Two in One – Can it work?
Readability and Translatability by means of Controlled Language

Ursula Reuther
Institute of Applied Information Sciences (IAI)
Martin-Luther-Str.14
D-66111Saarbrücken
ursel@iai.uni-sb.de

Abstract

According to common understanding the definition of a Controlled Language (CL) depends on its intended use: if improvement in text quality in terms of readability, clarity and comprehensibility is the primary goal of a CL, its rules will differ from CL rules which aim foremost at the improvement of translatability in an automated translation processing environment. In this article the following questions will be addressed: what is the relationship between the two CL approaches and rules? Are there any overlaps, contradictions, or interactions? Are the two approaches compatible and, if so, can they be combined? The findings presented are based on both the outcome of a CL research project and experiences in industrial CL applications.

1 Backdrop

Traditionally, Controlled Languages (CL) fall into two major categories depending on their intended application scenario, cf. (Lehrndorfer 1996) and (Spyridakis et al. 1997):

One approach towards CL aims at the improvement of readability, comprehensibility, clarity, and consistency of text. The addressees of documentation written according to such CL guidelines are human readers. Sometimes non-native speakers, as in the case of the AECMA Simplified English, and sometimes native speakers, usually in a more general technical documentation setting. In this context a CL is used with the aim of producing high quality documentation, be it in the service domain or in the area of customer documentation. In any of these scenarios, the definition and the design of the CL are driven by text linguistic and cognitive aspects.

In the other approach towards CL the improvement of translatability, especially in the case of automated translation, is the main objective. In this CL application it is important to take into account the translation tool, since both the type of tool (Translation Memory system vs. Machine Translation system) and tool specific characteristics have an impact on a related CL, as pointed out in (Bernth&Gdaniec 2001).

A further dimension, orthogonal to the above aspects, is the way the documents produced along the lines of a CL are further processed. For both the processes of "reading" (understanding) and "translating", the processing sys-

tem might either be human or an automated system.

2 Controlled Language user scenarios

	Readability	**Translatability**
Human processing	Human reader	Human translator
Machine processing	Automated language procesing (monolingual)	Automated translation (Translation Memories and MT systems)

Table 1. Types of CL consumers

Cross-classifying the two dimensions "purpose of the CL" and "processing system" as in Table 1 results in four types of "CL consumers" or user scenarios.

In order to point out possible consequences and implications a CL might have in different user scenarios, some linguistic phenomena for German modelled within a CL are examined more closely and are judged with respect to their relevance in the different scenarios.

2.1 Lexical level

At a lexical level the following phenomena are dealt with in the CL guidelines:

- **Spelling variants** like *Lambdasonde* vs. *Lambda-Sonde* do not affect human processing, whereas during machine processing they are relevant to processing, supposing that both variants are not in the respective dictionary, which is unlikely in a CL context.
- **Morphological variants** like *Abkühlungsvorgang* vs. *Abkühlvorgang* do not necessarily hinder the human reader in understanding the text. However, the human translator might have doubts as to whether to translate the two lexical items with the same concept or not.
- **Synonym variants** like *Kältetest* vs. *Kälteprüfung* might already have some impact during the human "understanding process", i.e. depending on his background knowledge, the human reader might or might not know whether the two lexical items denote the same concept.

2.2 Formatting level

Although not belonging to the core concept of a CL, it is more and more the case that CL guidelines include rules about formatting issues. This is mainly related to specific translation software input, but also some general formatting guidelines can be identified.

- **Punctuation marks** are very sensitive with respect to all applications where linguistic processing is done automatically, e.g. *„ABC“* vs. *`ABC´*.
- **Spacing** is also a problematic area for machine processing, e.g. *Be- / Entladen* vs. *Be-/Entladen.*
- **Typographic elements** are considered relevant for machine processing rather than for human processing. E.g. different ways of representing lists.

So, it comes as no surprise that all kinds of restrictions belonging to this rather format-oriented class of CL rules are relevant for ma-

chine processing and not for human language processing.

2.3 Phrase and sentence level

Most rules of a CL go beyond a lexical level and apply to word clusters, phrases and sentences. Some of these rules are referred to in the following.

- **Ambiguous structures** affect all above-mentioned user scenarios in a negative way, i.e. both the reading and translation processes are not straight forward - in other words, this phenomenon is in any CL use case relevant.
  Example: Der Verkauf der Firma war Gewinn bringend.
  Problem: The genitive could be subject or object of the nominalisation 'Verkauf'.

- The same holds for **pronouns**, since they are referential, thus posing a problem for comprehensibility and translatability.
  Example: Den Doppelkugelhahn vom Halter abschrauben. Er kann wieder verwendet werden.
  Problem: The pronoun 'er' could refer to either of the nominal antecedents in the previous sentence.

- **Complexity** involves problems for both human and machine parsing, so readability and translatability in the context of MT are affected, whereas the human translator, once he has understood the text, should have no problems in translating the text. A TM should also have no problems in translating a complex sentence when it is already in its memory. However, the more complex a sentence, the more unlikely that an exact match will be found in the TM.
  Example: In diesem Zusammenhang müssen zur Erhöhung der Betriebssicherheit und längerfristigen Abdeckung des benötigten Leistungsbedarfs Erweiterungen und Umbauten im Bereich der Stromversorgung durchgeführt werden.
  Problem: Too much information is presented at once (by means of nominalisation).

- The same as for complexity holds for **elliptical constructions**. Human and machine parsing mechanisms have to reconstruct the missing elements, which results in readability problems or, in the case of MT systems, in failed parses.
  Example: Ist die Betriebstemperatur erreicht, erlischt die Kontrolllampe.
  Problem: Omission of the conjunction 'wenn' at the beginning of the subordinate clause leads to difficulties in 'parsing'.

- The **order of elements** within a sentence might cause some problems in the context of readability, but for all other processing mechanisms this CL rule is of no importance.
  Example: Lassen Sie helle Farbe in das Farbwerk einlaufen, um die Walzen zu justieren.
  Problem: The objective of the action should be mentioned before the action itself.

- General **stylistic** recommendations, such as the use of the passive, the use of future tense, the use of negation etc. affects the reader to a great extent due to the cognitive aspect of human text processing; the human translator and a TM system are indifferent to this kind of guidelines, since their translation process is not affected. As for the relevance concerning the translation by an MT system, system-specific characteristics influence the translation of these stylistic phenomena.
  Example: Es ist darauf zu achten, dass alle Ventile geschlossen sind.
  Problem: ‚sein'+'zu'+ infinitive and extraposition of the subordinate clause should be avoided.

To sum up, it is quite obvious that the user scenario of a CL is crucial for the content and the design of the CL itself. This implies that CLs might differ more or less depending on their user scenario. And, finally, the factors to be taken into account occur along the two dimensions of "CL purpose" and "processing system".

3 Definition of a CL

3.1 Project context

In the context of a research project, cf. (TETRIS 2002), an approach to CL for German has been developed, designed and integrated into the automated checking tool MULTILINT (described in more detail in (Reuther&Schmidt-Wigger 2000)). One of the project objectives, as far as the CL component is concerned, was to provide empirical evidence as to whether and how CL rules intended to improve readability (hereinafter R-rules) relate or even coincide with CL rules aiming at the improvement of translatability of texts (hereinafter T-rules).

In a first step, existing rules have been collected and implemented, and on the basis of usability tests they have been validated according to their usefulness with respect to text quality, readability etc. In a second step, the rules have been examined with respect to their usefulness for translation purposes in general and, in addition, for a specific TM application. The outcome of this two-step validation as well as the findings about the relation between R- and T-rules will be presented in the following paragraphs.

3.2 Collection of rules

Generally acknowledged writing rules in the domain of technical authoring as well as some general but also specific "writing principles" have been collected from the literature, from style guides etc. and have been classified. The result was a set of 70 rules in total, distributed over seven main categories (see Table 2).

Rule Category	**Number of rules (in total 70)**
Typographical rules	7
Rules avoiding ambiguity	3
Lexical rules	16
Rules avoiding ellipsis	3
Rules avoiding complex structures	17
Rules dealing with word and constituent order	10
Stylistic rules	14

Table 2. Overview of rule categories

3.3 Rules for readability

Rule Category	Obligatory rules (42 / 70)	Percentage (60%)
Typography rules	2 / 7	28%
Rules avoiding ambiguity	2 / 3	67%
Lexical rules	12 / 16	75%
Rules avoiding ellipsis	1 / 3	33%
Rules avoiding complex structures	11 / 17	65%
Rules dealing with word and constituent order	6 / 10	60%
Stylistic rules	8 / 14	56%

Table 3. Number and percentage of obligatory rules per category entering the basic rule set

Starting out from this set of rules, the rules have been judged by different experts in technical writing (proof readers, trainers, authors, etc.) according to their status for being "a must" in a basic set of rules aiming at improving readability and comprehensibility (cf. also (TETRIS 2002)). The criteria according to which the rules' importance were judged relate to commonly acknowledged findings in cognitive science and can be found as well in training literature on technical writing, or even in already existing style guides of some companies. Common to all these resources where human information processing is the most central issue, is a collection of factors which do hinder this process in one or the other way. These factors are

- complexity of information (the reader is confronted with too many information chunks at a time)
- ambiguity (the reader must select the appropriate meaning)
- ellipsis (the reader has to reconstruct information)
- (text) pragmatic issues (information must be presented in a way that corresponds to the reader's situation)

These were the main criteria which have been taken as a basis when judging if and how readability and comprehensibility is affected, i.e. all the rules which account for the above phenomena in some way were given high priority and are chosen to be included in the basic rule set.[1]

Table 3 shows the percentage and the number of rules per category which are to be considered obligatory in a CL application aiming at comprehensibility and readability. In total, 60% of the rules are considered to be obligatory, thus building a basis for a CL for German technical documentation. As for the importance of the single categories, the ranking which results from the above percentages leads to the following conclusions:

1. The most important category seems to comprise the **lexical rules**.[2] Although very specific in their scope, this kind of rule seems to have a high degree of general applicability. Furthermore, from a data processing point of view, these rules are very simple and efficient, thus representing a

[1] A further, rather secondary, criterion, but which should be mentioned as well, was the (efficient) implementability of the rules in the checking software.

[2] Note: Lexical rules in this context do not include specific terminology rules, but rather restrictions on general vocabulary.

relatively 'cheap' method of standardising language.

2. Although the total number of rules for the category dealing with **ambiguous structures** is relatively small (only 3), its share is relatively high, showing its importance for comprehensiveness.
3. The fact that **complex structures** appear in third place underpins the assumption that these structures hinder comprehensibility.
4. **Word and constituent order** have an impact on quick (human!) information processing. This is shown clearly by the percentage of 60%.
5. In some document types (tables, charts, etc.) or in some information elements (e.g. headings), **elliptical constructions** are admitted, sometimes even unavoidable. Therefore this category reached only a percentage of 33%.
6. The least important category for readability seems to include the **typography related rules**. The rules cover too many, divergent items, so that no general guidelines which could contribute to a better understanding of the text could emerge.

3.4 Validating readability rules for translatability

In a next step the total set of rules (70 rules) was judged and scored independently with respect to translational relevance (indicating high or low priority), especially for a given TM application. The scoring was carried out by professional translators who have experience with TMs and by translators working for the TM provider, i.e. people who know the system well. The rating criteria applied to each rule were reflected in a score ranging from 1 (very relevant for translation) to 3 (no relevance for translation).

It came as no surprise that the number of rules considered obligatory for the improvement of translatability (T-rules) was higher than the number of rules in the readability context (R-rules).

59 out of the overall of 70 rules are judged as useful and necessary in this special translation context. Looking closer at the two resulting rule sets, the following observations can be made:

- All R-rules (42) are subsumed by the set of T-rules (59).
- 10 out of the 17 T-rules which are not included in the R-rule set are rated with low priority, i.e. their relevance for the TM system is not that high.
- The remaining 7 T-rules (out of 17) rated with high priority, however, lead to incompatibilities between the two scenarios, i.e. for translation purposes they are considered a must, whereas for the purpose of readability they are judged as irrelevant. However, this incompatibility can be observed in one direction only. This fact confirms the assumption that rules that apply to readability also apply to translatability, but not vice versa: a high degree of translatability is not always necessary in achieving a high degree of readability.

Although emerging from a specific translation scenario, the above results do not differ too much from those which have been obtained with a more general view on the translation process.

The rules with high priority for translation and no or less relevance for readability are the following:

1. Avoid complete sentences in brackets.
2. Avoid ambiguous genitive constructions.
3. Avoid parenthesis starting with *d.h.* (corresponding to *i.e.*).
4. Avoid additional plural forms in brackets.

5. In a condition/action sentence the condition part should precede the action part.
6. Avoid passive constructions (without by-agent)
7. Avoid double negation.

In three cases (number 1, 3, and 4) the rules are important in a TM scenario, because they aim at avoiding foreseeable segmentation problems. In two cases (number 6 and 7) the rules are intended to circumvent grammatical parsing problems and in only one case (number 2) does the respective rule address a 'real' translation problem.

If we look at the whole set of T-rules (59), the ranking differs to the one for the R-rules presented above in 3.3. In the translation context the most important categories are the ones which deal with **ambiguous structures, ellipsis** and **stylistic issues** with a percentage of 100% each. The categories dealing with **complex structures** (88%) and **word and constituent order** (80%) are important as well, whereas the **lexical rules** (75%) remain with the same score as before, but are less important in relation to the other categories.

These findings correlate to some extent with the conclusions of Bernth&Gdaniec (2001), although the results can only be compared on a more abstract level, since some rules describe language specific phenomena. However, the general phenomena (or rather some instances of them) of ambiguity, ellipsis and complexity (here: sentence length) also figure in the above ranking as one of the most important categories.

3.5 Conclusions

From a design point of view, it can be stated that a CL consisting of the set of T-rules is more restrictive than a CL consisting of the set of R-rules. More restrictive means in this case more rules. In more abstract terms, one might say that the requirements which must be fulfilled for a high quality translation do not correspond exactly to those which must be fulfilled for a high quality text in terms of comprehensibility and clarity. However, most of the requirements related to good writing practices result in a positive impact on translatability, as stated as well by Harkus (1997). The reverse effect, i.e. a high degree of translatability leading to high text quality is not that obvious, as shown above.

4 Automated processing environments

4.1 Automated checking of CL rules

If checking whether CL rules are adhered to during the authoring process or not is carried out automatically by means of some checking tools, the question arises whether the automated checking process can be "tuned" towards the one or the other purpose of the CL, i.e. readability or translatability.
In any case, the design of the automated checking process depends on how the relationship between R-rules and T-rules is defined. Possible relations are listed below:

- Starting out from the hypothesis that readability and translatability imply each other reciprocally and under the assumption that there is a uniform and contradiction-free rule set serving both application scenarios, an automatic checking of whether the respective rules are adhered to or not can be carried out by one program component and by one program call without further distinguishing between the checking purposes.
- Based on the assumption that there is no one-to-one relation between R-rules and T-rules but rather an intersection of the two rule sets, a second option would be to have two different checking processes based on two different rule sets (although with a possible overlap) to be called by the user by two different program calls depending on the checking objective.

- The third option builds on the assumption made above (cf. 3.5), namely that there is only a small number of T-rules which are not included in the set of R-rules, so that, also for efficiency reasons, the checking of both objectives could be done within one checking process. The results of the two checking purposes, however, are displayed to the user by different mechanisms and in different options in the GUI. For instance, one could have first displayed the results of a readability check, indicating the text units which are problematic in the sense of readability and, as an additional feature a kind of translatability scoring which is based on the same rules as the readability check plus some translational relevant rules.

This latter approach has been judged the most adequate one by the users and has been implemented within the project.

4.2 Controlled language and the use of TM

TM systems depend heavily on the reference corpus stored in the memory of the system which is used for finding identical or rather similar text segments in comparison to segments of the new text to be translated.

The more the memory is fed with controlled input, the better the quality of the translation output. This ideal combination of CL and TM is described in more detail in (Brockmann 1997). However, this rather simplistic view is only true when you only start using a TM in the translation workflow when the input to the memory consists already of controlled structures.

Otherwise, and this is more often the case in industry, if a TM has already been used in the translation process for some time, the reference material consists of uncontrolled, varying input both in the domain of terminology as well as in the stylistic domain. If a CL is not introduced in such a scenario from the very beginning producing controlled input, the results do not necessarily lead to an improvement of the translation process, since controlled input and uncontrolled reference material do not match. As a consequence, the TM cannot augment its hit rates, by the contrary they will be lowered.

The expected enhancements can only be achieved if the TM is fed with controlled input from the beginning, in other words, the reference material should consist of controlled input exclusively.

5 Conclusions

In this article we have shown that both readability and translatability of texts in the domain of technical communication can be improved by using a CL. However, a lot of different factors come into play and have to be taken into account when the respective user scenarios for a CL have to be defined.

As far as the relation between rules for readability and rules for translatability is concerned, we have shown that the two rule sets are not too different and that readability rules are a subset of translatability rules, in other words, translatability ensures readability. The reverse statement is only true to some extent.

Finally, on the basis of the above findings, we argued in favour of a common automated processing of both readability checking and translatability checking.

6 Acknowledgements

The author would especially like to thank Claudia Fottner-Top for her contributions (on behalf of BMW Group) in the TETRIS project in the framework of which the above evaluation was carried out and for her valuable comments on a draft version of this paper. We

also would like to thank the anonymous reviewers for their helpful comments and suggestions.

References

Bernth, Arendse & Gdaniec, Claudia. 2001. *MTranslatability*, in: Machine Translation 16. p.175-218.

Brockmann, Daniel. 1997. *Controlled Language & Translation Memory Technology: a Perfect Match to Save Translation Cost*, in: TC Forum 4-97.

Harkus, Susan. 1997. *Writing For Translation*. URL: www.multilingualwebmaster.com/library

Lehrndorfer, Anne. 1996. *Kontrolliertes Deutsch*. Gunter Narr, Tübingen.

Reuther, Ursula & Schmidt-Wigger, Antje. 2000. *Designing a Multi-Purpose CL Application*, in: Proceedings of CLAW 2000.

Spyridakis, Jan et al. 1997. *Measuring the Translatability of Simplified English in Procedural Documents*, in: IEEE Transactions on Professional Communication. Vol.40, No.1, March 1997.

TETRIS Abschlussbericht. 2002. URL: www.iai.uni-sb.de/docs/TETRIS_Bericht.pdf