Test and also for the BDF in the near future.

4 Thinking aloud and retrospective data

So far we have only said that the C-Test provides, at least to some extent, the same information as the BDF, but we have deliberately refrained from telling what it is exactly that it does measure. Although correlations between C-Test scores and scores from, for instance, other language tests or intelligence tests provide valuable information, correlational methods are not sufficient to settle the question of what the C-Test really measures. It is our opinion that we also need information about the psycholinguistic plans and processes operating in the language learner when he or she is doing a C-Test. Cohen (1984) points out correctly that due e.g. to flaws in the test or due to certain test-taking strategies, the learner might produce wrong solutions for the right reasons or correct solutions for the wrong reasons. That means that we are judging the learner's underlying competence by drawing conclusions solely from his or her performance data without taking into account what goes on inside the learner. If we succeeded in revealing the processes activated in the learner when working on the C-test, we might be able to gain a better insight into the learner's underlying competence and his or her ways of performing on the basis of his or her competence. And finally, revealing these psycholinguistic processes might enable us to draw conclusions about the demands which the C-Test makes on the learner. The principal purpose of this part of the paper is (a) to present a certain method of data collection well worth discussing;(b) to report on some of the findings obtained on the basis of this method; and (c) to give a prospective view of further investigation of our data.

In our investigation we used the learner as an informant and combined thinking aloud and retrospective methods of data collection. The subjects were given three C-Test texts of different levels of difficulty and were asked to say everything they were thinking and feeling while filling in the gaps. The thinking aloud data were recorded on tape. Subsequently, the subjects listened to the tape recording together with the experimenter in order to comment on their own utterances and also to explain those parts of the recording which were not understood by the experimenter. In this way we obtained two verbal reports, one consisting of thinking aloud data and the other of retrospective data. Since the former were produced at the same time as the subjects were doing the C-Test, these data can also be characterised as "periactional". Analogously, the retrospective data can be characterised as "postactional" (cf. Huber/Mandl 1982:18 for this terminology). Up to now, we have obtained data from 10 subjects, but only that from 3 subjects has been transcribed and tentatively analysed. This preliminary qualitative analysis will now be discussed.

Let us start by taking a look at our learner and the situation he or she encounters when dealing with the C-Test. As soon as the learner focuses on certain elements of the text, the reception process has been started. This involves activating certain knowledge which we will refer to as concept. Following Beaugrande/Dressler (1981:85) we define concept "as a configuration of knowledge that can be recovered or activated with more or less consistency and unity".Further, the authors define the content of a concept "as an ordered set of hypotheses about accessing and activating cognitive elements within a current pattern" (87). According to Hörmann (1976) the components which constitute a concept can be explored with regard to three processes: acquisition,

storage, and utilization.The investigation of such components will be one of our major interests in future analyses. Rather than trying to "decompose" such concepts into more basic units a priori, we will follow Beaugrande/Dressler's suggestion and start from empirical evidence by taking our thinking aloud and retrospective protocols as a basis for description. Descriptions of concepts within the framework of cognitive science have concentrated on two types of knowledge, i.e. declarative and procedural knowledge (cf. e.g. Anderson 1976; Winograd 1975). Whereas declarative knowledge ("knowing that") refers to a language user's underlying knowledge about linguistic structure, procedural knowledge ("knowing how") can be viewed as the sum of procedures operating on an individual's declarative knowledge in the performance of mental behavioural acts. Declarative knowledge can be differentiated according to linguistic levels and according to different languages. With regard to the mental organisation of declarative knowledge, it might be said that at least parts of the declarative knowledge of a multilingual speaker and also of a language learner are marked as language specific and hence may be stored separately from other parts of the declarative knowledge. It is important now to distinguish between how these types of knowledge are mentally stored and the way they are activated for problem solving. In this article we will focus on the way declarative and procedural knowledge is activated.

Before we take a closer look at the activation processes, let us once again look at our learner in his or her role as recipient of a defective text. In principle, a recipient may first identify small units such as minimal phonological segments which are then gradually added together to larger units (e.g. words, sentence constituents etc.). This type of process is referred to as bottom-up-processing. Alternatively, the recipient may take as a starting point higher level units (e.g. the context of a text) and work his or her way downwards. This process is often called top-down-processing. Transferring these two processing types to our special C-Test situation, it seems obvious that both processes are involved in text re-construction. In the light of what Porter (1983:74) reports as to the predictability of particular words in Cloze Procedure, it seems particularly promising to investigate the learner's use of the co- and context and its effectiveness with regard to the set task. As a result of his empirical study, Porter concludes that

> "any argument for Cloze Procedure as a technique for assessing general language proficiency in a language, or for assessing an important part of that general proficiency, based on that technique's presumed requirement of the subject that he uses wide-ranging linguistic constraints in restoring deletions, must be seriously weakened. It is possible that the subject does use information from such constraints - but he does not have to".

This argumentation is also supported by Cohen (1984:74), who reports on various course papers done by different authors with regard to the Cloze Test. He states that the majority of students taking Cloze Tests looked for a clue to the answer in the same sentence containing the deletion. With regard to reading ability, the classical Cloze Test seems to be "more of a measure of word- and sentence-level reading ability than of discourse-level reading" (75).

Hence, the use of bottom-up and top-down processing in our special problem solving situation becomes another important field of investigation. Some questions put forward will be: to what extent these processes are used strategically, to what outcome such use might lead and in which way bottom-up and top-down processes interact with each other.

After having presented some factors involved in the reception process we will now take a closer look at possible activation processes. In its most unproblematic way, concept activation is done "automatically", which implies according to Ericsson/Simon (1980:225) "that intermediate steps are carried out without being interpreted, and without inputs and outputs using short-term memory". While automatic processing is predominantly parallel and does not interfere with other processes, controlled processing is mostly serial and requires attention. (For a detailed discussion of "automatization" cf. Shiffrin/Schneider 1977; Shiffrin, R.M./Dumais 1981; Neves, D.M./Anderson 1981). In case the learner cannot activate a concept easily in order to find the appropriate item, he or she might try to solve these problems by consciously or unconsciously employing different plans and strategies.Referring to Miller/Galanter/Pribam (1960:16), we define plan as "any hierarchical process in the organism that can control the order in which a sequence is to be performed". Following Faerch/Kasper (1980:60) we generally define strategy as "a potentially conscious plan for solving what to the individual presents itself as a problem in reaching a particular goal". Which method is to be used to identify problem-solving strategies in our data depends on whether they are employed consciously or unconsciously. In the first case, the learner him- or herself can give us the clue by telling us what he or she is or was doing. But we should be aware here that what the learner is telling us that he or she is or was doing might not necessarily correspond to what he or she is or was actually doing. We might identify those problem-solving strategies which are unconsciously employed by observing what the learner is actually doing without explicitly saying so.

From the verbal protocols of our first subjects we have so far identified the following list of plans/strategies. This inventory, however, must be regarded as only a preliminary listing.

Reception Strategies

- analysing linguistic knowledge
- translation
- watch for formal indicators
- recourse to other languages
- read back in mental store
- recall of text and world knowledge

Evaluation Strategies

- checking via formal indicators
- checking via other languages
- checking via co- and context

- checking via structural analysis
- checking via "feeling for a language"
- checking via frequency of occurrence

Application Strategies

realization strategies

- interlingual transfer
- intralingual transfer
- recourse to interlanguage
- recourse to co-text
- recourse to formalism
- recourse to structural analysis

retrieval strategies (cf. also Glahn 1980)

- searching via text
- searching via other languages
- retrieval via semantic field
- retrieval via formal similarity
- retrieval from passed situations

Reception here includes both, perception - referring to the processing of speech units and comprehension - referring to the assignment of meaning to incoming linguistic information. Hence, reception strategies are aimed at solving problems arising when the learner tries to activate concepts during the reception process. We will refer to application strategies as being oriented towards the conversion of the activated concept into a linguistic form. Evaluation strategies might be used by the learner in order to check the appropriateness or inappropriateness of the activated concept or its converted linguistic form.

Since it will not be possible in this paper to discuss in detail the components of a concept so far identified, we will now give as examples at least some illustrations from our peri- and post-actional verbal protocols. (For presentation purposes the original German spoken data are translated into English. Because of the nature of our data only an approximate translation is possible. The examples will refer to text 4 and text 5 from the C-Test).

Text 4

A en croire les statistiques, le Français consacre plus de deux heures et demie de sa journée à regarder la télévision, ma_____ ce chi_____ n'indique qu'_____ moyenne gêné_____ masquant l_____ différences en_____ les mo_____ de v_____ en par_____ déterminés par_____ les situa_____ professionnelles e_____ familiales. En_____ l'enfant e_____ l'adulte, l_____ mère d_____ famille e_____ le retr_____, l'ouvrier e_____ le ca_____ supérieur, l_____ rural e_____ le citadin, le temps que chacun passe chez soi diffère comme la manière d'employer la part de détente laissée libre par les occupations domestiques.

Text 5

"Mes filles aiment beaucoup la publicité télévisée. Elles sont très jeunes et c'est vraiment ce qu'elles préfèrent. Je dir_____ même q_____ cela l_____ passionne. J_____ pense q_____ c'est pa_____ qu'elles retro_____ des scé_____ de l_____ vie quoti_____ qui, po_____ elles, repres_____ quelque ch_____. Les pet_____ films public_____ les intér_____ parce qu'_____ voit u_____ maman fa_____ sa vais_____ ou pas_____ son aspir_____, une famille à table, un enfant manger son fromage ou sa crème glacée."

Example 1

(thinking aloud)

Learner 1: je pense que c'est (...)

(text 5, line 3/4)

Example 2

(thinking aloud)

Learner 2: que cela leur passionne I always say or I say that it or that this - is their passion or something like that don't know passionne don't know what to do with this ehm - je pense I think que c'est ehm pa what's that pa (...)

(text 5, line 3)

Example 3

(a) thinking aloud

Learner 3: the next sentence --je p- yes probably je - je pense -- je pense que c'est -- wow that's really (...)

(text 5, line 3/4)

(b) retrospection

Interviewer: what I'd also like to know now - if you can still remember is if je and je pense que je and que was that already written down

Learner 3: well yes whenever I went on to the next sentence I filled in what I thought I knew

Interviewer: spontaneously]

Learner 3: I knew spontaneously]

Examples 1-3 reflect typical learner behaviour concerning the items je and que. Both items were filled in correctly in 99% of all cases. In all three examples these two items can be regarded as being produced automatically. With regard to example 2 it should be noted that after pronouncing the words je pense the learner translates them

into her mother tongue without pausing and directly continues with que c'est etc. Here we might hypothesize that after concept activation had been done automatically, the learner used translation as an evaluation strategy. Indicators of the activation or evaluation strategies in example 3 are the pauses before pronouncing the items and the mother tongue utterance "yes probably je". In the subsequent interview the learner confirms that the solution to these items occurred to her spontaneously.

The next item we want to look at was not filled in at all by 39 and filled in inappropriately by 30 of our 115 subjects.

Example 4

(a) thinking aloud

Learner 1: (...) son aspi - well I really don't have any idea what that could be - son aspi - aspirine (laughs) - aspiron mmh can't think of anything aspiration - no I don't know that word - (...)

(text 5, line 9)

(b) retrospection

Learner 1: I tried to recall all kinds of words starting like this like aspi well yes I somehow tried to do brainstorming

Here the learner tries to recover the item by reading back all the items starting with aspi in her mental store, but without success. In the interview the learner confirms this strategy. Note that the learner activates actual as well as non-existent French words. The learner in this example does not complete the item. Examining how this item was completed in our data set of 115 students, we find aspiration to be the most frequently used filler.This might indicate the use of a realization strategy such as e.g. intralingual transfer or recourse to interlanguage.

The next example is taken from text 4.

Example 5

(thinking aloud)

Learner 1: okay now let's try to continue entre l'enfant et l'adulte la mère - de famille --- e le - mmh --- whom does she find again (laughs) in this context - le then what does this le refer to - entre l'enfant et l'adulte et l'adulte - la mère de famille -- there should be a comma - but there isn't any if this continues continued elle le retrouve ----- la mère de famille -- mmh yes with an e that could also be an adjective which refers to famille -- entre l'enfant et l'adulte la mère de famille -- e le re - retrouve ---- -- well yes - could also be that this word starting with retr is a noun which would then fit in here into

this listing -------- la mère de famille --- well if there'd be a comma then it would be pretty easy then I'd write elle le retrouve though in such a case this le also had to be plural cause it all refers to l'ouvrier e de cades supérieur le rural et le citadè citadin -- mmh ---- (sighs) how difficult (...)

(text 4, lines 6-8)

Let us take a closer look at the gap starting with retr. In the first instance this gap was completed spontaneously with retrouve. As this word already occurred in text 5 (which in our test was completed earlier than text 4) and was filled in correctly by our learner, one could hypothesize that because of the first letters being similar - namely ret - the learner automatically completes this item by unconsciously recalling the previously activated concept. The learner then activates an evaluation strategy by referring to the context: "mmh whom does she find again in this context - le then what does this le refer to -....". As the learner does not find a satisfactory answer, she tries to evaluate her production with the help of a formal indicator, i.e. a comma. But since such an indicator is lacking, the testee questions the appropriateness of the activated concept and after further structural analysis concerning the item le as presenting an object, finally rejects it. A new concept is put forward - again by taking recourse to structural analysis. Instead of regarding the item as a verb, the learner first thinks of it as an adjective and then as a noun. The activated concept is now evaluated by structurally analysing the co-text: as a listing of nouns is already given, it seems possible to the learner that the uncompleted item also is a noun.

Our last example is again taken from text 4.

Example 6

(a) thinking aloud

Learner 1: - l'ouvrier e le - ca [ka] supérieur just a minute this ca [ka] here cades - that was treated once in my French classes I remember that quite well - ehm well how was that called again was kind of a higher employee cades supérieur cades or yes something like that -- mmh whether this really meant cades well yes I think so yes les cades supérieur ...

(text 4, line 9)

(b) retrospection

Learner 1: here I didn't know whether that was cad or something like that cad or
Interviewer: mmh did you know what that was supposed to mean
Learner 1: uhum later on I said
Interviewer: ah we'll hear that later on
Learner 1: that means a higher employee or something like that

Interviewer: ah yes mmh
(cassette continues)
and you remembered that from your French classes
Learner 1: uhum
Interviewer: long ago isn't it
Learner 1: yes uhum
Interviewer: that's something - what made you remember that - was there anything special or
Learner 1: no we talked about professions with l'ouvrier and cades supèrieur something like that and - somehow I remembered that

We can say that we are dealing here with the "tip of the tongue" phenomenon, i.e. a state in which one has difficulty recalling or cannot quite recall a familiar word but can recall words of similar form and meaning (cf. also Brown/McNeill 1966). The learner tries to recall the item cadre by bringing to mind a past situation (i.e. recalling an other experience as Stevick (1984) would put it) and the semantic field in which the item occurred. The learner here gets very close to the actual target word. The question to what extent thinking aloud and retrospective methods provide insights into mechanisms by which such words and their mode or storage in memory are recalled seems to be another field worthy of investigation.

The learner's problem-solving behaviour in the above examples takes us back to some considerations discussed in the first part of this article. It was said that statistical analyses alone will not suffice when trying to achieve more insight into what the C-Test actually measures. Example 5 above gives clear evidence that the learner here frequently takes recourse to specific aspects of her declarative knowledge in order to find the appropriate items. At a particular point it is rather obvious that the learner activates stored interlanguage knowledge about the use of the indirect object. But other learner behaviour in this example also makes clear that the question whether the declarative knowledge activated can be regarded (1) as being neutral with regard to the interlanguages possibly involved in the processing or (2) as being marked as interlanguage specific, still needs further investigation.

Another point put forward earlier referred to the problem of getting access to the mental processes involved. Nobody would deny that there are certain limits. But at the same time the above examples should also have made it clear that thinking aloud and retrospective reports might give us valuable insights into the learner's state of mind. Though no explicit example was given here, it should be mentioned that later in the interviews, when asking the learner about a particular long pause, we sometimes succeeded in eliciting very detailed verbal explanations about what the learner thought was going on in his or her mind during that time.

To summarize this last part of our paper, we think that we can agree with Ericsson/Simon (1980:247) who believe that

"... verbal reports, elicited with care and interpreted with full understanding of the circumstances under which they were obtained, are a valuable and thoroughly reliable source of information about cognitive processes ..."

We must add that we feel rather sure that the question of what the C-Test actually measures cannot be answered without qualitative analyses of the verbal reports of language learners as C-Test takers.

5 Future Perspectives

This paper should be regarded as only a preliminary step in the analysis of the C-test principle. Some further research has already been started.

- The C-Test data basis has been extended to more than 400 subjects.
- The test items have been classified according to content and function words and a preliminary analysis has been done.
- Various background data on our C-Test-taking subjects have been collected.
- The starting point for the deletion of the items was shifted from position n to position n+1 or n-1. In so doing we obtained a second version of the C-Test, which will be investigated as to whether it is parallel with regard to content, difficulty, etc.
- More than 80 subjects who had done the BDF and the C-Test were interviewed in French and rated according to their oral proficiency in order to find out to what extent the C-Test results and the oral proficiency ratings correlate.

As further investigations, we are planning the following.

- The dimensionality of the C-Test together with the BDF will be investigated with the help of the CLA model proposed by Moosbrugger and Müller.
- The possibility of an advance prediction of C-Test text difficulty (Klein-Braley 1984) will be studied.
- The construct validity of the C-Test with regard to different proficiency levels will be analysed.
- More thinking aloud and retrospective data will be collected including data for English and Spanish C-Tests.
- The problem-sloving behaviour of the subjects will be investigated more thoroughly and related to the subject's actual production.
- The role the context plays in this particular problem-solving situation will be studied.
- It will be investigated to what extent procedural and declarative knowledge can be distinguisned in our data and whether both types of knowledge are marked as being neutral or interlanguage specific.

Note

1 Since we have analyzed only a preliminary version of the BDF, we have deliberately refrained from doing a more sophisticated reliability analysis (for a good review of some further aspects of the reliability problem see Krzanowski/Woods 1984).

Bibliography

Alderson, J.Ch. 1979. "The cloze procedure and proficiency in English as a foreign language". TESOL Quaterly 13/2: 219-227.

Anderson, J.R. 1976. Language, memory and thought. Hillsdale, etc.: Erlbaum.

- (ed). 1981. Cognitive skills and their acquisition. Hillsdale, etc.: Laurence Erlbaum Associates.

Beaugrande, R.-A., Dressler, W.U. 1981. Introduction to text linguistics. New York, etc.: Longman.

Brown, R., Mc Neill, D. 1966. "The 'tip-of-the-tongue' phenomenon". Journal of Verbal Learning and Verbal Behavior 5 (4): 325-337. Reprinted in: Gardiner, J.M. (ed). 1976. Readings in human memory. London: Methuen + Co.Ltd.

Cohen, A.D. 1984. "On taking language tests: what the students report." Language Testing 1: 70-81.

Ericsson, K.A., Simon, H.A. 1980. "Verbal reports as data". Psychological Review 87 (3): 215-251.

Faerch, K., Kasper, G. 1980. "Processes and strategies in foreign language learning and communication". Interlanguage Studies Bulletin 5: 47-118.

Grotjahn, R. "Der Bochumer Diagnostiktest 'Französisch'". Paper presented at 15th Annual Conference of the Gesellschaft für Angewandte Linguistik, GAL e.V., Berlin, September 1984. (To appear in the Proceedings of the Congress).

Grotjahn, R., Stemmer, B. 1984. "Entwicklung und Einsatz eines C-Tests 'Französisch'". Kühlwein, W. (ed). Sprache, Kultur und Gesellschaft. Tübingen: Narr.

Guttman, L. 1945. "A basis for analyzing test-retest reliability". Psychometrik 10: 255-282.

Hörmann, H. 1976. Meinen und Verstehen. Frankfurt: Suhrkamp.

Huber, G.L., Mandl, H. 1982. "Verbalisationsmethoden zur Erfassung von Kognitionen im Handlungszusammenhang". In Huber, G., Mandl, H. (eds). 1982. Verbale Daten. Weinheim, Basel: Beltz. 11-42.

Hughes, A., Woods, A. 1983. "Interpreting the Performance on the Cambridge Proficiency Examination of Students of Different Language Backgrounds". In Hughes, A., Porter, D. (eds). 1983. Current Developments in Language Testing. London, New York: Academic Press. 53-62.

Hull, C.H.N. (ed). 1981. SPSS Update 7-9. New Procedures and Facilities for Releases 7-9. New York: Mc Graw Hill.

Klein-Braley, C. (1981). Empirical Investigations of Cloze Tests. An Examination of the Validity of Cloze Tests as Tests of General Language Proficiency in English for German University Students. Unpublished Ph.D. Dissertation. Duisburg.

- 1984. "Advance prediction of difficutly with C-Tests". In Culhane, T., Klein-Braley, C., Stevenson, D.K. (eds).Practice and problems in language testing 7. Proceedings of the Seventh International Language Testing Symposium of the Interuniversitäre Sprachtestgruppe (IUS), Colchester, October 1983.Colchester: University of Essex.

Klein-Braley, C., Raatz, U.(1984). "Survey report on the C-Test". Paper presented at the 7th World Congress of Applied Linguistics, Brussels, August 1984.(To appear in Language Testing 1985).

Krzanowski, W.J., Woods, A. (1984). "Statistical aspects of reliability in language testing". Language Testing 1: 1-20.

Miller, G.A., Galanter, E., Pribram, K.H. (1960). Plans and the Structure of Behavior. Holt, Rinehart, Winston Inc.

Moosbrugger, H. (1982). "Dimensionalitätsuntersuchungen von FPI-Skalen mit dem Klassischen Latent-Additiven Testmodell (KLA-Modell)". Zeitschrift für Differentielle und Diagnostische Psychologie 3: 241-264.

Moosbrugger, H., Müller, H. (1982). "A Classical Latent Additive Test Model". The German Journal of Psychology 6: 145-149.

Neves, D.M., Anderson, J.R. (1981). "Knowledge Compilation. Mechanisms for the Automatization of Cognitive Skills". In Anderson, J.R. (ed). (1981. 57-84.

Oller, J.W.,Jr. (1976). "Evidence for a general language proficiency factor: an expectancy grammar". Die Neueren Sprachen 76: 165-174.

Oller, J.W., Streiff, V. (1975). "Dictation: A Test of Grammar-Based Expectancies". English Language Teaching Journal 30: 25-36.

Porter, D. (1983). "The Effect of Quantity of Context on the Ability to make Linguistic Predictions: a Flaw in a Measure of General Proficiency". In Hughes, A., Porter, D. (eds). (1983). 63-74.

Raatz, U. (1982). "Language theory and factor analysis". In Lutjeharms, M., Culhane, T. (eds). (1982). Practice and problems in language testing III. Proceedings of the Third International Language Testing Symposium of the IUS. Brüssel: Vrije Universiteit. 30-56.

- 1984. "The factorial validity of C-Tests". In Culhane, T., Klein-Braley, C., Stevenson, D.K. (eds). (in press). Practice and problems in language testing 7. Proceedings of the Seventh International Language Testing Symposium of the Interuniversitäre Sprachtestgruppen (IUS), Colchester, October 1983. Colchester: University of Essex.

- 1984b. "Better Theory for Better Tests?" Paper presented at the Academic Committee for Research on Language Testing (ACROLT) Conference, "Authenticity and Language Testing" held 16-18 May 1984 in Quiryat Anavim, Israel.

Raatz, U., Klein-Braley, C. (1982). "The C-test - a modification of the cloze procedure". In Culhane, T., Klein-Braley, C., Stevenson, D.K. (eds). (1982). Practice and problems in language testing IV. Proceedings of the Fourth International Language Testing Symposium of the IUS. Colchester: University of Essex. 113-138.

- 1983. "Ein neuer Ansatz zur Messung der Sprachleistung. Der C-Test: Theorie und Praxis". In Horn, R., Ingenkamp, K., Jäger, R.S. (eds). (1983). Tests und Trends 3. Jahrbuch der Pädagogischen Diagnostik Weinheim, Basel:Beltz. 107-138.

Raju, N. S. (1977). "A Generalization of Coefficient Alpha". Psychometrika 42: 549-565.

Shiffrin, R.M., Dumais, S.T. (1981). "The Development of Automatism". In Anderson, J.R. (ed). (1981). 111-140.

Shiffrin, R.M., Schneider, W. (1977). "Controlled and automatic human information processing: I. Detection, search, and attention". Psychological Review 84: 1-66.

- 1977. "Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory". Psychological Review 84: 127-190.

Stevick, E.W. (1984). "Memory, Learning and Acquisition". In Eckman, F.R., Bell, L.H., Nelson, D. (eds). (1984). Universals of Second Language Acquisition. Rowley, Mass.: Newbury House. 24-35.

Süßmilch, E. (1984). "Language testing with immigrant children". In Culhane, T., Klein-Braley, C., Stevenson, D.K. (eds). (in press). Practice and problems in language testing 7. Proceedings of the Seventh International Language Testing Symposium of the Interuniversitäre Sprachtestgruppe (IUS). Colchester: University of Essex.

Winograd, T. (1975). "Frame representation and the declarative-procedural controversy". In Bobrow, D., Collins, A. (eds). (1975). Representation and Understanding: Studies in Cognitive Science. New York: Academic Press. 185-210.

Woods, A. (1983). "Principal Components and Factor Analysis in the Investigation of the Structure of Language Proficiency". In Hughes, A., Porter, D. (eds). (1983). Current Developments in Language Testing. London, New York: Academic Press. 43-52.

Wottawa, H. (1979). Grundlagen und Probleme von Dimensionen. Meisenheim: Hain 1979.