MA thesis report.
Charles University.
Department of English and ELT methodology

Martin Sedláček

The Efficiency of Graded Readers for Teaching Vocabulary: A Combination of Two Approaches

This thesis focuses on the acquisition of lexical items in L2 speakers using graded readers. The student tested 16 L2 speakers of English - L1 Czech in their retention of lexical items introduced with graded readers, and compared performance of reading only vs reading and listening. Retention was tested 3 times: right after the reading session, then after one week and finally after one month. The result shows that additional listening has an overall positive effect on retention, while comparisons of retention over time are mixed, depending on the type of test used and the group chosen. The thesis is well written and the topic is scientifically interesting. There are some limitations in the structure and in the statistical analysis.

The title mentions the use of 2 approaches. In the abstract the authors mentions 2 factors, but it looks like one is part of the other: factor 1 is the use of graded readers, factor 2 is the frequency of the individual items. In addition, the abstract mentions the existence of 2 experimental groups, one that is presented with an aural version of the book, and one that reads only. It is not clear from the abstract in what the two approaches consists of. After reading the whole thesis I got the conclusion that the term "two approaches" refers to the use of *reading only* vs *reading + listening*, but this is not made clear anywhere in the text (I might in fact be wrong in my interpretation).

Introduction and Theoretical Background appear well written and complete. The student presents milestones in the scientific literature on this topic and organises his references in a structured way. It is a bit unusual to present the theoretical concepts first and the

experimental evidence separately, at the end (see section Hitherto research), but this appears like a minor problem: The information needed is there.

The methodology is a faithful report of the experiment of Waring & Takaki (2003) that is at the foundation of this thesis, and is appropriate. The presentation of the research questions, instead, was a bit confusing to me. The student reports 4 research questions:

A. How many new words are learned from reading a GR and retained over time?

B. Are words that appear frequently in the text more likely to be learned than words which appear less frequently?

C. At what rate are the words forgotten (i.e., how many of the words known at a previous test time were not known later)?

D. Do different test formats yield different gain scores?

Why isn't there any reference to the effect of listening on retention. Isn't *that* the second approach mentioned in the title and the crucial research question of your thesis?

The analytical part (i.e. the results) is the most problematic part of the thesis. I will list my comments using pages as a reference. I will only focus on analyses of group and administration effects, but the same applies to analyses of frequency effects.

Table 7, page 58. It is rather odd to separate mean values from standard deviations. Normally, only one table is used, with the mean followed by the SD in parenthesis.

Page 59, the text says: "The data shows peak results in the first administration for both groups (reading group 18.88, listening group 19.25). After one week (administration 2), the reading group recognized 16.75 words on average, while the listening group recognized 18.00 on average."

The student uses these values as being meaningful for the discussion of the hypothesis, for example by using the term "peak results". However, the data is not analysed using any statistical test (here), and as such it is not appropriate to interpret the different administrations as yielding different results. In other words, despite an apparent peak in the first administration, we have no evidence that the first score is statistically higher than the second or the last one. Statistically speaking, the 3 values may be the same. A similar problem occurs a few lines below:

"Participants of the original study by Waring and Takaki (2003) recognized on average 15.3 items in the immediate tests, 11.1 after one week and 8.4 after three months. Their study shows lower recognition rates in all sessions, as well as gradual decline (unlike the results of the present thesis, which show decrease in in administration 2, and increase in administration 3)."

The student might be right in saying that the two studies show different patterns, but without statistical analysis this claim is not grounded (he might be right, he might be wrong, we don't know for sure).

This problem arises also with the multiple choice recognition test (page 61) and with the meaning test (page 63).

In page 65 the student addresses a rather interesting issue, by saying that a crucial aspect of his result is the variation in standard deviation across different conditions. The student notices that standard deviation increases over time - at the first administration participants are clustered together and they gradually spread in their performance. I got a bit confused by the structure at this point: this section has the title "statistical significance", so I expected at this point some statistical analysis on the standard deviations. Instead, the student dismisses the analysis of the SD just introduced and performs some ANOVAs. The student describes the ANOVAs as:

"To measure the differences between the two respective groups, two-factorANOVA with replication was used. The two factors refer to two groups (reading and listening)."

Are reading and listening the only two factors included in the ANOVA? So, basically, these are not ANOVAs, these are t-tests. It is not clear at all what the student actually included in the analysis, since a few lines below he states:

"The word-form recognition test did not show significant difference in scores over time ($F = 0.9$, $p > .05$; visualized in Fig. 1), and we therefore failed to reject the null hypothesis."

This suggests that also different times of administration were included as a factor, and not just groups, as stated above. This finding is, incidentally, rather important for the discussion I mentioned in the beginning of this section (see page 59). If, as the student says, we fail to reject the null hypothesis, this means that, indeed, there is no "peak" in the first administration: the 3 administrations are statistically not different.

A few lines below the students presents ANOVAs for the other tests, but in this occasion he seems to be comparing groups only:

"In contrast, the word-form meaning recognition test did show significant difference ($F = 9.25$, $p < .005$; Fig. 2), and so did the meaning (translation) test ($F = 8.84$, $p < .005$; Fig. 3). We rejected the null hypothesis and accepted the alternative hypothesis, i.e. the difference between the performance of the two groups is statistically significant and combining the two forms of input leads to different results.". Again, if only groups are included in the analysis, these are t-tests and not ANOVAs (technically speaking a t-test is a kind of ANOVA, but this is not the standard way of reporting it). Degrees of Freedom are also missing in all reports.

Discussion and Conclusion. At page 76, the text says:

"The hypothesis of the present thesis was that the listening group will outperform the reading group". This is the first time the student uses the word hypothesis (we are the end of the thesis, after the discussion!), and it is the first time the student makes clear what is the crucial manipulation of his study. This statement is inconsistent with the research questions presented at the end of the methodology, while research questions and hypotheses should go hand in hand. Interestingly, at page 73, where the student is answering to the list of research question, there is a research question that does not appear in the beginning: "4.6.5 Is audio assisted reading more efficient than silent reading?". Including this question together with the others at the beginning would have saved a lot of trouble.

Questions for the defence:

1. What factors are included in your ANOVAs? How did you perform these ANOVAs [what software?]? How did you organise your data [can you bring the spreadsheets]?
2. What are the two approaches mentioned in the title?
3. If you were to test students with lower proficiency (such as B1), how would you expect them to perform?

In summary, despite some methodological limitations, I believe this thesis is a good quality piece of work, well within the standards for an MA. I suggest the thesis to be accepted with a grade of **one** or **two**, depending on the performance at the defence.

Luca Cilibrasi, PhD
29 August 2018