[From: *CLAW ’96*, Universiteit Leuven]

# Issues in Conducting Empirical Evaluations of Controlled Languages[1]

Heather Holmback, Boeing Information & Support Services
Serena Shubert, University of Washington
Jan Spyridakis, University of Washington

## Introduction

Simplified English (SE) is one of several restricted language standards that have been developed to reduce ambiguity and provide greater consistency and readability in technical documents. Proponents of SE (and other controlled languages) have claimed that using a restricted English standard makes documents easier to read and understand and easier to translate accurately into other natural languages. SE was designed to be applied to both procedural and descriptive writing, but in practice it has primarily been applied to procedural technical documents. The purpose of this paper is to discuss methods for assessing the claims about SE as used in procedural documents. First, we briefly overview the methodology and results of two experiments we conducted to test SE vs. non-SE documents.[2] Next, we discuss some difficult and important issues that arose in designing and conducting the experiments and analyzing the results. Finally, we conclude with some recommendations for further empirical studies of SE.

## Overview of Empirical Studies

For the past three years The Boeing Company has sponsored research conducted with the University of Washington that tested the comprehensibility and translatability of SE vs. non-SE in airplane maintenance manual procedures. We conducted two studies, both of which used actual Boeing maintenance procedures written by maintenance manual writers at Boeing Commercial Airplane Group (BCAG) Customer Services Division. SE versions of the procedures were produced after BCAG began distributing manuals in the AECMA SE standard in 1990. Non-SE versions were produced prior to that time.

[1] The authors wish to thank James E. Hoard and Richard H. Wojcik for helpful comments on an earlier version of this paper.

[2] For more details on the background, methodology, results and discussion of the experiments, see the two papers by Shubert et al. in the references.

## Comprehension Study

The first experiment (see Shubert et al., in press) tested 130 subjects (undergraduate engineering students) who were randomly assigned to read one of four documents. The four documents consisted of two procedures, called Procedure A and Procedure B, each with an SE and a non-SE version. After reading the assigned document, each subject then took a comprehension test. The test consisted of 20 two-part questions: one part assessed comprehension and the other assessed the reader's ability to identify the location of the content in the document.  The subjects were given a time limit of 30 minutes for reading one document and taking the relevant test.

We hypothesized that subjects reading the SE versions of the procedures would score higher on the comprehension test, correctly identify more content locations, and complete the task faster than subjects reading non-SE versions. An alpha level of .05 or less was adopted for all statistical tests. The most meaningful results stem from the significant interactions: readers of SE Procedure A comprehended significantly more and correctly identified significantly more content locations than readers of non-SE Procedure A. These significant effects occurred for both native and non-native speakers of English although non-native speakers consistently benefited the most. There were no significant effects for Procedure B.

Motivated by the different results for the different procedures, we re-examined the documents. Originally we wanted to choose two procedures with similar characteristics in order to generalize our results across documents.  Although we tried to select good criteria to achieve this, upon further examination we discovered that Procedure A was in fact more complex or difficult than Procedure B (i.e. the non-SE version of Procedure A received higher difficulty ratings on some readability measures and the procedural task was more complicated). This led to our preliminary conclusion that, with relatively complex documents, the use of SE will significantly improve comprehension.

## Translation Study

The second experiment, using the same four documents as the comprehensibility experiment, was designed to test the hypothesis that SE documents are more accurately and easily translated into another natural language than non-SE documents (see Shubert et al., manuscript).  To test this hypothesis, we recruited native speakers of Spanish (n = 15), Chinese (n = 17), and Japanese (n = 6) and randomly assigned them to read and translate one of the four documents. They had three hours to translate the document into their native language.

We also recruited three native speakers of each target language to rate the translations. The raters, who were graduate students at the University of Washington, were trained in the rating task.
The translations were rated on the following variables:

- Accuracy of the translation (1-5 scale)
- Style match to the original document (1-5 scale)
- Ease of comprehension (1-5 scale)
- Number of mistranslations (major and minor)
- Number of omissions (major and minor)

The translatability results were less clear-cut than the comprehension results, but they provided valuable information nevertheless. We examined the results of the three target languages combined as well as each target language separately. An alpha level of .05 or less was adopted for all statistical tests. For the three languages combined, translations of the SE versions of the procedures produced significantly higher ratings for style match and significantly fewer minor omissions than translations of the non-SE versions. For the Spanish speakers, translations of the SE versions of the procedures resulted in significantly higher ratings on accuracy, style match and comprehensibility, and significantly fewer minor mistranslations than translations of the non-SE versions. Chinese speakers' translations revealed no significant differences, and Japanese speakers' translations could not be analyzed with inferential statistics because of small cell sizes. A possible explanation for the difference in significance between the Spanish and Chinese results is the relative linguistic similarity of Spanish to English. The benefits of SE relative to non-SE might be more directly applied in translating documents from English to a language like Spanish than in translating from a language like Chinese.

Beyond these significant differences, the data also revealed a clear pattern of better scores for translations from SE versions than from non-SE versions for all the languages on almost all of the variables. As there were no significant differences, or even non-significant patterns, between Procedure A and B, we could not draw any further conclusions about procedure differences.

## Important Issues Arising in Both Studies

In this section we discuss some interesting issues which we had to face in conducting the research described above. These issues appear to be relevant to the study of any controlled language standard.

### Materials Selection

One of the most important decisions in conducting our experiments comparing SE and non-SE concerned what materials to use to represent SE and non-SE. As there is no compliance standard for SE, there is no universal agreement as to what counts as "good," "typical" or even acceptable SE. And even if such a compliance standard existed and we tested documents conforming to it, there would still be the question of whether anyone could reasonably comply with such a standard in practice, so the results of our study would not be directly applicable to use of SE in industry (which was our main interest). Therefore we decided not to construct our documents, so as not to test our conception (or anyone else's) of ideal SE. We used documents that occurred naturally, i.e. were produced by Boeing maintenance procedure writers who, in the natural course of their job, were attempting to write procedures that conformed to SE.[3] Using naturally occurring aircraft maintenance procedures also ensured some degree of relevance to Boeing and the rest of the aerospace industry.

We also did not want to create the non-SE documents. Again we did not want to compare SE to either bad non-SE or good non-SE, but to naturally occurring non-SE. Fortunately, at Boeing we had access to the procedures that had both a pre-1990 non-SE version and a post-1990 SE version, produced in compliance with the SE standard. The main problems that arose with our decision to use these types of documents were: (1) finding procedures that were long and complex enough to test but not so complex that our subjects would not be able to understand them, (2) selecting procedures where the content was the same or almost the same between the two versions, (3) ensuring that the SE version had enough SE-related differences from the non-SE version so we could test what we intended. With our limited resources we could test only two procedures, so we had to spend some time on the selection process, which was quite time-consuming. With the help of BCAG, we found two suitable procedures. We had hoped that they were evenly matched, but as it turned out the two procedures differed as to task complexity and the non-SE versions differed on certain readability measures. While not part of our original hypothesis, this difference turned out to be important to our conclusions and will be discussed further below.

[3] Maintenance procedures in SE at Boeing are created, in part, by the use of the Boeing Simplified English Checker.

The fact that we identified significant differences and clear patterns in the data between the SE and non-SE documents in both studies is of greater importance considering that we used naturally occurring documents than if we (or an SE expert) had created the documents. The problem of how to define compliance to the SE standard remains, but at least we have empirically supported the claim that SE in practice, not just in theory, can produce more readable and, to some extent, more translatable documents. To pinpoint more specifically exactly what it is about SE that makes documents more comprehensible, one might study artificially created SE and non-SE language examples, along the lines of some psycholinguistic experiments.

Another important issue for any type of writing that is found to be more comprehensible is how conformance to it will be achieved. At Boeing, maintenance procedures are written by engineers who are not necessarily professional technical writers. The use of the Simplified English Checker helps achieve conformance to SE (see Hoard et al., 1992).

## Generalizing to the Population of Interest

We had a concern about the relevance of using university students as subjects rather than actual users of maintenance documentation, the population that Boeing and others in the aerospace industry are most interested in. We justified the validity of using our subject pool on two grounds: (1) They were engineering students who had some familiarity with procedural writing and (2) the claims about SE (and other controlled languages) are usually made with respect to the general public, so it is not unreasonable to evaluate such a controlled language using university students. A more recent study by Chervak et al. (1996) tested the comprehensibility of SE and non-SE aircraft maintenance workcards using aircraft maintenance technicians for subjects. Their results were similar to ours in that the SE versions of the more complex procedures were significantly more comprehensible than the non-SE versions. The conclusions, thus, seem to hold for the population of interest as well as university students. Unfortunately, there were very few non-native English speakers among the aircraft maintenance technicians in the Chervak et al. (1996) study, but the performance of the non-native speakers was much better when using the SE documents. It would be useful to test more non-native English speaking aircraft maintenance technicians to support these conclusions. It is also of further research interest whether there are certain groups of individuals for which the conclusions from the above studies do not apply.

While we found that SE documents improved the comprehensibility of procedures for non-native English speakers, it would be helpful to know more about the relationship between English ability and the usefulness of SE.

We did not assess English ability (beyond native vs. non-native) as a variable in our comprehension study, so we cannot draw any specific conclusions about this. Chervak et al. did give a reading ability test to all subjects and found a moderately positive correlation between reading ability and performance on the comprehension test. But again, level of English ability among the non-native subjects was not an independent variable in the study, and any interaction between English reading ability and the SE/non-SE variable was not reported. In our translatability study, we were not able to vary different degrees of English ability either, but we suspect this would make a difference. Furthermore, there may be an important difference between non-native English speakers living in the United States (or other English-speaking countries) and non-native English speakers who reside in non-English-speaking countries. This was not addressed in any of the SE studies and would be an important next step in assessing the usefulness of SE (along with some measure of literacy in English), since many of the actual users of maintenance manuals are non-native English speakers living in non-English-speaking countries.

### The Role of Time

In our comprehension study, we had a time limit and we timed the reading and test-taking tasks, but there was no time pressure or real incentive for subjects to finish as quickly as possible. We did not find any significant differences for time between the SE and non-SE documents. It would have been difficult to make time a critical factor in this study. The subjects may not have been as cooperative under time pressure. The same is true for the translation experiment, where there was a three-hour time limit, but no individual times were analyzed. In the real use of the maintenance procedures, however, time is a factor. And it appears to be at least an implication, if not an outright claim, that information written in SE can be understood more quickly (i.e. "more readable" or "comprehensible" often implies quicker understanding of content). We expected to see faster times on the comprehension test of subjects using the SE documents, and we concluded that the lack of time pressure could have been a factor in the lack of a difference. It would be useful in further studies of SE to apply some time pressure or to find a useful way to measure the time it really takes to understand a document or perform a task from SE vs. non-SE instructions.

### Explaining Document Differences

Our comprehension study found that positive effects of SE related to document differences: SE was more comprehensible than non-SE for Procedure A than for Procedure B. Upon closer inspection of the documents, it appeared that the non-SE version of Procedure A was more difficult or complex (according to general readability measures and type of task) than the non-SE version of Procedure B. We concluded that SE's ability to improve comprehension and location accuracy is greater for relatively complex or difficult documents than for relatively simple or easy documents. This seems a reasonable conclusion, but it was empirically based on only one set of documents. Chervak et al. (1996) looked more closely at complexity (defined as a combination of superficial readability measures and task complexity, as judged by a Boeing technical editor) and found that, "for the two Easy workcards there was no significant change in accuracy between Simplified English and non-Simplified English versions, but for the two Difficult workcards, Simplified English gave clearly superior accuracy." In our translatability study, we did not find significant differences between the simple and complex documents, though we did find that for Chinese translations, the differences between SE and non-SE source documents were greater for Procedure A than for Procedure B. This was not true for the Spanish translations where there was no clear pattern of differences between Procedure A and Procedure B.

The issue of how to define complexity of naturally occurring documents to use in such empirical studies remains problematic. There are different types of complexity, and it is not always easy to classify a document. Even deciding "task complexity" is difficult, as a task may be hard to do, but still be quite easy to describe, e.g., sink a basketball from mid-court. While it is fairly intuitive that the use of a controlled language such as SE would be more beneficial for something complex like a difficult aircraft maintenance procedure than for something simple like using a microwave oven, research has not yet shown just what level of complexity SE is useful for. There is also the question of whether SE is any more beneficial than good quality, professional technical writing that does not conform to the SE standard. Our studies compared naturally occurring SE and naturally occurring non-SE, not SE and the best technical writing.

## Further Issues in the Comprehension Study

### Content of the Test

We constructed our comprehension tests to focus on the SE-related differences between the SE and non-SE versions of the procedures. This was quite difficult and time-consuming, but it was necessary since we were specifically interested in the comprehension differences related to the linguistic aspects of SE. It would be interesting, however, to have a test or task that did not specifically focus on the discrete "SE-motivated" differences between two documents. We should expect that the benefits of SE would hold for the overall comprehensibility and readability of the document as a whole. Of course, testing this would require using procedures long and difficult enough and having sufficient time limits and pressure for overall comprehensibility and readability to make a measurable difference.

### Wording of the Test

We took great care not to bias the wording of the test, using a neutral term in those cases where the SE document and non-SE document used different words. While, again, this was quite difficult and time-consuming, it was very important to use a completely unbiased test.

One characteristic shared by our subjects was that they were all first-time readers of SE, so they were not accustomed to seeing one or the other term. If we had used subjects familiar with SE, this might have been a problem since they might expect the SE-sanctioned term. In some ways it was an advantage in our study that we did not sample the target population, which might have included regular users of SE, since it might have been open to the criticism that the users did better with the SE documents because they were more familiar with that type of language. Ideally, we need to test many types of users.

## Further Issues in the Translatability Study

### Measuring "Translatability"

The standard quantitative measures of the accuracy of a translation are the number of major and minor omissions and the number of major and minor mistranslations. These, however, seem to measure translation only at the word level (though understanding the context is important to get the words right) and do not really reflect how accurate or understandable the overall translation is. We wanted more accuracy measures. The goal was to establish measures of the notion that the translation "conveys the same information" with the same general tone or style as the source text. Furthermore, the quantitative measures for translation do not address the parameter of "ease of translation," which appears to be involved in the claims about SE. For this, measures are needed to reliably indicate the difficulty in translating a given text. In our translatability study, we supplemented the standard quantitative measures of accuracy with three other qualitative measures that we felt were more global indicators of quality and ease of translation, but these were not perfect, nor have they been validated in general. We have concluded that it is very difficult to measure translatability. One way to measure the relative quality of translations would be to have subjects take a comprehension test or perform a task using the translations and compare the results. Perhaps applying time pressure or closely observing the process of translation might be other ways to better get at the ease of translation. The use of machine translation systems might also produce some reliable measures, though it is not clear that what is easy for machine translation is easy for human translation, and vice versa.

### Recruiting Subjects and Raters

In our translation experiment, it was very difficult and time-consuming to recruit subjects to do the translations. We had to pay the subjects and the raters, and we had limited funds. The reality of the situation both delayed the experiment and reduced the number of subjects we could use. In the case of the Japanese translations, we did not even have enough subjects to run meaningful inferential statistics.  It is possible that the lack of significance in the patterns in our data is due to the small number of subjects per cell in the experiment. This was not a big concern for the type of pilot study we were doing, but in a future study it would be necessary to have more subjects.

Another aspect of recruiting is identifying raters who understand and can perform the task in as objective a manner as possible. To compensate for subjectivity, we held a training session for the raters, and in the end they all agreed with each other fairly well. This is an important issue, and anyone doing such a study must find a way to recruit and train raters who can approach the task in a reliable and objective manner.

### What Counts as a Good Translation

In doing a translatability study, one must be aware of certain issues in the field of translation itself. For example: Which words in a technical document should be left in the source language? Should the style and organization of the original document be preserved? Some languages do not have agreed-upon terms for English technical words and, even if they do, not everyone knows them.  People also have different subjective feelings about linguistic borrowings. Furthermore, in some cultures using simpler language is not valued by educated individuals, and this might affect the way a document is translated and how it is rated.

Many of these translation issues apply equally to SE and non-SE documents, so they probably did not greatly affect our results. But it is important to be aware of these issues in designing a study, giving instructions to the translators, training the raters, and examining the results.  In our study, we had to design measures other than the standard quantitative measures of "mistranslations" and "omissions" in part because of the general disagreement on whether or not English terms should always be translated into the target language.

## Recommendations for Future Studies

We want to stress that, while our studies showed that SE can significantly improve comprehension and can to some extent improve translation ease and quality, more empirical studies need to be done on SE and other controlled languages. While the claims made about controlled languages seem reasonable, they are nevertheless empirical claims and should continue to be tested, especially as they concern how controlled languages are actually used in practice. Even if those of us who work with and advocate controlled languages do not need to be further convinced of their usefulness, the idea of using a controlled language will be more readily embraced by others if there is more empirical evidence (experimental as well as anecdotal) to support the claims.

We want to conclude with a handful of recommendations for collecting such empirical evidence, based on our experience in conducting the studies described here on the comprehensibility and translatability of SE overall.

### 1. Systematically Vary the Level of Subjects' English Ability

Since a major (though not the only) claim about SE and other controlled Englishes is that conforming to them will improve the usability of documents for non-native speakers of English, it would be interesting to assess the relative usefulness of SE depending on the level of a person's English ability or literacy. Organizations considering adopting SE or a similar controlled language might want to know the minimum level of English necessary for users to understand typical documents written in SE and how and whether that utility diminishes with increasing English ability.

### 2. Make Time a More Critical Evaluation Factor

In our experiments, there was no evidence that documents written in SE could be understood or translated more quickly than documents written in non-SE. While there were time limits, it was not feasible for us to design our experiments so that time pressure could be realistically applied. It would be useful to design an experiment using time pressure to better test for any SE vs. non-SE differences in the speed of using a document. Any empirical evidence about speed of comprehension would be useful in assessing the advantages of using SE or other controlled languages.

### 3. Test Different Levels of Complexity and Different Types of Documents

In our comprehension study, we explained the document differences as differences in document complexity: Comprehension differences between SE and non-SE are greater with documents that are relatively more complex, where complexity is defined using general readability measures and impressions about task complexity. The study by Chervak et al. (1996) strongly supported this claim. This intuitively plausible claim could be tested further to describe some threshold of complexity where using a controlled language like SE is worthwhile for improving comprehensibility (and possibly, translatability). This would give more information on when and how a company might want to use a controlled language.

### 4. Test Ability to Perform the Procedure

Taking a comprehension test is probably not the best way to measure comprehension of a procedure. Rather, actually performing the procedure under some kind of time pressure would be a more relevant task. It is more difficult and probably more costly to set up such an experiment. But it is a more realistic experiment, since procedures are written to be performed, not quizzed.

### 5. Objectify the Rating of Translations

One purpose of our translation experiment was to further determine how the claims about translatability could be measured. Our measures were primarily qualitative. The quantitative measures we had (number of mistranslations and omissions) measured only a part of the translation quality. Our measure of "comprehensibility" of resulting translation would be more objective if we actually tested subjects using the translated procedures rather than asking the opinion of a rater. Further, the differences between the translations from SE vs. non-SE documents might be greater with greater time constraints so that the subjects would have to translate as quickly as they could. The time constraint might also give more evidence about "ease of translation," which is very hard to measure. With some effort and ingenuity, it should be possible to improve upon the measures we used or define new ones to evaluate "translatability."

### 6. Test More Translation Subjects in More Languages

A possible reason for the lack of significance in our results in the translation study is that there were not enough subjects. A larger study might reveal more empirical evidence of the usefulness of SE for translation. One of the more interesting outcomes of this study was the difference between the Spanish results and the Chinese results. Other languages should be included in translation experiments to see if linguistic similarity to English makes a consistent difference in the ease and accuracy of translation from SE texts and to identify any problem areas for translating from SE source materials into any given natural language.

## References

*AECMA Simplified English Standard* (1995). AECMA Document PSC-85-16598. Issue 1. Brussels, Belgium.

Chervak, S., Drury, C.G., and Ouellette, J.P. (1996). Field Evaluation of Simplified English for Aircraft Workcards. FAA report.

Hoard, I.E., Wojcik, R., and Holzhauser, K. (1992). An Automated Grammar and Style Checker for Writers of Simplified English. P.O. Holt and N. Williams (eds.) *Computers and Writing, State of the Art.* Oxford, England: Intellect, 278-296.

Shubert, S.K., Spyridakis, J.H., Holmback, H.K., and Coney, M.B. (in press). The Comprehensibility of Simplified English in Procedures. *Journal of Technical Writing and Communication.*

Shubert, S.K., Holmback, H.K., and Spyridakis, J.H. (manuscript). Measuring the Translatability of Simplified English in Procedures.