Summary of decisions re C-Tests
following Attrition Workshop, 15-16/12/03 Amsterdam

Test construction

C-Test consisting of 5 texts with varying degrees of formality ranging from concrete texts through more abstract to colloquial texts. The degree of formality will be established using a calculation of type-token ratio (? - other ideas?). The texts will be arranged in ascending order of difficulty starting with the easiest.

Extra care needs to be taken not to choose texts containing specialized vocabulary (particularly in formal texts) or texts on topics/from text types that subjects are unlikely to have encountered at all or very rarely in the past. This therefore excludes information included with medication and some formal letters. Instruction leaflets are a possibility but tend to exhibit highly stereotypic language making them relatively easy. However, that might make them a good starting point and would allow us to get at least some indication of how people deal with different text types. No final conclusion has been reached on this issue, particularly also with regard to what the final text might be like.

Number of gaps per individual text in test (my proposal): Although there is only a slight tendency for longer texts to be more difficult, there is an argument for keeping the texts that bit longer. The fact that on average only 8% of gaps created real problems (i.e. were solved correctly by less than 10% of test takers) translates to less than two words worthy of discussion in each text. In comparison, a text with 65 or 70 gaps would provide over 5 such potential problems. A longer text provides more opportunity for problems to occur. It also gives occasion to testing a greater range of vocabulary and structures (albeit possibly within the same semantic field) which would be of interest particularly if the C-test was to be used as the only device for testing general language proficiency. In addition, the face validity of the test may be low if it contains too little 'meat'. In order to ensure consistency with the demands this type of test poses to non-native speakers, it would seem reasonable to aim for an overall average test time of 30 minutes (5-6 per text). Texts between 52 and 78 gaps (average 65 gaps) took between 5 and 6 minutes to complete although great differences between individuals were also observed. The precise number of gaps made hardly any difference. Hence, our texts might feasibly range in their number of gaps between 50 and 70.

Test items: Since the vast majority of items were solved correctly by everybody in all circumstances, we ought to concentrate on a subset of them with a view to achieving better discrimination between the control and the experimental groups. Thus, we should identify 20 items from each text, which incidentally corresponds closely enough to the 40% of items which sometimes or always yielded problems. So we'd take the 'top 20 items' from each text (i.e. the ones most commonly creating problems) bearing in mind that in a text of 60 gaps the second 10 'most difficult items' will have been solved with greater than 60% accuracy by the control group. Based on these items we would calculate the range of error incidence in both groups. This gives us an idea of the respective proficiency levels in each cohort on a finer scale than the overall success rate. (We already have average total success rates from the pre-test with non-attriters.) These scores won't be directly comparable to scores achieved in C-tests conducted with foreign language learners but the advantage is a better degree of discrimination. While the score will be based on the 20 selected items only, all errors should however be recorded for each subject in case of attriters exhibiting problems where the control group didn't.

Since no difference was found with respect to the ease with which mutilated function words vs contents words were reconstructed their respective share won't be taken into account in the text selection.

Compounds will not be mutilated according to the classical procedure but will have the first letter of the second component left standing since they otherwise produce inordinately many errors. This change should in turn reduce the number of problematic items and should increase the overall score - another reason for keeping the texts slightly longer.

Pre-testing

In order to ensure a sufficiently good discrimination between the control and the experimental groups, a slightly lower cut-off point for allowing texts than is usually the case has been agreed: Texts which are solved with 87%-90% accuracy will be deemed acceptable, and the text chosen will be ideally from this bracket. Failing that, texts with better degrees of accuracy may be used.

Due to the remarkable consistency with which the tests were completed, a smaller number of pre-testers than originally thought will suffice (around 10).

Test administration

Timing: This wasn't discussed but if attrition manifests itself as reduced access and slower processing we could either

a) set a time limit (for 55-75 gaps up to 7 minutes would seem reasonable (average time at the higher end: 6 minutes, although we need to be aware of the great degree of inter-individual differences) or

b) ask subjects to record the time taken for solving each text.

Order and recursiveness of solving the task: Again, we did not discuss this but I think subjects should be instructed to deal with one text at a time in the order they are presented and should not be allowed to return to an earlier text. (In reality, this only happened if people had plenty of time left over.) This has been shown to improve reliability.

Instructions will be kept simple and won't include reference to the manner in which the gaps were created. For example: 'In the following texts, some parts of words have been omitted. Please reconstruct/fill in the missing parts.'

Scoring

Each correctly filled gap will receive one mark. Possible alternatives need to be established a priori at the pre-testing stage/on the basis of the native-speaker baseline data. If such a solution implies agreement changes with respect to other gaps, and these are filled accordingly, then these gaps are also deemed to be correct albeit differing from the original word.
We suggest the following classification:

0 = empty

1 = incorrect lexical stem and incorrect word class

2 = incorrect lexical stem but correct word class

3 = correct lexical stem but incorrect word class

4 = correct lexical stem, correct word class, agreement error

5 = all of above correct, but still slightly wrong

6 = acceptable variant with spelling error

7 = correct word spelling error

8 = acceptable variant

9 = correct word

If you want to use a binary right/wrong taxonomy this can then easily be reccoded by counting 6-9 as correct and 1-5 as incorrect.