Reihe:
FREMDSPRACHEN IN LEHRE UND FORSCHUNG (FLF)

Im Auftrag der Ständigen Kommission des Arbeitskreises der Sprachenzentren, Sprachlehrinstitute und Fremdspracheninstitute (AKS)

herausgegeben von

Klaus Vogel und Bernd Voss

Band 31

Coleman, James A., Grotjahn, Rüdiger & Ulrich Raatz (eds.)

University Language Testing and the C-Test

AKS-Verlag, Bochum

2002

Koller, G. And R. Zahn (1996), 'Computer based construction and evaluation of C-Tests', in R. Grotjahn (ed.) (1996a), 401-418

Meißner-Stiffel, M., and U. Raatz,(1996), '_____ or _____? Zwei Grundlagenuntersuchungen zum C-Prinzip bei L1-Lernern' in R. Grotjahn (ed.) (1996a), 173-81.

Oller, J. W. (1976), 'Evidence for a general language proficiency factor, an expectancy grammar', *Die Neueren Sprachen*, 75, 165-74 (reprinted in Oller (1983), 3-10).

Oller, J. W. (ed.) (1983), *Issues in Language Testing Research*, Rowley, Mass., Newbury House.

Raatz, U. (1984), 'The factorial validity of C-Tests' in T. Culhane, C. Klein-Braley and D. K. Stevenson (eds.) (1982), 124-39.

Raatz, U., (1985a), 'Better theory for better tests?', *Language Testing*, 2, 60-75

Raatz, U. (1985b), 'Investigating dimensionality of language tests - a new solution to an old problem', in V. Kohonen, H. Von Essen and C. Klein-Braley (eds.), *Practice and problems in language testing* 8, 123-136, Tampere, AFinLA.

Raatz, U. (1985c), 'The factorial validity of C-Tests', in C. Klein-Braley and U. Raatz (eds.) (1985a), 42-54.

Raatz, U. and C. Klein-Braley (1982), 'The C-Test - a modification of the cloze procedure', in T. Culhane,C. Klein-Braley and D. K. Stevenson (eds.) (1982), 113-38.

Raatz, U., and C. Klein-Braley (1983), 'Ein neuer Ansatz zur Messung der Sprachleistung. Der C-Test, Theorie und Praxis', in R. Horn, K. Ingenkamp and R. S. Jäger (eds.) *Tests und Trends 3. Jahrbuch der Pädagogischen Diagnostik*, Weinheim and Basel, Beltz. 107-38

Raatz, U., and C. Klein-Braley (eds.) (1985a), *Fremdsprachen und Hochschule, 13/14, Thematischer Teil, C-Tests in der Praxis*, Bochum, AKS.

Raatz, U., and C. Klein-Braley (1985b), 'How to develop a C-Test', in C. Klein-Braley and U. Raatz (eds.), (1985), 20-22.

Raatz, U., and C. Klein-Braley (1992), *CT-D 4. Schulleistungstest Deutsch für 4. Klassen. Test und Beiheft mit Anleitung und Normentabellen*, Weinheim, Beltz.

Roos, U. (1994), 'The C-Test in Japanese' in R. Grotjahn (ed.) (1994), 61-113.

Roos, U. (1995), *Ein C-Test für Lerner der japanischen Sprache. Entwicklung, Erprobung und Validierung*, Bochum, AKS.

Sigott, G., and J. Köberl (1996), 'Deletion patterns and C-Test difficulty across languages', in R. Grotjahn (ed.) (1996a), 159-72

Spolsky, B. (1973), 'What does it mean to know a language or how do you get someone to perform his competence?', in J. W. Oller and J. C. Richards (eds.) (1973), *Focus on the learner*, Rowley, Mass., Newbury House, 164-76.

Stemmer, B. (1991), *What's on a C-Test taker's mind, Mental processes in C-Test taking*, Bochum, Brockmeyer.

Stemmer, B. (1992), 'An alternative approach to C-Test validation' in R. Grotjahn (ed.) (1992), 97-144.

Seven: C-Tests and language processing

Rüdiger Grotjahn and Brigitte Stemmer

Usually the construct validity of a language test is investigated by statistically correlating the observed test scores of the examinees with their observed scores on other purported measures of human abilities, such as other language tests, intelligence tests, or school grades.

It has been argued by various authors (cf. e.g. Grotjahn, 1986, 1994; Messick, 1989) that this approach is unsatisfactory for several reasons. One major argument put forward is that without an understanding of the individual cognitive processes and task-specific mental operations on which the observed performance depends, we cannot really understand what a (language) test measures. Therefore, in addition to the directly observable test performance, the covert individual mental processes in the test-taking subject also need to be analyzed. (A more detailed justification can be found in Anderson, Bachman, Perkins and Cohen, 1991; Grotjahn, 1986, 1994; Messick, 1989.)

1. Mental processes in C-Test taking: the introspective approach

1.1 The Bochum C-Test project

For nearly a decade the Bochum C-Test project has focused on investigating the mental processes going on in a subject when he or she is working on a C-Test. Preliminary results have been published, for example, by Grotjahn and Stemmer (1985) as well as Feldmann and Stemmer (1987). The most comprehensive and detailed account is given by Stemmer (1991, 1992).

One of the objectives in this project was to investigate *how* the subjects tried to solve C-Tests. To pursue this goal, data were collected by using introspective and retrospective methods, and a model of analysis, which relies on the assumption that C-Test solving is a cognitive task, was developed and applied to the data.

1.2 Data collection

Data collection was carried out as follows. Thirty German native speakers with an average of 5.7 years of French instruction were tested individually in separate sessions. The subject was instructed to think aloud while solving a French C-Test consisting of three texts. The utterances were audiotaped. Immediately after the thinking-aloud phase, the interviewer and the subject jointly listened to the audiotaped utterances and the subject was encouraged to comment spontaneously

on his or her thinking-aloud utterances during the previous test phase and answer the interviewer's questions. This retrospective phase was audiotaped as well. Both the thinking-aloud and retrospective verbalizations were then transcribed literally. A similar design was used with a Spanish C-Test consisting of four texts. For Spanish, data were collected from 50 subjects. However, since only a very small amount of the Spanish data has been analyzed in detail, we will restrict ourselves to the French data (a detailed discussion of the data collection procedure can be found in Grotjahn, 1986, 1987 and Stemmer, 1991, 1992).

1.3 Analyses

1.3.1 Model of analysis

Since no adequate model of analysis existed, the major difficulty was to establish a consistent analytical model to describe the *reported* processes and to reconstruct *non-reported* processes. The model finally adopted by Stemmer (1991) consisted of three main parts: (1) a task analysis; (2) an analysis of the verbal report data, and (3) an analysis of the performance data, i.e. of the completed test. The details of the model read as follows:

(1) Task Analysis
  (1a) Analysis of mutilated words
    Analysis of word beginning
    Word frequency
    Syntactic characteristics (grammatical properties and classification into function words and content words)
    Cohesive ties
  (1b) Text analysis
    Propositonal analysis
    Coherence/cohesion analysis
(2) Analysis of verbal protocols
  (2a) Problem-solving behaviour
  (2b) Taxonomy of problem-solving strategies
  (2c) Verbal Protocol Mapping Graph (VPMG)
(3) Analysis of performance Data

(1) Task analysis
As a first step, a task analysis was performed. An *a priori* task analysis can indicate what kind of cognitive processes are involved in the current C-Test task, such as, for example, the particular knowledge involved in producing a solution to an item,

i.e. in filling in the missing part of a word. In this way task analyses can support data interpretation and validation.

(1a) Analysis of mutilated words
Research in the areas of word perception and word recognition suggested including
- an analysis of a C-Test item (i.e. the mutilated word) with regard to the letters given and the extent to which they convey " meaning";
- an investigation of the frequency with which a C-Test item occurred in the French language;[1]
- an examination of the syntactical properties of the items; and
- an analysis of the cohesive ties of the items.

(1b) Text analysis
In addition, an analysis of the non-mutilated text was carried out. The rationale for this procedure was based on the assumption that by performing a propositional analysis, including a cohesion/coherence analysis, it was possible to obtain information about the semantic units in each text and about the position of an item within the propositional and textual structure. This analysis is described in some detail below.

(2) Verbal protocols

(2a) Problem-solving behaviour
The next step in our approach was an analysis of the verbal protocol data. In this way we attempted to gain insight into how the problem was approached and which type of knowledge was activated. Central to this analysis was the concept of (problem-solving) strategy.

(2b) Taxonomy of problem-solving strategies
Based on the analysis of 30 verbal protocols, 32 different types of strategy, falling into two major categories, were identified (cf. Stemmer, 1991, Appendix 6; 1992, Appendix 2; cf. also Grotjahn and Stemmer, 1985; Feldmann and Stemmer, 1987). The two major strategies are (1) recall strategies, employed to retrieve an item, and (2) evaluation strategies, employed to check the appropriateness or correctness of a retrieved item. We will not discuss the taxonomy of problem solving strategies but rather focus on the propositional analysis of both the non-mutilated texts and the

[1] Cf. the recent study by Köberl and Sigott (1994) who found for an English C-Test administered to native speakers of English that 'more frequent items indeed tend to show higher item facility and vice versa' (1994: 60).

verbal protocols.

Propositions and text comprehension

Propositions are considered elementary units of meaning. Formally, they are represented as a relation-concept (predicate) which is connected to specific content-concepts (arguments). Propositional analyses have been widely used, for example in analysing recall protocols and narrative texts, and ample evidence has been accumulated that propositions can be regarded as psychological processing units (cf. Schnotz, 1994: 150ff.; van Dijk and Kintsch, 1983: 38ff.).

Both Kintsch and van Dijk (1978) and van Dijk and Kintsch (1983) assume that because of the comprehender's limited cognitive processing capacities, a text is processed in several cycles. During each cycle a certain number of phrases is taken into working memory and processed in the form of propositions. While reading a text, new propositions are connected to previously processed ones. In this way, the comprehender tries to construct a *text base*, i.e. a coherent 'semantic representation of the input discourse in episodic memory' (van Dijk and Kintsch, 1983: 11). In addition, by means of inferencing, a *situation model*, that is 'the cognitive representation of the events, actions, persons, and in general the situation, a text is about' is constructed in episodic memory (van Dijk and Kintsch, 1983: 11f.; cf. also Schnotz, 1994: 177ff.).

According to van Dijk and Kintsch (1983) text comprehension can thus be regarded as the strategic process of construction, elaboration, evaluation and revision of text bases and situation models - a view adopted by Stemmer (1991, 1992) in her propositional analyses.

Propositional analysis and C-Test solving

If one assumes that in most cases at least a minimum of comprehension is involved in C-Test solving, according to the model the subject will extract information from the C-Test text in such a way as to form a propositional text base. However, the propositional text base constructed by the subject may not necessarily coincide with that of the original non-mutilated C-Test text. Therefore what is needed is a way to compare both, the meaning represented by the non-mutilated C-Test text and the meaning inferred by the subject from the mutilated C-Test text. One way to accomplish this is to compare the propositional structure of the non-mutilated C-Test text with the propositional structure of the mutilated C-Test text as constructed by the subject. The propositional text base extracted from the non-mutilated C-Test will thus serve as a kind of norm or default against which the propositional text as constructed by the subject will be compared. Therefore, a first step is to perform a propositional analysis of the non-mutilated C-Test text. We will thus obtain information about the semantic units the text is composed of as well as information

about the position of an item within the propositional structure. This in turn allows us to investigate whether the particular position of an item within the propositional structure influences in any way the retrieval of the item.

Propositional analysis of the non-mutilated C-Test texts

Based on Frederiksen's (1986) model, a propositional analysis was performed on on all three non-mutilated C-Test texts yielding a so-called propositional graph for each text.[2] The propositional graphs were then used in the analysis of the thinking-aloud data. To illustrate this type of analysis, Text 1 and its propositional graph are displayed in the Appendix.

(2c) Verbal Protocol Mapping Graph (VPMG)

In the next step, a so-called verbal protocol mapping graph for each of the 30 subjects and each text was constructed by the following procedure: from the propositional text base of the non-mutilated text, all atomic propositions were extracted and for ease of analysis coded with numbers, i.e. each number at the left of the Verbal Protocol Mapping Graph refers to an atomic proposition. As soon as the subject uttered the first word when thinking aloud, the utterance was then mapped onto its corresponding atomic proposition and marked as the starting-point of a reading cycle. Whenever the subject left out more than one word in a sequence, the end-point of the cycle was reached. In such a case, the subject might, for example, jump back to the beginning of the text, or jump forward two or more words, or even go to a new sentence or a new paragraph. It was further noted whether the subject solved an item correctly or incorrectly and whether he or she disregarded an item during the reading cycle. An example of a Verbal Protocol Mapping Graph (Text 1, Subject 19) is given in the Appendix (see also Stemmer, 1991: 357; Stemmer, 1992: 144).

---

[2] See Stemmer (1991: 348f.; 1992: 141). For a recent application of Frederiksen's model to the analysis of think-aloud protocols of the writing process, see also Bracewell and Breuleux (1994). Since 1986, Frederiksen's model of discourse processing has been advanced considerably. The model views discourse as a form of situated cognition and provides hypotheses about the representation and processing of verbal and non-verbal information. In recent studies, the model has been used to describe narrative discourse processing in normal and brain-damaged individuals by comparing the discourse structure obtained from the analysis of narrative recall protocols of the subjects with the default discourse structures obtained from the analysis of the original narrative (Frederiksen and Stemmer, 1993; Stemmer, Joanette, Frederiksen and Marchand, 1994). This procedure yielded new insights into the discourse processing of brain-damaged subjects.

By constructing the Verbal Protocol Mapping Graph, we hoped to answer the following questions:

- Do the starting- and end-points observed in a subject's reading behavior reflect the propositional organisation of the text?
- Are there any starting- or end-points that occur more often than others? And if so, why?
- Is there any connection between item retrieval and the propositional embedding of the item?

1.3.2 Connecting strategies and propositions

Van Dijk and Kintsch (1983) argue that comprehension is a *strategic* process with propositional micro- as well as macrostructures built *strategically*. If one follows this view, the strategies identified from the thinking-aloud protocols can be interpreted as an indirect indication of the subjects' building of propositional micro- and/or macrostructures. In other words, we identify the strategy a subject employs and look at the propositions such a strategy entails. We can then compare the propositions realized in the strategy with the propositional default structure of the non-mutilated C-Test text. If the propositions identified in the strategy only operate on elements of one proposition of the non-mutilated text base, this would indicate an orientation towards low level comprehension at this specific point. On the other hand, if the strategy operates across propositional boundaries this would indicate that higher level comprehension is involved. Relating the propositional content of various types of strategies to the propositional structure of the non-mutilated C-Test text thus gives us some insight into the level of comprehension involved.

To sum up, a complete analysis of strategies and propositions is intended to indicate:

- which type of strategy is used with which item;
- the preference for strategies to operate on an inter-propositional or an intra-propositional level;
- the relationship between strategy type, intra- and inter-propositional operation of strategies, and success in supplying correct items.

1.3.3 Problems

There are various problems in the approach outlined so far. Obviously, the thinking-aloud data reflect only a portion of what is going on mentally. We do not know what other processing occurs at the same time. Nor do we know what kind of processing occurs when the subject hesitates or makes pauses. For example, one might ask whether the starting- and end-points identified in the Verbal Protocol Mapping

Graph reflect the psychological reality we interpret it to be. There is always the possibility that the subject continues reading silently, thus leaving us without a clue as to the end-point of the reading cycle. There was indeed some evidence in the protocols that exactly this had occurred. The subject sometimes seemed to simply 'skip' a word and continue reading after the skipped word without hesitation. In such cases, the skipping of one word was disregarded, and it was assumed that reading continued at this point. Of course the same thing could happen with much larger portions of the text. Here, however, the reading was assumed to be interrupted, which is reflected in the Verbal Protocol Mapping Graph as a new reading cycle.

As the difficulty indices of .71 (Text 1), .59 (Text 2) and .53 (Text 3) reported in Stemmer (1992) show, the C-Test texts were not too difficult for our subjects. Therefore, it is reasonable to assume that the process of reading a C-Test test does indeed involve building some kind of a propositional text base and thus also involves at least a certain level of understanding. Nevertheless, there is the possibility that a subject solves an item without understanding it. The subject may complete the item by mere guessing, or by simply adding letters to the mutilated word. In addition, with very proficient subjects and very easy texts, the blanks may be filled in more or less automatically without a deeper understanding of the text.

1.3.4 Statistical analysis of strategy use and proposition boundary crossing

Recall Strategies

We shall now present some results of the statistical analysis of the type of strategies employed and of whether a particular strategy involved propositional boundary crossing.[3] The most frequent type of recall strategy was "repetition of preceding/following words or clause" without propositional boundary crossing (33%). This strategy was employed nearly twice as often as the next type of recall strategy in rank order, which is "repetition of item beginning" (17%). The strategy types which operate on an interpropositional level occupy the end of the rank list. This suggests that very often only the immediate context was used to recall an item, and that integration in the sense of the construction-integration model as outlined by Kintsch (1988) plays only a minor role in item recall.

Evaluation Strategies

The situation looks somewhat different in the case of evaluation strategies. Here the leading type of strategy is "translation of co-text +/- item" (31%) closely followed by "repetition of item" (26%). The first strategy type indicates that the subject is

[3] A detailed presentation and discussion of the results of the various statistical analyses can be found in Stemmer (1991, 1992).

more concerned with incorporating the meaning of the context, whereas the second strategy type indicates that the subject does not concentrate upon meaning.

Content vs. Function Words

If we investigate the ratio of strategies to content and function words for each text, we note a tendency to use more strategies with content words and less strategies with function words. Furthermore, the more difficult the text becomes, the more strategies per content word and the less strategies per function words occur.[4] This can be interpreted as an indication that at least with more difficult texts the solving of a C-Test involves a considerable amount of semantic processing.

Reading cycles and boundary crossings

We shall now consider the length and number of reading cycles performed by the subjects, the number of propositional boundary crossings and whether there is any correlation between these statistics and text difficulty measured by the number of correct item solutions.

Stemmer (1992) found the following Pearson correlations between the test scores and the *length* of the reading cycles measured by the number of atomic propositions: -.26 ($p > .10$) for Text 1, .37 ($p < .05$) for Text 2, and .67 ($p < .01$) for Text 3 (two-tailed probabilities). Thus in the case of the more difficult texts 2 and 3, the longer the reading cycles performed by a subject, the higher the subject's score on the text. For Text 1, however, which is the easiest text, no such tendency can be observed. Instead, the opposite tendency was observed: the shorter the reading cycles, the higher the score.

Test scores and the *number* of reading cycles are positively correlated for Text 3 ($r = .41$; $p < .05$) and negatively for Text 1 ($r = -.34$; $p < .10$). Text 2 shows a slight (non-significant) positive correlation ($r = .29$; $p > .10$). This means that the more reading cycles were produced, the higher the score on Text 3, while the situation is reversed for Text 1 where fewer reading cycles led to a higher score.

Performing longer reading cycles entails including more context. This suggests that the subject tries to integrate more information in order to retrieve an item. Increasing at the same time the number of reading cycles may reinforce this process. For the more difficult texts 2 and 3 this approach seems rewarding as reflected in the higher scores. For easy texts, however, like Text 1 in the present C-Test, this procedure seems less than optimal. One reason might be that an easy text is solved to a large

[4] Note that Sigott (personal communication) found verbs to be the most difficult C-Test items. It would be interesting to find out whether this holds primarily for verbs with a large number of semantic and structural relations (e.g valencies in the sense of dependency grammar, or cases in the sense of case grammar).

extent automatically, so that it is unnecessary to perform lengthy and numerous reading cycles.

Furthermore, the data suggested that easy C-Test texts tended to be solved by operating within one propositional boundary while solving the more difficult C-Test text involved more boundary crossings (cf. Stemmer, 1992: 124f.).

1.4 Preliminary conclusion

We thus have evidence from several sources that in C-Test solving the subject operates predominantly within one meaning unit. If we differentiate between recall and evaluation strategies, this tendency is stronger for the retrieval than for the evaluation of items.[5] This can be interpreted as indicating that in item retrieval - and to a lesser degree in item evaluation - comprehension plays an inferior role. This tendency appears to be more pronounced in the case of easy C-Test texts. Stemmer (1992: 126) therefore came to the following conclusion: 'If general language proficiency is meant to include high level comprehension then our results suggest that the C-Test cannot be regarded as a measure of general language proficiency.'

Similarly, Kamimoto (1992) claimed that

> frequent deletions seem to force learners to concentrate on micro-level processing ... C-Tests tend to prompt learners to activate their vocabulary competence and grammatical competence more frequently than their previous knowledge or schemata. ... In brief, what a C-Test tends to measure seems to be rather limited to discrete-point skills. (1992: 77)

2. Does C-Test solving involve primarily lower-level processing? Some further evidence

There is some evidence from studies using a different methodological approach supporting the results from our introspective studies. We will briefly comment on Grotjahn and Tönshoff (1992), Germann and Grotjahn (1994), and Grotjahn (1996). In Grotjahn and Tönshoff's (1992) study, immediately after solving a C-Test consisting of only one text, the subjects had to recall the content of the text. This procedure revealed that, although quite often the subjects were able to fill in the blanks, they had not understood the text. These results would corroborate Stemmer's findings. Germann and Grotjahn (1994) used the computer for an on-line recording

[5] The fact that more context is taken into account in evaluating items may account for Klein-Braley's (1996) finding that in a relatively high number of cases the scores on fairly distant items tended to be correlated.

of certain aspects of the subjects' overt C-Test processing. They found among other things that some subjects appeared to have tried to solve quite a number of items disregarding the context almost completely. However, when rereading the texts, the same subjects seemed to take more context into account.

These results are partly in line with the following criticism of the C-Test by Cohen, Segal and Weiss Bar-Siman-Tov (1985: 126) based on the analysis of the processing of a C-Test in modern Hebrew: 'Supplying the first part of a word allows the respondent to do local word-level guessing, without even considering either the contextual meaning or the syntactic context.'.

Finally, the issue of whether C-Test solving involves mainly low-level processing was investigated by Grotjahn (1996) in a series of controlled experiments. In these studies, the sequential structure of three French C-Test texts with 28, 52 and 56 items was disordered by scrambling the clauses. The original C-Test texts, together with two scrambled versions differing from each other in the amount of scrambling, were then administered to several samples of students of French at the University of Bochum. The hypothesis to be tested was: if the macro-context is taken into account in C-Test processing, the more scrambled a C-Test text is, the more difficult it will be to complete the mutilated words.This should at least hold for longer texts with a clear sequential structure.

It could be shown that in the case of texts longer than those normally used in C-Tests scrambling appeared to have an effect - although a weak one - on C-Test difficulty. Thus with longer texts the subjects seem to make use of the macro-context in solving the C-Test items. In the case of C-Tests of canonical length, however, there is not much macro-context which can be used in item solving. Here scrambling appeared to have no effect.

3. Final conclusion

In conclusion, we want to comment briefly on Stemmer's conclusion that the C-Test cannot be regarded as a measure of general language proficiency because higher-level comprehension processes appear to play an inferior role in item solving. We think that this conclusion may be too far-reaching. As has been found again and again in L2 reading studies, L2 readers very often tend to perform mainly low-level processing. This holds at least for not very advanced learners, and if the text is easy, even for very advanced learners (cf. Bernhardt and Kamil, 1995; Grotjahn, 1995). Thus possibly the subjects investigated in the Bochum Project were merely displaying their normal L2 reading behaviour to a large extent. Rather, we believe that, at least with longer and more difficult texts, the C-Test might well be able to

measure higher-level comprehension processes as well.[6] Thus the claim that the C-Test is a measure of general language proficiency is *not* invalidated by the studies discussed.[7] However we feel that the term "general language proficiency" needs some qualification: We suggest that with regard to the C-Test 'general language proficiency' be understood to refer primarily to the receptive component of Cummins's (1984; 1991) notion of Academic Language Proficiency (cf. also Daller, 1995 and the critical discussion of Cummins's theory in Collins, 1993, pp. 138-146). This means that the C-Test is primarily considered to measure the ability to cope receptively with context-reduced language in cognitively demanding tasks.

Bibliography

Anderson, N. J., L. Bachman, K. Perkins, and A. Cohen (1991), 'An exploratory study into the construct validity of a reading comprehension test: triangulation of data sources', *Language Testing*, 8, 41-66.

Baker, C. (1993), *Foundations of bilingual education and bilingualism*, Clevedon, Multilingual Matters.

Bernhardt, E. B. and M. L. Kamil (1995), 'Interpreting relationships between L1 and L2 reading: Consolidating the linguistic threshold and the linguistic interdependence hypotheses', *Applied Linguistics*, 16, 15-34.

Bracewell, R. J. and A. Breuleux (1994), 'Substance and romance in analyzing think-aloud protocols', in P. Smagorinsky (ed.), *Speaking about writing: Reflections on research methodology*, Thousand Oaks, CA, Sage, 55-88.

Cohen, A. D., M. Segal, and R. Weiss Bar-Siman-Tov (1985), 'The C-Test in Hebrew', *Fremdsprachen und Hochschule*, 13/14, 121-7.

Cummins, J. (1984), 'Wanted: A theoretical framework for relating language proficiency to academic achievement among bilingual students', in C. Rivera (ed.), *Language proficiency and academic achievement*, Clevedon, Multilingual Matters, 2-19.

Cummins, J. (1991), 'Conversational and academic language proficiency in bilingual contexts', *AILA Review*, 19, 75-89.

[6] Cf. also Mochizuki's 1994 study with four 120-item C-Tests, each constructed from one long text (367 to 413 words).

[7] Cf. also Klein-Braley (1995), who on the basis of a detailed analysis of the linguistic response behaviour of 200 subjects (German university freshers and English schoolchildren) to individual blanks in 30 different English C-Tests came to the following conclusion: 'The data presented here show that long-range constraints do operate, and that successful processing involves keeping these higher-level considerations in mind while simultaneously filling in blanks on the lower level ... I believe that the data presented here show clearly that C-Test performance can be validly interpreted in terms of general or overall language proficiency'.

Daller, H. (1996), 'Der C-Test als Meßinstrument alltagssprachlicher und akademischer Sprachfähigkeiten türkischer Remigranten', in R. Grotjahn (ed.), *Der C-Test: Theoretische Grundlagen und praktische Anwendungen* (vol. 3), Bochum, Brockmeyer, 343-66.

Feldmann, U. and B. Stemmer (1987), 'Thin____ aloud a___ retrospective da___ in C-te___ taking: diffe___ languages - diff___ learners - sa___ approaches?' in C. Færch and G. Kasper (eds.), *Introspection in second language research*, Clevedon,Multilingual Matters, 251-67.

Frederiksen, C. H. (1986), 'Cognitive models and discourse analysis', In C. R. Cooper and S. Greenbaum (eds.), *Studying writing: linguistic approaches*, Beverly Hills, Sage, 227-67.

Frederiksen, C. H. and B. Stemmer (1993), 'Conceptual processing of narrative discourse by a right-hemisphere brain-damaged patient', in H. Brownell and Y. Joanette (eds.), *Narrative discourse in normal aging and neurologically impaired adults*, San Diego, Singular Press, 239-78.

Germann, U. and R. Grotjahn (1994), 'Das Lösen von C-Tests auf dem Computer. Eine Pilotuntersuchung zu den Bearbeitungsprozessen', in R. Grotjahn (ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (vol. 2), Bochum, Brockmeyer, 279-304.

Grotjahn, R. (1986), 'Test validation and cognitive psychology: some methodological considerations', *Language Testing*, 3, 159-85.

Grotjahn, R. (1987), 'On the methodological basis of introspective methods', in C. Færch and G. Kasper (eds), *Introspection in second language research*, Clevedon, Multilingual Matters, 54-81.

Grotjahn, R. (1994), 'Psychologie cognitive et évaluation en langue étrangère', in Association de didactique du français langue étrangère (ASDIFLE) (ed.), *Les cahiers de l'ASDIFLE No 5: Certifications linguistiques en Europe: Problématique, instrumentations, méthodologies*. Actes des 11e et 12e Rencontres, Paris,ASDIFLE, 166-79.

Grotjahn, R. (1995), 'Zweitsprachliches Leseverstehen: Grundlagen und Probleme der Evaluation', *Die Neueren Sprachen*, 94, 533-55.

Grotjahn, R. (1996), '"Scrambled" C-Tests: Untersuchungen zum Zusammenhang zwischen Lösungsgüte und sequentieller Textstruktur', in R. Grotjahn (ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (vol. 3), Bochum, Brockmeyer, 95-125

Grotjahn, R. and B. Stemmer (1985, 'On the development and evaluation of a C-Test for French', *Fremdsprachen und Hochschule*, 13/14, 101-20.

Grotjahn, R. and W.Tönshoff (1992), 'Textverständnis bei der C-Test-Bearbeitung. Pilotstudien mit Französisch- und Italienischlernern', in R. Grotjahn (ed.), *Der C-Test. Theoretischen Grundlagen und praktischen Anwendungen* (vol. 1), Bochum, Brockmeyer, 19-95.

Kamimoto, T. (1992), 'An inquiry into what a C-Test measures', *Fukuoka Women's Junior College Studies*, 44, 67-79.

Kintsch, W. (1988), 'The role of knowledge in discourse comprehension: a construction integration model', *Psychological Review*, 95, 163-82.

Kintsch, W. and T. A. van Dijk (1978), 'Toward a model of text comprehension and production', *Psychological Review*, 85, 363-94.

Klein-Braley, C. (1996), 'Towards a theory of C-Test processing', in R. Grotjahn (ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (vol. 3), Bochum, Brockmeyer, 23-94

Köberl, J. and G. Sigott (1994), 'Word frequency, transitional probability and item facility in C-Tests', *Language Testing Update*, 16, 56-62.

Messick, S. (1989), 'Validity', in R. R. Linn (ed.), *Educational measurement* (3rd ed.), New York, American Council on Education/Macmillan,1-103.

Mochizuki, A. (1994), 'C-Tests: Four kinds of texts, their reliability and validity', *JALT Journal*, 16 (1), 41-54.

Schnotz, W. (1994), *Aufbau von Wissensstrukturen. Untersuchungen zur Kohärenzbildung beim Wissenserwerb mit Texten*, Weinheim, Beltz and Psychologie Verlags Union.

Stemmer, B. (1991), *What's on a C-Test taker's mind: Mental processes in C-Test taking*, Bochum, Brockmeyer.

Stemmer, B. (1992), 'An alternative approach to C-Test validation', in R. Grotjahn (ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (vol. 1), Bochum, Brockmeyer, 97-144.

Stemmer, B., Y. Joanette, C. Frederiksen, and N. Marchand (1994), 'Investigating discourse processing of brain-damaged individuals: The use of a stratified model of discourse processing', oral presentation at the *International Neuropsychological Society (INS)*, 16th European Conference, Angers, France, July.

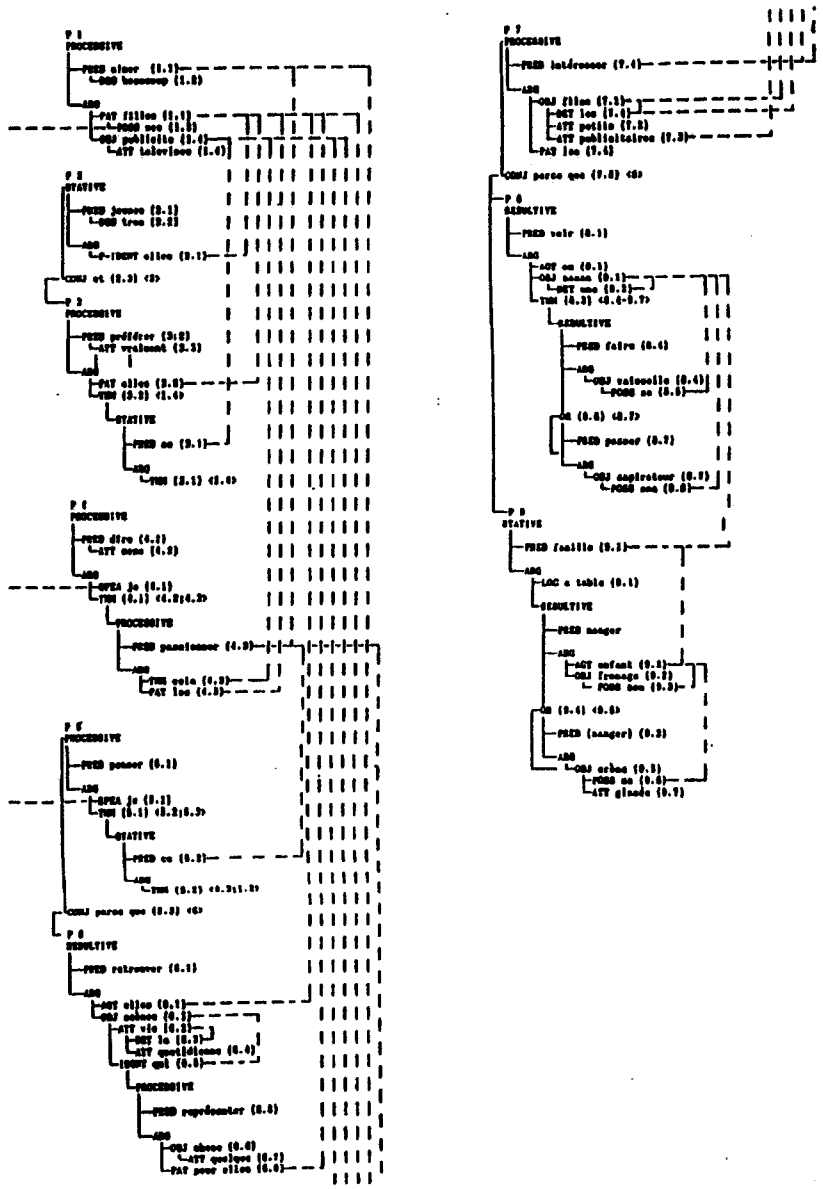
van Dijk, T. A. and W. Kintsch (1983), *Strategies of discourse comprehension*, New York, Academic Press.

APPENDIX

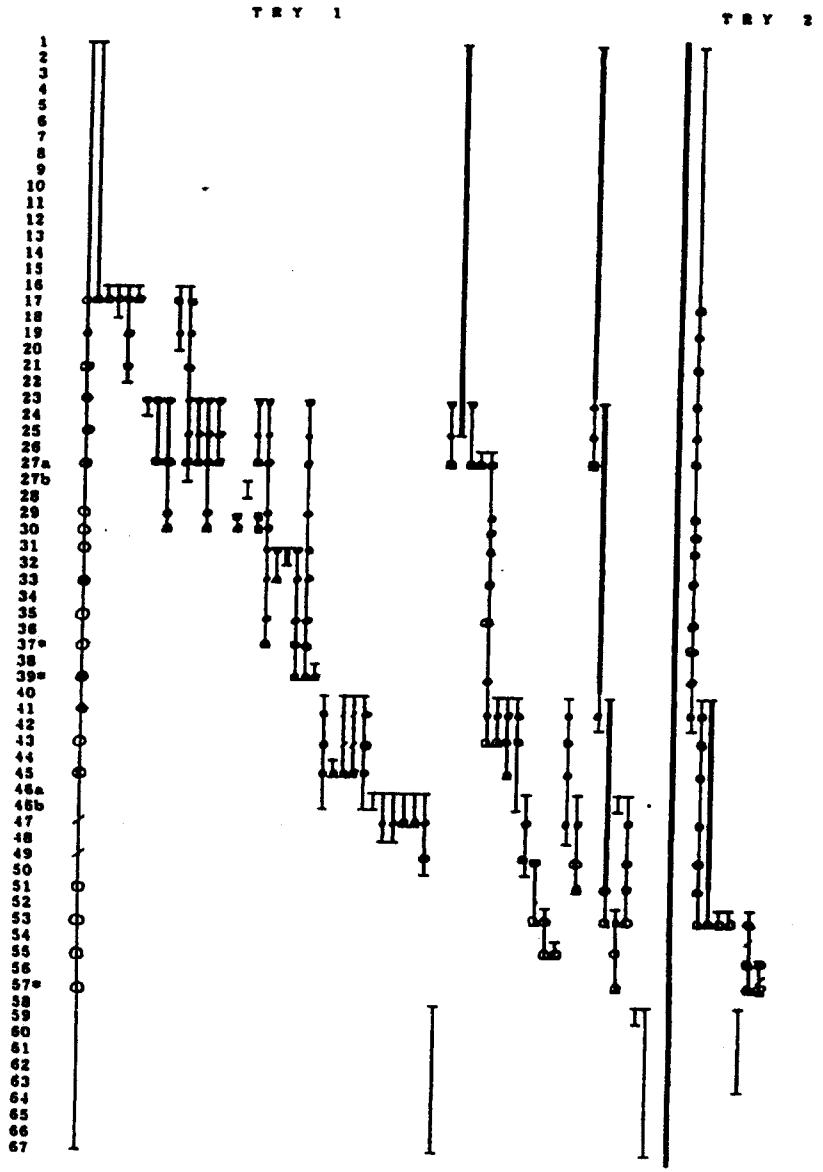
Text 1

'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 scè_____ de l_____ vie quoti_____ qui, po_____ elles, représ_____ 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.'

Propositional Graph of Text 1



Verbal Protocol Mapping Graph of Text 1 and Subject 19



Explanation of the Verbal Protocol Mapping Graph

1. Reading aloud is represented by an uninterrupted line. The beginning of each line is marked by $\top$, the end by $\perp$. One uninterrupted line with a beginning and an end mark is called a reading cycle.
2. The numbers to the left are referring to the atomic proposition number of each element.
3. Repetition of *one* single item is not marked specifically. Similarly, repetition of an item while the subject is writing it down is not taken into account.
4. Pauses are not marked.
5. Movement to a new reading cycle is performed when sequential reading is discontinued, such as: subject regresses to previous text elements and re-reads, translates etc.; or subject continues reading the text but utters word which does not sequentially follow the last word which was uttered. If *one* item was skipped during reading aloud and the subject uttered the next word after the one which had been skipped, this was not taken as a criterion to start a new reading cycle. If, however, more than one element was skipped, this was taken as an indication to start a new reading cycle.
6. Subject-item interaction during reading aloud is marked as follows: white circle = item was not completed (only item beginning was uttered); black circle = item was completed correctly; crossed circle = item was completed incorrectly.
7. Try 2, Try 3 etc. denote that the subject works through the text once again after having worked on one of the other texts.

* = item discussed in case study
|| = translation
/ = word was skipped

Eight: Psycholinguistics of C-Test taking†

Christine Klein-Braley

One of the most important questions about a language test for the psycholinguist is how far the test elicits authentic language behaviour from the test subject. Indeed, one of the main criticisms of multiple-choice language tests is that putting crosses in boxes has no relationship to real-life language use. One of the questions most frequently asked about C-Tests is whether the behaviour elicited by the mutilated texts can be genuinely viewed as representing a sample of the test takers' general language proficiency. One possible answer to this question is pragmatic: C-Tests have high intercorrelations with many other criteria which are viewed as representing real-life language use: teacher grades, other tests and examinations, pupils' self assessments, oral interviews, essay tests, reading and listening comprehension tests and so on. Therefore, if the other criteria represent real language use, so does the C-Test. For the psycholinguist, however, this statement, while true, is not satisfying. What goes on when a person takes a C-Test?

One way of looking at the behaviour shown by C-Test takers is the investigation of test-taking processes. Feldmann *et al.* (1986; see also Grotjahn, 1986) suggest three possible approaches: *statistical item analysis*, *text linguistic item analysis* and *analysis of individual performance*. The research in Duisburg conducted by Raatz and myself has primarily used the classical statistical techniques of test and item analysis. The Bochum group under the leadership of Grotjahn have concentrated their efforts on investigations of individual performance while subjects are completing C-Tests. This paper will present evidence derived from the linguistic analysis of test-taker responses. A more detailed account of the research summarised here can be found in Klein-Braley 1996 (in Grotjahn, 1996).

1. Investigation of test-taking strategies: think-aloud protocols

The researchers in Bochum (Grotjahn, 1987; Stemmer, 1991, 1992; Feldmann *et al.*, 1986) have chosen to use think-aloud procedures to investigate C-Test taking behaviour. There is no doubt that analysis of these protocols has substantially increased our understanding of what goes on inside the subject during test-taking. The methodology involves asking subjects to verbalise their thoughts while taking the tests. After the test has been completed, the tapes are then replayed to the subject who can answer any questions put by the investigator, clarify remaining uncertainties, or add comments if necessary. But there are problems in using this technique. One major drawback for anyone trained in a quantitative tradition is the *small number of subjects involved.* Feldmann *et al.* (1986) report results for a total