Controlled Language and Readability

Sharon O'Brien

Introduction

Controlled Language (CL) rules specify constraints on lexicon, grammar and style with the objective of improving text translatability, comprehensibility, readability and usability. A significant body of research exists demonstrating the positive effects CL rules can have on machine translation quality (e.g. Mitamura and Nyberg 1995; Kamprath et al 1998; Bernth 1999; Nyberg et al 2003), acceptability (Roturier 2006), and post-editing effort (O'Brien 2006). Since CL rules aim to reduce complexity and ambiguity, claims have been made that they consequently improve the readability of text (e.g., Spaggiari, Beaujard and Cannesson 2003; Reuther 2003). Little work, however, has been done on the effects of CL on readability. This paper represents an attempt to investigate the relationship in an empirical manner using both qualitative and quantitative methods.

If CL rules do indeed reduce ambiguity, then we might reasonably assume that a secondary consequence is that the readability of the text is improved. Our first hypothesis is that texts written according to CL rules will be "easier to read" than those that have been written without such controls. Our second hypothesis is that, when controlled texts are translated by an MT system, the readability and acceptability of the translated text is greater than that of uncontrolled texts.

In seeking to research these hypotheses, one of the challenges is how to define and measure *readability*. Crystal (1992: 326) defines the concept as "the ease with which written language can be read with understanding." Miller and Kintsch (1980: 335) view it more as a

property that is dependent on the reader's own *abilities*: "an interaction between a text and a reader's prose-processing capabilities." Research at IBM in the 1980s had a slightly different view again, citing "grammatical complexity, overuse of jargon, and poor uses of graphics as readability defects" (Hargis 2000: 128). One can, therefore, view readability as being primarily dependent on the properties of text, or as being a function of understanding or as being determined by the reader and his or her level of education and processing capabilities. Not having one agreed definition of readability is a problem when one wants to research the concept. For the purposes of this paper, we will primarily view readability as being the property of a text which contributes to or detracts from reading ease, as measured by eye movements. We are interested in writing practices and readability. Quoting Klare (1963), Giles and Still make an important point regarding the differences between reading and writing, i.e., a high level of readability does not necessarily imply good writing (2005: 49). CL rules guide writing practices, but do they necessarily lead to a higher level of reading ease?

Much of the key literature on readability is quite dated (e.g., Flesch 1948; Dale and Chall 1948; Gunning 1968). The aforementioned readability scholars developed formulas for the automatic measurement of text readability (e.g., Flesch-Kincaid, Dale-Chall, Fog, Fry Graph, SMOG and the Automated Readability Index). Formulas were developed using different criteria, but typically they used sentence length, word frequency and number of syllables per word as criteria (DuBay 2004), and they were built upon "criterion" passages taken from texts used in the U.S. educational system, typically referring to the American grade school level one would need in order to read a specific text. According to some scholars, this makes their general applicability questionable (Giles and Still 2005; Connatser 1999).

While these readability formulae may be able to report on the general complexity of a text from a sentence length and lexical point of view, they tell us little about how much cognitive effort is involved in reading them, understanding them and translating them (whether by machine or by a human translator). Cognitive effort during reading can perhaps better be measured using eye tracking methodologies. Many studies on reading have been carried out using eye trackers (for an extensive review, see Rayner 1998 and Radach et al. 2004) and some have factored in text difficulty (Klare, Shuford and Nichols 1957). More recently, Moravcsik & Kintsch (1995) investigated the effects on recall of texts that had been written well and re-written poorly. Eye tracking has also been used to investigate translation processes (Göpferich et al. 2008) and interaction with translation tools (O'Brien 2006, 2008). However, to the best of our knowledge, no work has been done to date specifically on controlled language, readability and translatability using eye tracking as a method.

This study builds on a preliminary investigation carried out by Cadwell (2008), who investigated the following two questions: does the concept of readability have merit in the field of controlled language and can CL rules be shown to increase readability in texts? Cadwell selected a corpus of naturally occurring texts in the field of virus-checking software which functioned as training material for support personnel in the Symantec Corporation.[1] Using the Flesch Reading Ease, one of the most widely used, tested and reliable formulas (DuBay 2004), and the Flesch- Kincaid Grade Level score as readability indicators, he isolated three passages of text that corresponded to the following readability descriptors: very challenging to read, somewhat difficult, and fairly easy. Using a software program known as a Controlled Language Checker, he then created controlled language versions of his texts by applying pre-defined rules which focused primarily on improving readability (e.g., sentences should not exceed 25 words). In addition, he edited his texts according to rules flagged by the

[1] We gratefully acknowledge the help of the Symantec team in Dublin, Ireland.

CL checker, which are traditionally described as "translatability" rules in the CL literature. Readability scores were then obtained for the CL versions of the texts and, while the scores improved, they did so only marginally.[2] This marginal improvement is interesting and may explain some of the results we obtained in our study, and so we will revisit this topic later.

Cadwell compared the readability of the pre- and post-CL versions of his texts using a questionnaire and a record of reading time approach for both an expert group of readers (from Symantec Corporation) and a non-expert group. Building on the assumption that retention of keywords in short-term memory is an indicator of readability, Cadwell asked his respondents to select keywords that had appeared in their texts from a list containing terms that had occurred in the texts and three synonyms that had not occurred. He also asked them to give their opinions on which versions of each text were easier to read. Cadwell found that a majority of his expert and non-expert readers rated the CL versions as "easier to read" and more key vocabulary was retained for the CL versions for both groups. However, when measuring the speed of reading, his results showed that the reading speed for CL texts was, on average, lower than that of non-CL texts. Interestingly, Cadwell also found that non-native speakers of English in both his expert and non-expert groups seemed to find the CL versions easier to read. Cadwell's findings are interesting. However, as he himself points out, there were a number of weaknesses in his study. For example, the reading speed was measured simply by using a stopwatch and was, therefore, prone to inaccuracies. Also, he suggests that the difference in readability between his three texts was not significant enough. On the basis of his suggestions, we decided to conduct a second test of readability using just the two texts on either end of the readability scale (i.e., fairly easy and very challenging) and using an eye tracker to measure reading activity more accurately. By removing the text in the middle of the scale, we hoped to have a more striking contrast in reading ease between the two texts. The

[2] For example, Cadwell's pre-CL "easy" text only improved from 71.1 to 71.3 on the Flesch Reading Ease scale after CL rules had been applied and the "most difficult" text went from 17.8 to 19.2 on the same scale.

use of the eye tracker also allowed us to measure readability indicators such as number and duration of fixations, which are considered to be good indicators of reading difficulty (Clifton et al. 2007). We also expanded Cadwell's recall test to include cloze test sentences and one free recall question, as well as the list of key terms. One further addition to Cadwell's study involved investigating correlations between readability of the source texts in English and their readability, acceptability and (machine) translatability in three target languages (French, Spanish and Chinese). The precise methodology used is outlined below, but first we will briefly discuss the use of eye tracking in readability studies.

Eye tracking and readability

According to Kaakinen and Hyönä (2005), the eye engages in different types of movements when reading: forward first-pass, first-pass re-reading and look-backs. Each of these movement types reflects different text processing tasks. Forward first-pass fixations are seen to reflect the initial processing of textual information; first-pass re-readings reflect "integrative processing of text information" (242) or comprehension problems; and look-backs reflect the need of the reader to restore text information in working memory. Hyönä and Nurminen (2006) used these measures to categorise the reading behaviour of adult readers into fast linear readers, slow linear readers and topic structure processors. While different types of eye movements and different reader types is an important factor in reading research, it was outside the scope of this study to factor in those parameters, but we do return to this topic in our conclusions.

According to Clifton et al. (2007: 248), the length of time a word is fixated is influenced by the ease or difficulty of accessing the meaning of the word. Typically then, eye

tracking studies of reading difficulty focus on measurements such as number of fixations and duration of fixations, where the higher the number and duration the more difficult it was (we presume) for the reader to comprehend that piece of text. We adopt these assumptions in the study here by measuring fixation duration and number for controlled language and non-controlled language versions of the same text.

Methodology

Texts

As mentioned above, we reused two of Cadwell's three texts in order to see if we could replicate his findings. The texts were selected using the Flesch-Kincaid and Flesch Reading Ease formulae, with one text being classified as "easy" (Text 1) and the other as "difficult" to read (Text 3). The texts were from the IT domain, sub-domain of virus- checking software. Their primary function was to serve as informational training material for technical support staff at Symantec Corporation. These texts, then, are reasonably specialised and we discuss the implications of prior knowledge of the domain below.

Each text was checked by a CL checker for adherence to the Symantec CL rule set. The texts were edited while paying particular attention to the "readability" type rules, but also including "machine translatability" rules. The pre-CL texts are denoted throughout as "A" texts and the post-CL texts as "B" texts. It is worth mentioning here that the changes made to the texts by implementing the CL rules are more subtle than striking due to the fact that the pre-CL texts were not badly written in the first place, since authors typically adhere to style-guide rules when authoring this type of technical documentation. The subtle changes are reflected by the limited increases in the readability scores for before and after CL rules. We could, of course, have used some very poorly written texts to start with in order to make the

differences between the pre- and post-CL texts more striking. However, this would involve constructing an artificial situation: texts written by multi-national IT companies tend not to be very poorly written (in English at least) because this would reflect negatively on their image.

CL rules are typically applied to sentences or sub-sentential segments (e.g., entries in bulleted lists). Equally, MT output is normally evaluated on the sentential level and little attention is given in either domain to *text*. However, we were not interested here in a sentential or sub-sentential analysis. While we read words and sentences, it is also claimed that there is a strong top-down effect in reading. According to Kaakinen and Hyönä (2005: 253), "the current models of eye -movement control in reading emphasize the word-level processes … and the higher level guidance has so far been left aside." Thus, instead of concentrating on a sentential analysis for reading ease, we decided to focus on the textual level (although we do revert to sentential analysis for the translation evaluation). The texts were short (192 words on average), but conformed to the need of the readability formulae used to contain at least 100 words. Although the texts are relatively short, they are naturally occurring, self-contained units with a heading and one or two paragraphs about the topic. Their shortness was also appropriate for eye tracking methodology since scrolling of text can cause difficulties in eye tracking studies.

Participants were given a warm-up task which involved reading some similar text on the screen so as to familiarise themselves with the text type and the study set-up. They were informed that the research was about the ease of readability of the texts, but they were unaware of what treatment the texts had received.

It was decided to have all participants read the CL versions (B texts) first because we did not want to confuse any reading ease effect with familiarity effects on the second reading. In retrospect, we feel that it would have been better to have mixed the approach, giving some participants A texts to read first and others B texts first. This issue concerning the order of

presentation of texts came to the fore on the second reading when participants reported that the texts were "familiar" to them (see the *Recall Results* section). Cadwell's study allowed a time lag of seven days between reading sessions, and he concluded that this was insufficient to reduce or remove the familiarity effect. Therefore, we chose to have a time lag between reading sessions of four weeks. Despite this, participants reported familiarity with the texts. This issue will be discussed more fully in the Results section.

Participants

Participants volunteered from among the post-graduate student and academic community at Dublin City University. All were native speakers of English. Nineteen participants volunteered, but data for only fourteen (twelve female and two male) were analysed. As with all eye tracking studies, a number of participants were dropped from the study for various reasons (e.g., incompatibility with the eye tracker, voluntary reports of not having read the text in a "normal" way due to the fact that they were in a research environment, and so on). Five participants had a background education in Computing Science, while the remaining nine had an education in Humanities and Social Science.

Participants read in individual reading sessions. They were seated in front of the eye tracker (Tobii 1750, 50 Hz) and, following calibration, viewed the text in a browser which had no other items displayed except for the toolbar. The font size was set at fourteen (eighteen for the Headings). The analysis of the eye tracking data was carried out in the software package Tobii Studio (version 1.2.38).

Reading motivation, recall and comprehension

DuBay states that:

> The new research would establish that, along with vocabulary and sentence structure, the reader's reading ability, prior knowledge and motivation are powerful contributors to text readability. (2004: 28)

Kaakinen et al. (2003) investigated whether giving readers a "reading perspective" (i.e., a goal) would influence their text processing behaviour. Their previous research demonstrated that a reading perspective causes the reader to process perspective-relevant information longer and to recall more of this type of information than perspective-irrelevant information. They used both eye tracking and Think-Aloud-Protocols (TAP) as methods in this investigation. In our study, we did not provide readers with a specific reading perspective since we did not want them to concentrate on specific propositions in the text, at the expense of other propositions. However, we did try to provide them with a motivation for reading and comprehending the texts by informing them that they would be tested for recall and comprehension of text contents immediately after reading. Participants were made aware that this would involve recalling key terms and filling in blanks in sentences that had occurred in the text. The objective was to motivate the participants to read the text with comprehension in mind, rather than simply reading at surface level. Participants were advised to "read normally." At the same time, they were asked to read the text only once. While this did not preclude them from re-reading parts of, or entire, sentences, they were asked not to start reading the text again from the beginning. This was done so that all participants were exposed to the texts to the same extent.

However, we know that reading comprehension is complex and that the means we have at our disposal for testing it are indirect (DuBay 2004: 29). Recall tests, in particular, do not necessarily prove that a reader has *understood* the text (Kaakinen and Hyönä 2005). Sometimes it is necessary for the reader to employ prior knowledge in order to understand the

text. In the context of this study, while the readers would have been familiar with the notion of anti-virus software, none could be considered to be specialists or even to have prior knowledge in this domain. Thus, even if they retained portions of the text, as exemplified in the recall test, this does not mean that they "understood" the text. McNamara et al (1996: 7) point out that recall tests provide only a relatively superficial index of understanding based on the text base that has been constructed by the reader. On the other hand, even if participants do not retain text portions, we cannot automatically conclude that they did not understand the text since individual differences in Working Memory Capacity (WMC) could interfere with these results (Daneman and Carpenter 1980).

Despite the methodological difficulties presented by recall tests, we decided to include them here as we wanted to see if we could replicate Cadwell's findings on recall and, following Kaakinen and Hyönä (2005: 248), we included an additional "free recall" question which tested for the "gist of the meaning" of the texts, which would give us some indication of the participants' general comprehension of the text content. Finally, readers were instructed to read silently, since reading aloud is acknowledged as influencing reading time (242).

Prior knowledge

Kintsch's (1998) construction integration (CI) theory states that the reader's background knowledge plays a crucial role in text comprehension. Thus, when there is no or only little prior knowledge, the topic of the text is not available in the reader's knowledge base (long-term working memory) and the comprehension process has to rely on the use of short-term working memory, which has a limited capacity (Kaakinen, Hyönä and Keenan 2003: 448). Moravcsik and Kintsch (1995: 233) maintain that domain knowledge almost always

facilitates text comprehension. No prior knowledge will require extra time because a text representation will have to be constructed by the reader (Kaakinen, Hyönä and Keenan 2003). In Kaakinen and Hyönä's study (2005: 243), readers were selected with "no or very little prior knowledge" of the text content because this low prior knowledge condition showed a robust perspective effect in previous studies.

The readers for our study also conformed to a "low prior knowledge" condition. We chose this condition for two reasons: the texts had already been tested on a group of readers with high prior knowledge, although using different methodologies (Cadwell 2008), and we were interested in what general differences might be observed. Also, we assumed that with low prior knowledge there would be no masking of effects due to prior knowledge on the part of the participants. Drawing on several studies, Kaakinen and Hyönä (2003) mention that comprehension processes can be automatised for high prior knowledge conditions and are therefore not available for analysis (at least when using the think-aloud method).

Readers' perceptions

After having read the texts for the second time, we asked the participants for some feedback on their perceptions regarding their own recall and the readability of the texts. Participants were asked three questions which probed their familiarity with the texts, their perceived level of recall, and their ability to identify those texts that had been edited using CL rules (see Results).

Machine translation evaluation

One of the ultimate aims of CL rules, as mentioned in the introduction, is to improve the translatability (and, especially, machine translatability) of texts. It was therefore appropriate to include some MT evaluation in this study. All four texts were machine translated into

French, Spanish and Simplified Chinese using Systran Version 5, which had been tuned specifically for the domain. One evaluator per language was then asked to rate the acceptability and readability of each TL sentence. They were also provided with a detailed error classification and were asked to carry out an error analysis on all texts. They were unaware of which texts had been controlled and which were uncontrolled. One evaluator per language is clearly limited, but only one was used, as MT evaluation was not the main focus of the study. Moreover, each evaluator was highly qualified to perform the evaluation: all evaluators either had a Ph.D. in the subject of Controlled Language and Machine Translation evaluation, or was working towards a Ph.D. qualification in this topic. All three had extensive knowledge of the domain, text type and of MT evaluation procedures.[3]

Evaluators were told they could rate the texts in any order. For acceptability, they were given a four point scale:

1. *Ideal*: Sentence is grammatically correct and all information is included.
2. *Acceptable*: Sentence is not perfect, but definitely comprehensible, and has accurate transfer of all important information.
3. *Possibly Acceptable*: May be interpretable given context/time, some information transferred correctly.
4. *Unacceptable*: Absolutely not comprehensible and/or little or no information transferred accurately.

For readability, they were also asked to use a four-point scale as follows:

[3] I would like to thank the three evaluators who kindly gave of their time.

1. *Highly readable*: The segment reads as if it were written by a native writer. It is easy to read and you had no cause to pause during reading.
2. *Readable*: The segment is relatively easy to read, but you may have had to pause slightly for processing or to jump backwards once in the sentence to re-read something.
3. *Somewhat difficult*: The segment does not read as if it were written by a native writer and you may have had to pause once or twice during reading and/or jump backwards to re-read one or two phrases/words.
4. *Very difficult*: The segment is difficult to read because its structure does not conform to what is normally expected of a grammatical sentence in the TL. You would have to re-read it to make any sense of it.

The evaluators were also given the opportunity to make any comments they felt were appropriate in the evaluation spreadsheet. Having completed the acceptability and readability scores, they were then instructed to perform a detailed error analysis for each sentence. There are many examples of error categories in use for rating MT output. We chose to use the Vilar et al. (2006) error classification since it is quite detailed (without being unworkable) and the authors have also tested its relevance against automated evaluation metrics (e.g., Bleu, Nist etc.), which is more and more the norm in MT evaluation. In this framework, there are five main classes of errors: (1) Missing Words, (2) Word Order, (3) Incorrect Words, (4) Unknown Words, and (5) Punctuation. Each class can have several sub-classes, which will not be listed here, but which are listed in the results section under "MT Evaluation Results" (see Table 9). Again, evaluators were instructed to perform the error analysis in a random order. They were also told to count errors only once and, if they were unsure about which category an error belonged to, they were advised to assign the most obvious category or to select "None of the Above."

Results

In this section, we will first present the qualitative results on recall, followed by the eye tracking results and then the MT evaluation results.

Recall Results

The recall test involved answering three questions. Question one asked participants to tick the boxes for terms that occurred in the text they had just read. Six terms were provided in a list, and the number of correctly selected terms and the number of incorrectly selected terms were recorded. Question two provided participants with two sentences that had appeared in the text with two words blanked out for each sentence, and they were asked to fill in the blanks with the exact word that had occurred in the text. Question three was a free recall question, where participants were asked a general question about the content of the text, e.g., *what is the general function of the scanner described in the text*? The objective of question three was to establish whether or not the participants had generated a basic understanding of the text content. Their answers were analysed and given a judgement of (a) Yes, (b) Yes, but vague and (c) No. "No" was recorded when the answer was so vague that one could not be sure that the participant had understood the content; "Yes, but vague" was recorded when participants gave a reasonable answer, but which was sparse on detail, and "Yes" was recorded when they gave an answer that convinced the researcher that they had actually understood the gist of the text. While we are aware that judgements for question three are subjective, we nonetheless deemed it valuable to ascertain what, if any, differences there were in comprehension for the controlled vs. uncontrolled text. Our hypothesis was that there would be positive correlations between the level of recall and the controlled versions of the texts. Table 1 gives the average figures for the number of terms correctly and incorrectly recalled from the list (out of a total

of two and four respectively) and the average number of words correctly recalled within sentences (out of a total of four) for all four texts.

Table 1. Recall statistics

Text	**Average number of terms correctly recalled from list (out of 2)**	**Average number of terms incorrectly recalled from list (out of 4)**	**Average number of words correctly recalled in sentences (out of 4)**
A1	1.79	1.21	1.00
B1	2.00	0.71	0.64
A3	1.86	2.14	2.00
B3	1.93	2.14	1.57

The figures suggest that correct recall of terms from the lists was marginally higher for controlled texts (B1, B3) than for the uncontrolled versions (A1, A3) and that for Texts 1, incorrect recall is higher for the uncontrolled version, but this figure is equal for Texts 3. Thus, the controlled versions appear to have a marginal advantage when it comes to recall from a list of terms. However, for words correctly recalled in sentence context, the opposite is true, with an average figure of 1.5 for successful recall for the uncontrolled texts versus 1.1 for the controlled versions. Thus, our expectation that the controlled texts might demonstrate a recall advantage has only been partially fulfilled, with a stronger positive indication for recall from lists versus recall in sentential context. This finding reflects Cadwell's, where the CL texts showed a marginal advantage for recall of key terms in lists. However, Cadwell did not check for recall within full sentences.

Participants were asked three questions when they had completed both reading sessions. These questions were subjective, but give some insight into the participants' perceptions of the texts, reading task and their ability to recall. The questions were:

1. On the second reading, did you find the texts familiar, somewhat familiar or very familiar?
2. How would you rate your recall the second time around: better, worse or the same?

3. Would you rate the second batch of texts as easier to read, the same as the first batch, or more difficult to read?

Tables 2, 3 and 4 provide the results of this qualitative survey.

Table 2. Participants' self-rating of familiarity after second reading session

Question 1	Unfamiliar	Somewhat familiar	Very familiar	Mixed Answer: A3 somewhat familiar; A1 unfamiliar
	0	5	6	3

Table 3. Participants' self-rating of recall after second reading session

Question 2	Better	Worse	Same	Mixed Answer: A3 better; A1 the same or worse
	5	0	7	2

Table 4. Participants' rating of text readability after second reading session

Question 3	Easier	Same	Mixed Answer: A1 more difficult; A3 the same	Mixed Answer: A3 more difficult; A1 the same
	4	4	3	3

The survey shows that participants found the texts to be, on average, somewhat or very familiar to them on the second reading. However, three participants found that there was a difference in familiarity between texts A1 and A3, with A1 being rated as unfamiliar, a somewhat surprising response since this is the text that had the highest reading ease, according to the Flesch-Kincaid scoring system. Participants also voluntarily reported that they found texts A1/B1 difficult to understand, compared with texts A3/B3. This again seems to contradict the predictive reading score by Flesch-Kincaid.

Seven participants rated their recall as "the same" for the second reading, while five said "better" and 2 had mixed views depending on the text. As we have already reported,

recall was only marginally better for the CL texts and then only so for recall from lists. We can see that the perceptions about recall are mixed, as are the actual recall success rates. We can conclude that the participants were not convinced that their recall was substantially better on the second reading and, in fact, many voluntarily reported that they were surprised at how little they could recall immediately after having read the texts. So, despite the fact that the participants reported the texts as "familiar" on the second reading, they still had limited capacity to remember exact keywords that occurred in the text. This again raises the issue, mentioned earlier, of different levels of comprehension. We will return to this in the Conclusions section.

The survey results regarding the participants' perceived readability of the second batch of texts are ambiguous, with four rating them as easier, four as the same as the first batch, and six giving contradictory mixed views depending on the text. It is clear that the uncontrolled texts did not appear to be more difficult to read than the controlled texts, from the participants' viewpoint.

Cadwell (2008: 40) also asked his participants which version of text they found "easier to read." A majority of both his expert and non-expert groups selected the CL texts in response to this question. However, our own findings suggest a much weaker link between perceived reading ease and CL texts. We will now present results from the eye tracking data that will allow us to comment further on this.

Eye Tracking Results

Total Task Time

The total task time is a combined measurement of fixation gaze time, saccadic movement and non-gaze time. It gives us an indication of how much time was spent processing each text (Table 5).

Table 5. Average total task time per text

Average Total Task Time (Seconds)	Text A1	Text B1	Text A3	Text B3
Mean	58.08	73.51	69.35	66.26
Median	53.48	69.67	71.99	67.34
Standard Deviation	03.26	02.71	01.86	00.76

Text B1, the controlled version of the "easy" text, took more time to process, on average, compared with Text A1. On the other hand, Text B3, the controlled version of the "difficult" text took less time to process than Text A3. Since the sequence in which texts B1 and B3 were presented to subjects was randomised, we cannot attribute the longer task time for B1 to acclimatisation of the task, text type or domain. We could tentatatively assume that the lower processing time for B3 could be attributed to the CL rules.

Fixation Count

The Fixation Count is the number of fixations within an Area of Interest (AOI). In this study, the area in which the text occurred on screen was defined as an AOI and the number of fixations per participant was calculated. Researchers do not always agree on what should qualify as a "fixation" when reading text. For this study, we decided to use the fixation settings used in the "Eye-to-IT" research project, i.e., forty pixels and a minimum duration of one hundred milliseconds.[4] Table 6 shows the mean and median number of fixations per text and the standard deviations.

Table 6. Fixation count for all texts

Fix. Count	Text A1	Text B1	Text A3	Text B3
Mean	173	197	210	175
Median	171	182	210	167

[4] The Eye-to-IT project sought to develop human-computer monitoring and feedback tools for the purposes of studying cognition and translation - http://cogs.nbu.bg/eye-to-it/?home

Standard Deviation	39	63	34	51

If CL texts were easier to read, we would expect the mean fixation count to be lower for the CL (B) texts when compared with the non-CL (A) texts. This is not the case for A1 to B1. However, it is true for A3 to B3, and this difference was found to be significant ($p<0.01$, $t=2.417$, and $df=13$), again suggesting that B3 may have benefited from revision by CL rules.

The number of words and characters differs in each text. When we look at fixation count as a function of characters per text (without spaces) (Table 7), we see that B3 has the lowest fixation count per character, followed by A3, again suggesting that the CL rules had a positive effect on B3 in terms of readability.

Table 7. Mean fixation count per character

Fix. Count	Text A1	Text B1	Text A3	Text B3
Mean	173	197	210	175
Characters (no spaces)	776	828	1162	1172
Mean fixation count per character	0.22	0.24	0.18	0.15

Fixation Duration

Fixation duration is the length of fixations (as per the fixation filter settings) in milliseconds within an AOI. Table 8 shows the mean fixation duration for each text.

Table 8. Mean fixation duration in milliseconds

Mean Fix. Duration (in ms)	Text A1	Text B1	Text A3	Text B3
Mean	245	237	249	235
Median	241	247	244	241
Standard Deviation	3	7	3	4

The differences between the mean and median fixation durations from A1 to B1 are slight. While the mean drops between A3 and B3, the medians are very close. In addition, there is no dramatic difference between the average fixation durations for the "easy" texts and the "difficult" texts. Therefore, while we saw some evidence of a positive effect on text B3 for fixation count, no effect is evident either from uncontrolled to controlled text or from easy to difficult texts for the fixation duration measure.

Discussion of eye tracking results

Texts A1 and B1 were rated as "easier to read" compared with texts A3 and B3, according to the readability scores generated by Cadwell (2008) (A1= 71.1 Flesch Reading Ease; B1= 71.3; A3=17.8; B3=19.2). If eye tracking data accurately reflected the readability indicators, we would expect a lower fixation count and total fixation length for texts A1 and B1 compared to A3 and B3. While text A1 meets our expectations for fixation count and total task time, B1 does not.

When comparing controlled and uncontrolled versions of the texts, we would again expect the eye tracking data to demonstrate that the controlled versions (the "B" texts) were easier to read when compared with their uncontrolled versions (the "A" texts). For Fixation Count, B3 is significantly lower than A3. However, there is little difference in Fixation Duration between the two. Nonetheless, there is some evidence that the CL rules had an impact on reading ease for B3.

Our expectations are not met for A1/B1 since it is B1 (supposedly easier to read) which has higher fixation counts and duration, suggesting that it did not benefit from the treatment of CL rules.

In general then, the eye tracking data suggest that Text B3 might have improved in readability, thanks to the application of CL rules, but the "easy" text, B1, did not. Is it

possible that CL rules have more of an impact on more complex texts? This is something we will return to in the Conclusions.

MT evaluation results

As mentioned under "Methodology", we also undertook an evaluation of the MT output of all four texts for three target languages: French, Spanish and Simplified Chinese. We wanted to find out if there were differences for the controlled and non-controlled texts in the perceived acceptability and readability of the machine translated texts. Since this evaluation was based on subjective judgments by one (nonetheless experienced) evaluator per language, we supplemented the evaluation with an error analysis using Vilar et al.'s (2006) detailed error classification. We expected that the number of errors reported in the controlled input texts would be lower than the number for the uncontrolled texts. We also expected that, on average, readability and acceptability would be higher for the controlled texts.

Readability/acceptability of machine translated texts

Figure 1 shows the acceptability and readability scores for all four texts across the three target languages. For both acceptability and readability, the highest possible score was 4 (Ideal/Highly Readable) and the lowest possible score was 1 (Unacceptable/Very Difficult to Read). The scores were applied on a sentence basis and were then averaged for each text. "FR-A" is the average score per French text for "Acceptability", while "FR-R" is the average for "Readability" and so on across the three target languages.

Figure 1. Acceptability/readability: three machine-translated target languages

@@ Insert Figure 1 here

For French, acceptability increases between A1 and B1, but decreases between A3 and B3. Readability increases between A1 and B1 and, contrary to the acceptability measurement, increases between A3 and B3. We could make a very tentative claim, then, that

the controlled language rules have contributed to a slightly higher level of acceptability/readability for the French machine translated text.

For Spanish, acceptability remains the same for A1 and B1, while it increases slightly between A3 and B3. The readability measure shows the same trend for A3/B3, but drops for A1/B1. Thus, we can see a slight, positive effective of the CL rules for the more complex text, but a neutral effect for the easy text.

For Simplified Chinese, acceptability drops between A1 and B1, but increases between A3 and B3. The same is true for readability. Similar to Spanish, we see a slight positive effect of CL rules for the complex text, but a negative effect for the easy text.

In general, acceptability/readability increases slightly for the more complex texts, but not for the easy texts. This reflects the conclusions deduced from the eye tracking data.

Error Analysis of Machine Translated Texts

Table 9 shows the number of errors reported for each text, as well as the total number of errors. “FA1” refers to text A1 translated into French; “ZHA1” refers to text A1 translated into Chinese; “SA1” refers to that text translated into Spanish, and so on. For Text 1 translated into French, we can see that there was a small reduction in the errors when CL rules were applied (22 vs. 25). However, the error count remained the same for Spanish (21) and actually increased for Chinese (20 vs. 26). For Text 3, we see a very small decrease in errors (40 vs. 39) for Spanish, Chinese stays the same (19) and the number of errors in French actually increases this time (45 vs. 51).

The proportion of errors recorded is considerably higher for the more complex text (Text 3) than for the “easy” text for both French and Spanish (25 vs. 45 for French and 21 vs. 40 for Spanish). However, no such effect is apparent for Chinese (20 vs. 19).

Table 9. Error Analysis for MT Output

Error Types/Texts	FA1	SA1	ZHA1	FB1	SB1	ZHB1	FA3	SA3	ZHA3	FB3	SB3	ZHB3
No Error	2	3	1	3	2	1	2	2	0	2	2	0
Missing Words: Content Words	0	3	0	0	2	0	4	2	0	1	0	0
Missing Words: Filler Words	2	0	1	2	0	1	1	0	2	4	2	0
Word Order: Word Level: Local Range	3	1	2	2	2	1	0	2	0	0	1	0
Word Order: Word Level: Long Range	0	0	0	0	0	0	0	0	0	0	0	0
Word Order: Phrase Level: Local Range	2	0	3	0	0	5	0	0	1	0	0	1
Word Order: Phrase Level: Long Range	0	0	4	0	0	7	0	1	5	0	1	3
Incorrect Words: Sense: Wrong Lexical Choice	3	0	2	4	1	2	7	2	3	10	5	10
Incorrect Words: Sense: Incorrect Disambiguation	5	9	1	5	7	2	6	12	0	3	12	0
Incorrect Words: Incorrect Form	2	1	0	2	5	0	5	0	0	11	1	0
Incorrect Words: Extra Words	0	2	3	1	3	2	4	12	0	6	12	0
Incorrect Words: Style	2	1	2	0	1	3	4	3	0	7	4	0
Incorrect Words: Idioms	0	0	0	0	0	0	0	0	1	0	0	1
Unknown Words: Unknown Stem	0	0	0	0	0	0	0	0	0	0	0	0
Unknown Words: Unseen Forms	0	0	0	0	0	0	0	0	0	0	0	0
Punctuation	1	0	1	1	0	1	1	4	0	1	1	0
None of the above	5	4	1	6	9	1	6	1	0	2	1	0
Total Errors	25	21	20	22	21	26	45	40	19	51	39	19

Discussion of MT Evaluation Results

The application of CL rules appears to have had only a small positive effect for the three target languages, but primarily for the more complex text. We conclude from the error analysis that the application of CL rules has not reduced the error count significantly and this may account for the limited effect we have seen in the subjective ratings of acceptability and readability. The proportion of errors is greater (for French and Spanish, but not for Chinese) for the text that was rated as more complex by the readability indicators. Contrary to the acceptability/readability results and the eye-tracking results, we do not see an effect of CL rules on the number of errors in MT output.

Conclusions and Future Work

By partially replicating Cadwell's experiments using eye tracking, we hoped to gain a greater insight into the relationship between readability indicators and reading ease, as reported in eye tracking data as well as the effect that CL rules have on readability. Our hypotheses were that (1) texts written according to CL rules will be "easier to read" than those that have been written without such controls, and (2) when controlled texts are translated by an MT system, the readability and acceptability of the translated text is greater than that of uncontrolled texts.

In Cadwell's study, a majority of his readers selected the CL texts as being easier to read. Our eye tracking data suggested that the application of CL rules did indeed increase reading ease, though only marginally and only for the text identified as "difficult" by the readability formula. Our acceptability/readability evaluation of those texts machine translated into three target languages also demonstrated a small

positive effect on subjective ratings of readability and acceptability, though again this was limited to the more complex text, while the error analysis did not provide evidence of any improvement in translation output for either text.

Given these results, we could claim that our hypotheses have been partially supported, though the effects of CL rules appear to be limited to more complex texts. Future research could investigate more fully how the readability level correlates with the level of text complexity. It would also be interesting to test correlations with other readability indicators.

The improvements in readability observed in this study appear to be quite limited in nature. This brings us back to questions of methodology. Should we have used texts that had a greater differentiation in readability, as it is defined by specific readability indicators? We could have taken this approach. However, on the assumption that most writers, especially technical communicators, seek to create readable texts in the first instance, this would have meant investigating an artificial situation, which was not appealing.

We hoped that a four week gap between readings would be sufficient to make any familiarity with the texts negligible. However, many of the participants reported that the texts were familiar to them on the second reading though this view did not correlate with improved recall. Is an even longer gap between readings necessary? Or is it the case that one cannot actually erase the memory of a text from long-term memory? These are methodological questions which require further consideration when investigating the re-reading of the same, or edited, texts.

On the topic of recall tests, DuBay says “even advanced readers cannot correctly complete more than 65% of the deleted words correctly in a simple text”? Also, McNamara et al. (1996: 7) comment that recall tests provide a relatively

superficial index of understanding based primarily on the text base that has been constructed (as opposed to the deeper understanding achieved by constructing a situation model). This brings into question their use as a means of testing comprehension, which is one of the main indicators used in readability research. Some authors (e.g., Hargis 2000: 126) argue that readability research needs to extend beyond testing comprehension into testing the usability of information provided in the text. As Connatser (1999: 284) states:

> Most audiences of technical documents *read to do*. Therefore, usability testing of a document seems much more appropriate for measuring how effectively a text conveys technical information than a formula.

The next step in testing the effectiveness of CL rules for both comprehension of the source text and of the human or machine translated target text might be to examine how well the users of that documentation can carry out text instructions. A comparison between translated and non-translated instructions might also be interesting.

In some cases our data suggest a negative effect on readability when CL rules are applied. Could CL rules have a negative effect on the reading experience? By keeping all sentences, for example, to within twenty-five words are we creating a monotonous style and losing rich contextual clues provided by different types of clause relationships (Hargis 2000)? McNamara et al. (1996:5) warn us that, while a low-knowledge reader will benefit from a fully coherent text, a high-knowledge reader will learn better from a text that stimulates “more active processing”, i.e., that requires more gap-filling inferences. Could it be the case that CL texts have a

beneficial effect for non-native readers, but a detrimental effect for native readers, especially those with high levels of prior knowledge? Again, this is a question that merits further investigation.

Hyönä and Nurminen (2006) demonstrate that there are three different kinds of adult readers: slow linear readers, fast linear readers and topic structure processors, with the linear readers (slow and fast) being found to be in the majority. The behaviour of each of these types of readers in terms of look-back in texts is significantly different. They claim that the adopted reading style has consequences for the mental representation of the text as, for example, in their study topic structure processors were better at producing post-reading summaries of a text. Future studies could also factor in text presentation, the reader type and WMC.

In summary, this study has perhaps demonstrated that we need to go beyond the simple question of "Do CL rules improve readability and translatability?" to a new dimension which takes into account text users' requirements, knowledge states, reader type, working memory capacity and text presentation.

References

Bernth, Arendse. 1999. "EasyEnglish: A Confidence Index for MT". In *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation*, Chester College, England,120-127.

Cadwell, Patrick. 2008. *Readability and Controlled Language*. Unpublished M.A. Dissertation. Dublin City University.

Clifton, Charles, Staub, A., and Rayner, K. 2007. "Eye Movements in Reading Words and Sentences." In *Eye movements: A window on mind and brain*, R. V.

Gompel, M. Fisher,W. Murray, and R. L. Hill (eds.), 341–372. Amsterdam: Elsevier.

Connatser, Bradford R. 1999. “Last rites for readability formulas in technical communication.” *Journal of Technical Writing and Communication* 29(3): 271-287.

Crystal, David. 1992. *An encyclopedic dictionary of language and languages*. Oxford, UK; Cambridge, Mass., USA: Blackwell.

Daneman, Meredyth and Carpenter, P. 1980. “Individual differences in working memory and reading.” *Journal of Verbal Learning and Verbal Behavior* 19(4): 450-466.

Dale, Edgar and Chall, J. 1948. “A formula for predicting readability: Instructions.” *Educational Research Bulletin* 27(1): 11-28.

DuBay, William H. 2004. *The principles of readability*. Costa Mesa, CA: Impact Information.

Flesch, Rudolf. 1948. “A new readability yardstick.” *Journal of Applied Psychology* 32(3): 221-233.

Giles, Timothy and Still, B. 2005. “A syntactic approach to readability.” *Journal of Technical Writing and Communication* 35(1): 47-70.

Göpferich, Susanne, Jakobsen, A.L. and Mees, I. (eds). 2008. *Looking at eyes: Eye-tracking studies of reading and translation processing*. [Copenhagen Studies in Language 36]. Frederiksberg: Samfundslitteratur.

Gunning, Robert. 1968. *The technique of clear writing*. New York: McGraw-Hill.

Hargis, Gretchen. 2000. “Readability and computer documentation.” *ACM Journal of Computer Documentation* 24(3): 122-131.

Hyönä, Jukka and Nurminen A.M. 2006. “Do adult readers know how they read? Evidence from eye movement patterns and verbal reports.” *British Journal of Psychology* 97: 31-50.

Kamprath, Christine, Adolphson, E., Mitamura, T. and Nyberg, E. 1998. “Controlled language for multilingual document production: Experience with Caterpillar Technical English.” In Mitamura et al. (eds) *Proceedings of the Second International Workshop on Controlled Language Applications – CLAW ‘98*, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, 51-61.

Kintsch, Walter, Patel, V. L., and Ericsson, K. A. 1999. “The role of long-term working memory in text comprehension.” *Psychologia: An International Journal of Psychology in the Orient* 42: 186-198.

Kintsch, Walter. 1998. *Comprehension: A paradigm for cognition*. Cambridge, England: Cambridge University Press.

Klare, George. 1963. *The measurement of readability*. Ames, Iowa: Iowa State University Press.

Klare, George R., E. H. Shuford, and W. H. Nichols. 1957. “The relationship of style difficulty, practice, and ability to efficiency of reading and retention.” *Journal of Applied Psychology* 41: 222-226.

Kaakinen, Johanna and Hyönä, J. 2005. “Perspective effects on expository text comprehension: Evidence from think-aloud protocols, eyetracking, and recall.” *Discourse Processes* 40(3): 239-257.

Kaakinen, Johann, Hyönä, J. and Keenan, J. 2003. “How prior knowledge, WMC, and relevance of information affect eye fixations in expository text.” *Journal of Experimental Psychology: Learning, Memory and Cognition* 29(3): 447-457.

McNamara, Danielle, Kintsch, E., Butler Songer, N. and Kintsch, W. 1996. "Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding from text." *Cognition and Instruction* 14(1): 1-43.

Miller, James and Kintsch, W. 1980. "Readability and recall of short prose passages: A theoretical analysis." *Journal of Experimental Psychology: Human Learning and Memory* 6(4): 335-354.

Mitamura, Teruko and Nyberg, E. 1995. "Controlled English for Knowledge-Based MT: Experience with the KANT System." In *Proceedings of the 6th International Conference on Theoretical and Methodological Issues in Machine Translation*, Leuven, Belgium, 158-172.

Moravcsik, Julia and Kintsch, W. 1995. "Writing quality, reading skills, and domain knowledge as factors in text comprehension." In *Reading and Language Processing*, John Henderson, M. Singer, and F. Ferreira (eds.), 232-246. New York, London: Psychology Press.

Nyberg, Eric, Mitamura, T. and Huijsen, W.O. 2003. "Controlled language for authoring and translation." In Somers, H. (ed), *Computers and Translation: A Translator's Guide*, Amsterdam: John Benjamins, 245-282.

O'Brien, Sharon. 2008. "Processing fuzzy matches in translation memory tools: An eye-tracking analysis." In *Looking at eyes: Eye-Tracking studies of reading and translation processing*, Susanne Göpferich, A. L. Jakobsen, and I. Mees (eds.), 79-102. Frederiksberg: Samfundslitteratur.

O'Brien, Sharon. 2006. "Eye-tracking and translation memory matches." *Perspectives: Studies in Translatology* 14(3): 185-205.

Radach, Ralph, Kennedy, A. and Rayner, K. 2004. *Eye movements and information processing during reading*. Hove: Psychology Press.

Rayner, Keith. 1998. “Eye movements in reading and information processing: 20 years of research.” In *Psychological Bulletin* 124: 372-422.

Reuther, Ursula. 2003. “Two in one - Can it work? Readability and translatability by means of Controlled Language.” In *Proceedings of the Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop (CLAW 2003)*, 15th-17th May, Dublin City University, Ireland. 124-132.

Roturier, Johann. 2006. *An investigation into the impact of controlled English rules on the comprehensibility, usefulness and acceptability of machine translated technical documentation for French and German users.* Unpublished PhD dissertation. Dublin City University.

Spaggiari, Laurent, Beaujard, F. and Cannesson, E. 2003. “A controlled language at Airbus.” In *Proceedings of the Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop (CLAW 2003)*, 15th-17th May, Dublin City University, Ireland. 151-159.

Vilar, David, Xu, J., D’Haro, L.F. and Ney, H. 2006. “Error analysis of statistical machine translation output.” In *Proceeding of LREC-2006: Fifth International Conference on Language Resources and Evaluation*. Genoa, Italy, 22-28 May. 697-702.