[*Translating and the Computer 26*, November 2004 [London: Aslib, 2004]

Assessing a set of Controlled Language rules: Can they improve the performance of commercial Machine Translation systems?

Johann Roturier
Centre for Translation and Textual Studies,
Dublin City University.
E-mail: johann.roturier2(@mail.dcu.ie

Abstract

This paper presents a preliminary evaluation of a set of Controlled Language (CL) rules selected within the framework of a research project on Machine Translation (MT). The main objective of this project is to improve the machine translatability of an English corpus. This corpus has been obtained from a global Internet security technology company, Symantec. The corpus contains XML alert notifications that are generated from a SQL database. Due to the time constraints that are associated with this new type of communication medium, MT presents itself as a prospective candidate.

1. Introduction

Symantec's localisation department is currently facing the following challenge: finding an alternative way to translate a high volume of security-related contents generated from a SQL database. The traditional translation workflow does not appear to be the most suited to this task for two reasons. First of all, the information generated by the databases is perishable and must be rapidly delivered to worldwide subscribers via alert notifications. Translations should then be obtained as quickly as possible in the desired target languages (Japanese, French and German) so that the received information is not obsolete. Besides, a significant number of updates can be sent to the subscribers, sometimes on the very day of the initial notification. The main challenge is therefore to find a solution that can provide an extremely fast turnaround. This requirement is coupled with economic factors. If the traditional translation workflow were used, the cost of translating ephemeral information would be enormous. Automating the translation process by introducing MT was therefore considered as a prospective solution. A feasibility study is currently in progress in order to identify and understand the processes that would be part of a future implementation.

This paper will describe the preliminary findings that have been made with regard to some of these processes. Section 2 will focus on significant previous initiatives in the field of MT. In sections 3 and 4, the corpus and the test suite that have been designed to test the effectiveness of a selection of CL rules will be presented. In section 5, I will report on the translation results obtained with two commercial MT systems: Logomedia Translate Pro and Systran Premium 4.0, for which User Dictionaries (UDs) have been created. These results will be used to provide preliminary answers to the following two questions:

- How significant is the effectiveness of CL rules in terms of post-editing effort?
- Which CL rules have the best impact on the MT output?

I will also comment on the two evaluation methods that were used to assess the MT output (an automatic evaluation method and a manual evaluation method). The conclusion will provide recommendations with regard to the selection of CL rules. It will also outline the future directions of this study.

2. Conceptual framework

The limitations of MT are often epitomized by its inability to fully automatically translate unrestricted input, in order to obtain an output of high quality (Bennett & Gerber 2003). Raw MT output may be suitable in some cases for gisting, but a post-editing stage is necessary when the objective is to obtain a translation of publishable quality. Of course, the challenge lies in reducing this post-editing process to a minimum so that the two goals outlined above (time and cost) are achieved. Research in the last ten years or so has indicated that the quality of the MT output can be significantly improved if writers create documents with MT in mind (Bernth & Gdaniec 2001). Previous initiatives showed that some CL rules must be applied to the source text to achieve this objective. By applying lexical, syntactic and sometimes semantic restrictions, a CL attempts to improve the clarity of the source text so as to reduce ambiguities during the automatic translation process (Kamprath et al, 1998). One of the most successful collaborations in the field of MT and CL involved Carnegie Mellon University and Caterpillar in the mid 90s, where an interlingual MT system was designed. The objective of the present project is, however, different, since it aims to use an existing commercial MT engine that will satisfy the language pair requirements. As it was pointed out earlier, the contents of the aforementioned databases do not follow any corporate writing guidelines. This current lack of controlled input is reinforced by the fact that some of the discoverers of security issues are non-native speakers of English. The database editors currently have to pre-process what they receive from various authors before the alert notifications can be generated.

At the start of this feasibility study, an exceptional opportunity existed to work on a selected corpus and to edit it by using CL rules destined to improve the performance of various MT systems. The following figure presents the differences between the two approaches:

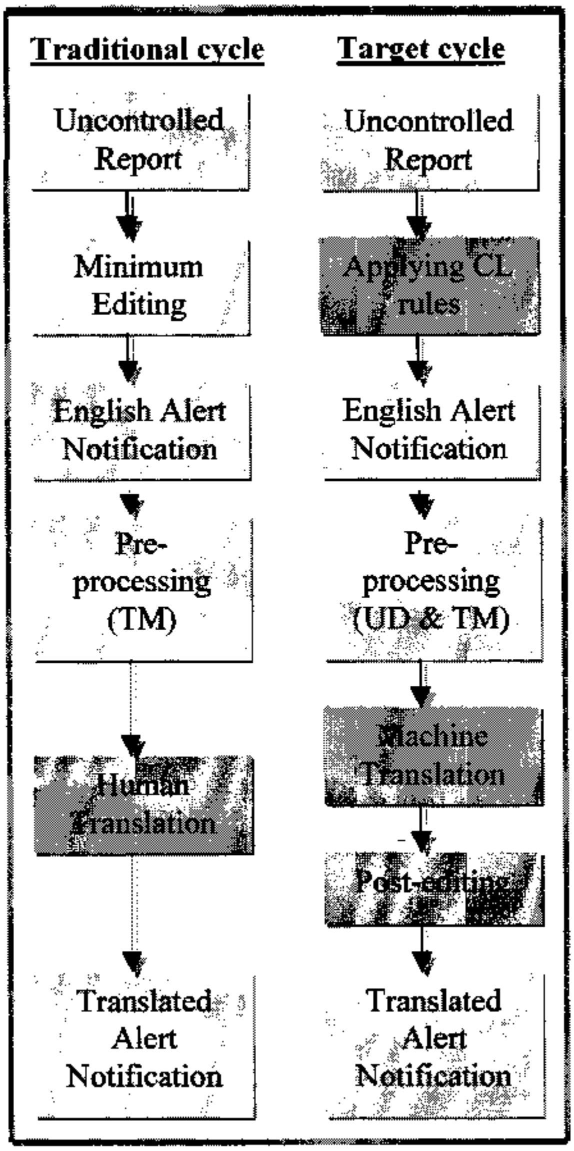


Figure 1: Current publishing cycle and traditional translation workflow vs. target publishing and translation workflow.

Even though the above diagram does not show the expected savings in terms of time (and subsequently of cost), the goal is to reduce the cost of the translation process by 2/3 of its current cost.

3. Selection of the corpus and first steps in designing CL rules

The first stage of this study was to isolate a restricted corpus that was representative of the overall contents of the databases. Since these contents refer to a very specific knowledge domain, the language that is used in the alert notifications may be regarded as a "sublanguage" (Kittredge, 1987). A sublanguage is characterised by a particular set of sentences whose usage is shared by some "community of speakers" (Ibid). In the case of security alert notifications, these speakers are the discoverers of security issues and the end-users. When it comes to the description of security vulnerabilities and threats, the lexical and syntactic patterns are often predictable. The same applies to the procedural steps that must be followed by the user in order to deal with a security issue. One of the characteristics of a sublanguage is that "it exhibits some form of closure whereby a finite set of words and grammatical constructions is used" (Pearson 1998). Since there is a lack of lexical and grammatical variation in sublanguage texts, the patterns of language would be likely to emerge more clearly from a small corpus. Consequently, a corpus of 30,000 words was gathered, based on the various types of vulnerabilities and threats that were included in the security alert notifications. Even though sublanguage texts show some form of lexical closure, their authors often use different words to refer to the same concept. Due to time constraints, editors cannot always rectify their creativity, which, as it has been suggested, sometimes goes to the point of inventing new terms for already defined concepts (Warburton, 2001). After gathering the corpus, the initial objective was to extract a set of candidate terms from the corpus. Several automatic terminology extraction tools were considered, but the results were disappointing due to the absence of bilingual reference corpora. A set of corpus linguistic tools (Wordsmith Tools 4.0[1]) was therefore used to perform some frequency tests and study the behaviour of candidate terms in context. Some lexical rules were then implemented to remove duplicates and false positives from the list of candidate terms in an effort to standardise the terminology. Even though Symantec's Editorial Styleguide contains a list of spelling and usage recommendations, this list is currently limited to the most common words used in Symantec documentation. The objective of the terminology standardisation phase was to expand this list in order to cover the terminology specific to alert notifications.

Duplicates can encompass a wide range of variants, as shown by the following examples that were found in the corpus:

- Spelling variants: *username* vs. *user name*
- Hyphenation variants: *password protected* vs. *password-protected*
- Capitalization variants: *trojan horse* vs. *Trojan horse*
- Plain form vs. symbol variants: *pipe character* vs. '|' *character.*
- Plain form vs. abbreviation variants: *Voice over IP* vs. *VoiP.*

Synonymy was also of particular interest in this study, and frequency was predominantly used as the determining factor to eliminate redundant terms. For instance, *'hostile'* (6 occurrences) and *'malevolent'* (2 occurrences) were discarded in favour of *'malicious'* (143 occurrences). The KWIC (Keyword in Context) tool

[1] http://www.lexically.net/wordsmith/

included in Wordsmith Tools 4.0 was used to make sure that all these adjectives referred to the same concept. Frequency was, however, not always the deciding criterion to select a term when syntactico-semantic features had to be taken into consideration. For instance, the term *'attack'* collocated more often with the verb *'to execute'* than *'to carry out'*. Yet, the transitive verb *'to execute'* had already been selected to collocate with inanimate artefacts such as *'programs'*. *'To carry out'* was therefore chosen to conform to the AECMA Simplified English (SE) rule: one word, one meaning (Farrington 1996). By using this method, a final list of terms was obtained by removing around 33% of duplicates. It is worth mentioning that this list did not try to select one part-of-speech (POS) per word, since several POS per word can be coded in UDs. The next stage involved finding equivalents in the desired target languages (Japanese, German and French) so as to create MT UDs. The list of terms was then sent to translators who were asked to bear in mind interlinguistic morpho-syntactic differences so as to respect a one-to-one part of speech equivalence. This shows that a certain control must also be applied on the target language prior to the creation of MT user dictionaries. Before creating a test suite to test the performance of an MT system, a regression cycle must be performed to make sure that the terms that have been imported into the user dictionaries do not create 'lexical noise' during the translation process (King & Falkedal 1990). Lexical noise can appear in the translation output when insufficient or conflicting linguistic features have been assigned to certain terms prior to the dictionary compilation. It must be noted that dictionary modules vary from one MT product to the next. The two dictionary modules that were used (Dictionary Browser for Logomedia Translate Pro and Dictionary Manager for Systran Premium 4.0) showed different capabilities with regard to their ability to deal with specific POS or types of words (such as neologisms) depending on the language pair. Despite the best efforts to control target terms, it is sometimes impossible to have a term properly inflected or used at all. These issues were reported to the developers so that they can hopefully be fixed in the next versions. For those reasons, it was decided to use Systran Premium 4.0 for French and German and Logomedia for Japanese. This selection was also performed according to the linguistic options offered by the MT systems. For instance, Logomedia allows the user to select different levels of formality for the Japanese output. Since this feature was not present in Systran Premium 4.0, Logomedia was chosen as the best candidate for Japanese. On the other hand, the absence of a 'preferred' option for German and French terms in the Logomedia Dictionary Browser did not appear to ensure consistent translations, hence the choice of Systran for those languages.

4. Designing a test suite to assess CL rules

4.1 Identifying the rules

The next step of this project concerned the selection of a set of linguistic rules to be applied to the corpus prior to the translation process. A recent study, (O'Brien 2003), assessed eight CL rule sets for English and showed that only one rule was common to all the CL rule sets under scrutiny. This finding tends to indicate that CL rules largely depend on the performance of the MT system that is used. As a result,

[2] One of the first CLs to be developed- its objective being the improvement of the readability of aircraft maintenance manuals.

existing CL rules or machine translatability guidelines (Bernth & Gdaniec 2001, O'Brien 2003) had to be gathered from the public domain and fine-tuned to meet the specific requirements of the source text. The following figure shows the breakdown of the CL sources that have been selected to create a test suite:

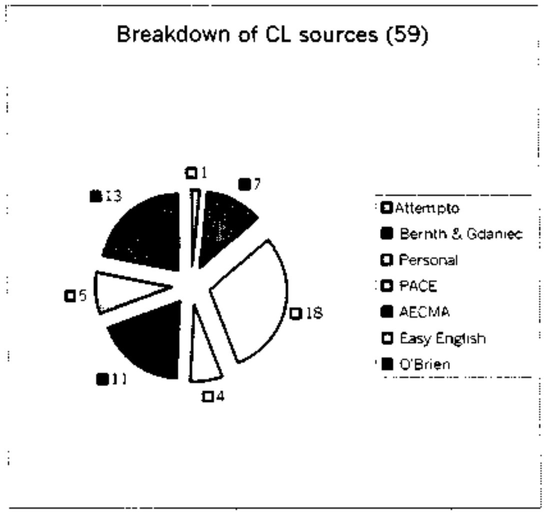


Figure 2: Breakdown of CL sources

Various sources have been used to identify CL rules from the public domain. Even though AECMA SE is often regarded as the reference source for CL rules, it was not designed with MT mind. As a result, we only selected the rules that were deemed likely to improve the machine translatability of the source contents. Besides, most of the CL rule sets have so far been limited to the automotive sector. For that reason, this study tried to find out whether a successful set of CL rules could be ported to the computer security industry by looking at new rules ('personal' category).

It should be mentioned that 17 of the rules that were selected are already present in Symantec's corporate writing guidelines. Apart from three human translation-related points, these guidelines have been designed primarily to improve the readability of the English documentation. If it emerged that these guidelines had a successful impact on machine translatability, they could easily be turned into rules. The following section will describe the method that was used to test the selected rules.

4. 2 Types of CL rules

The aforementioned study (O'Brien, 2003) divided CL rules into three categories: lexical, syntactic and textual. The test suite reflected this division with a breakdown of 8 lexical rules, 40 syntactic rules and 11 textual rules. Some of the lexical rules addressed the aforementioned problems of spelling, synonymy and morphology. These rules were tested to perform a final check on the effectiveness of the UD coding of the selected terms.

The largest category was that of syntactic rules, designed to deal with linguistic issues such as ellipses, referential phenomena, or aspectual discrepancies.

Finally, the rather high number of textual rules can be explained by the technical nature of the corpus, which at the moment does not clearly isolate non-translatable from translatable contents. As a result, several rules were designed to assess the effects of punctuation, formatting, typography, and style with regard to machine translatability.

4.3 Using test examples

The test suite consisted of two sets of examples (examples A and examples B), each containing at least one sentence. The test suite did not totally follow the uncontrolled/controlled pattern. As one of our initial objectives was to find the CL rules that could significantly improve the MT output, these rules had to be tested separately.

It appeared that the best way to test the rules was to adhere to the following procedure (adapted from King & Falkedal 1990):

- Find an example from the corpus that does not conform to the rule.
- Edit this example to make sure that it conforms to all the other rules under study (this example will be referred to as example A in the remainder of the paper).
- Reduce even further the linguistic complexity of the example to a minimum to make sure that no extra problems are introduced.
- Apply the CL rule under study to turn example A into what will be referred to as example B.
- Repeat this procedure twice so as to obtain 3 test examples A and 3 test examples B per rule.

This procedure yielded 177 examples (for a total of 205 sentences) to be machine translated for each category (A and B).

5. Evaluation of the results

5.1. Choosing evaluation methods

5.1.1 Automatic evaluation for an overall overview

Evaluation of MT quality is often regarded as a perilous exercise due to the subjectivity that is inherent to each evaluator. Several automatic translation methods have therefore appeared in the last few years: BLEU (Papineni et al, 2001) and N1ST (Doddington. 2002). These methods seem, however, more appropriate when dealing with large corpora within the framework of MT system development (Coughlin, 2003). Since the chosen test suite contained only 177 examples (for a total of 2,660 words), a simpler and quicker approach was required to get a clear overview of the effect of CL rules. A recent study (Hajič et al, 2003) presented an automatic evaluation method using the statistical matching of Trados TW 5.5's Analyze function. This method compares MT output against a reference translation of commercial quality contained in a TM. As Symantec is currently using Trados in its traditional translation workflow, this technology was readily available. However, a human translation was not used as the reference translation. It was decided to post-edit the MT output containing examples B to obtain a reference translation. The advantage of this strategy lies in the fact that human-translated reference translations could be syntactically or stylistically different from excellent MT outputs. These differences would affect Analyze's comparison process, and produce unsatisfactory results. It was therefore preferable to control the post-editing process by using strict post-editing guidelines.

5.2 Post-editing guidelines

Post-editing is also often regarded as a problematic activity depending on the level of translation quality that is expected (Allen, 2003). Due to the nature of the texts that constitute the test corpus, information accuracy prevails over stylistic considerations. For that reason, only minimum post-editing is required as long as the post-edited MT output is liable for the exactitude of the information it provides. In this light, the following guidelines were given to the post-editors (adapted from Wagner in Allen. 2003):

- Rectify what is grammatically deviant from an output of commercial quality.

- Modify what is lexically essential for the understanding of the target text (wrong or nonsensical words and phrases).

- Remember that the terminology has been imported into the MT user dictionaries. There is no need to use synonyms for the sake of originality.

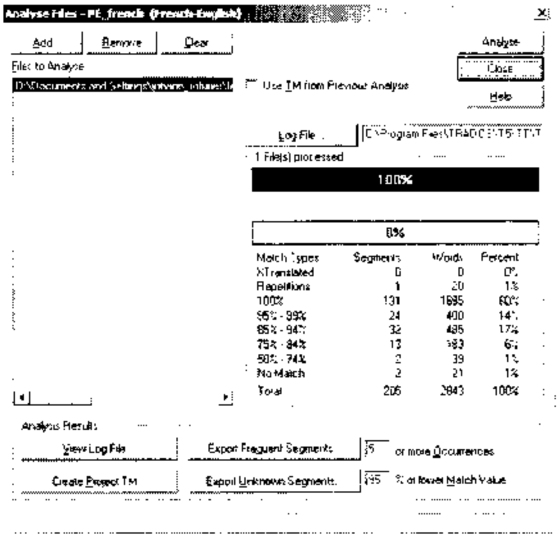
- Do not forget that all the words are probably present in the MT output (possibly in the wrong order).

- Do not forget that style does not matter (even when repetitive or pedestrian), but that information accuracy does.

- Do not spend too much time over a problem. If you cannot think straightaway of a means to improve the output, leave it unchanged (there is no point in trying a few alternatives and reverting eventually to the initial suggestion).

- Make sure that all information is accurately transferred.

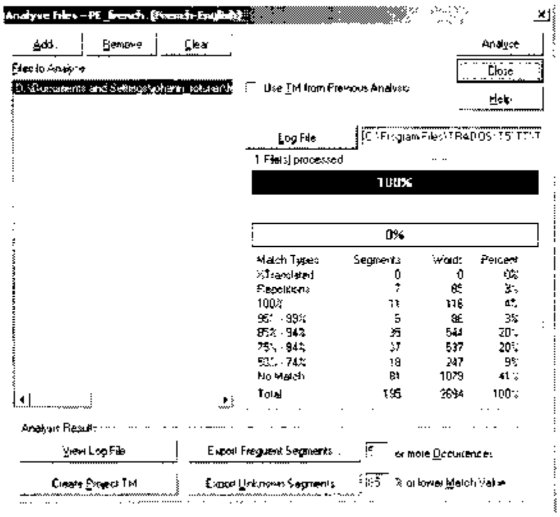
Due to time constraints, it was decided to use one post-editor per language. Once the MT output B was post-edited, its segments were aligned with the original segments to create a TM. This TM was then used to analyze the raw MT output B. The figures obtained with this type of evaluation provided an interesting overview of how close MT output B was from the post-edited one. For instance, 131 100% matches were obtained for the French MT output B, indicating that 60 % of the segments did not require any post-editing at all. Besides, high fuzzy matches (56 segments between 85% and 99%) also showed that only minor modifications were made to the MT output B.



Match Types	Segments	Words	Percent
XTranslated	0	0	0%
Repetitions	1	20	1%
100%	131	1695	60%
95% - 99%	24	400	14%
85% - 94%	32	485	17%
75% - 84%	13	183	6%
50% - 74%	2	39	1%
No Match	2	21	1%
Total	205	2843	100%

Illustration 1: Breakdown of matches for the French MT output B, using Trados 5.5's Analyse.

However, this strategy could not be used to compare the MT output A against the reference translation (post-edited output B). This was due to the issue of translation divergence that was outlined earlier, and shown by the following statistics:



Match Types	Segments	Words	Percent
XTranslated	0	0	0%
Repetitions	7	85	3%
100%	11	116	4%
95% - 99%	5	86	3%
85% - 94%	35	544	20%
75% - 84%	37	537	20%
50% - 74%	18	247	9%
No Match	81	1079	41%
Total	195	2694	100%

Illustration 2: Breakdown of matches for the French MT output A, using Trados 5.5's Analyse.

The particularly low number of 100% matches (11) does not necessarily

reflect the potential number of excellent segments present in the MT output A.
A human evaluation was therefore required to corroborate these figures and provide a more fine-grained analysis of each rule.

5.1.2 Human evaluation

A recent study (Coughlin, 2003) highlighted that the traditional dichotomy fluency/adequacy (or intelligibility/accuracy) can impede the human evaluation of the quality of an MT output. The metrics that were to be used had therefore to focus on the usability of the MT output and its influence on the subsequent post-editing process. It was decided to use the following four evaluation criteria:

***Excellent (E)**:* Read the MT output first. Then read the source text (ST). Your understanding is not improved by the reading of the ST because the MT output is satisfactory and would not need to be modified (grammatically correct/proper terminology is used/maybe not stylistically perfect but fulfils the main objective, i.e. transferring accurately all information).

***Good (G)**:* Read the MT output first. Then read the source text. Your understanding is not improved by the reading of the ST even though the MT output contains minor grammatical mistakes (word order/punctuation errors/word formation/morphology). You would not need to refer to the ST to correct these mistakes.

***Medium (M)**:* Read the MT output first. Then read the source text. Your understanding is improved by the reading of the ST, due to significant errors in the MT output (textual and syntactical coherence/ textual pragmatics/ word formation/ morphology). You would have to re-read the ST a few times to correct these errors in the MT output.

***Poor (P)**:* Read the MT output first. Then read the source text. Your understanding only derives from the reading of the ST, as you could not understand the MT output. It contained serious errors in any of the categories listed above, including wrong POS. You could only produce a translation by dismissing most of the MT output and/or re-translating from scratch.

These metrics put the emphasis on the usability of the MT output for the post-editor. Besides, they insisted on the acceptability of the output from a functional point of view rather than from a stylistic perspective. One of the main objectives of the alert notifications is to convey technical information that will be used by the end-users to remedy an issue. Style is not a high priority due to the short-lived nature of this type of text. Unlike reference materials that can be consulted over and over again, the essence of these alerts evaporates once the procedures they contain have been implemented. Besides, it is very unlikely that the end-users will engage in a thorough reading of the whole alert notification, but rather focus on the chunk of information that is the most relevant to them. To some extent, the chosen evaluation criteria may be regarded as a simplified version of existing metrics, such as the J2450 Translation Quality Metrics from the Society of Automotive Engineering (SAE). However, the chosen metrics provided the advantage of being quicker to use than the detailed SAE J2450. The time required for an MT output evaluation should not be underestimated, especially when the evaluators are not particularly familiar with a specific subject field such as that of security alert notifications. Due to time constraints and the objective of the activity (to obtain preliminary findings), it was decided that the chosen evaluators (all translation specialists) should work with basic metrics. One of the requirements in the selection of the evaluators was their availability to provide some feedback about the evaluation process. Besides the author, native evaluators were chosen for German and Japanese. Due to time constraints, it was decided to use only three evaluators per language so as to avoid obtaining conflicting opinions. The issue of evaluation consistency has often been mentioned in past reports (Coughlin, 2003; Rychtyckyj, 2002). This problem was confirmed by a certain number of internal discrepancies stemming from the results of the same evaluator. These discrepancies were corrected after getting some essential feedback from these evaluators. However, several evaluators per language pair are going to be required in the next stage of this study to corroborate the preliminary findings.

5.2 Findings

5.2.1 Overall evaluation

The objective of this study was not to compare the two MT systems with one another, but rather to assess the performance of the system that provided the best initial results for each language pair. Overall results suggest that the CL rule set had a very significant impact on the MT output quality. The following charts show the real improvement in terms of translation quality obtained with the French, German and Japanese MT output B. in all the charts, P stands for Poor, M for Medium, G for Good and E for Excellent:

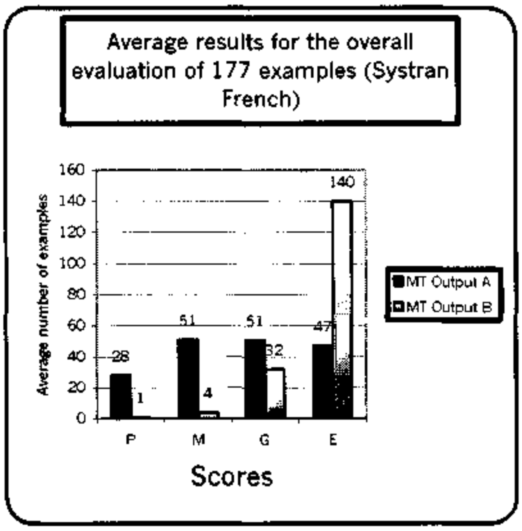


Figure 3: Average results for the overall evaluation of 177 examples (Systran French)

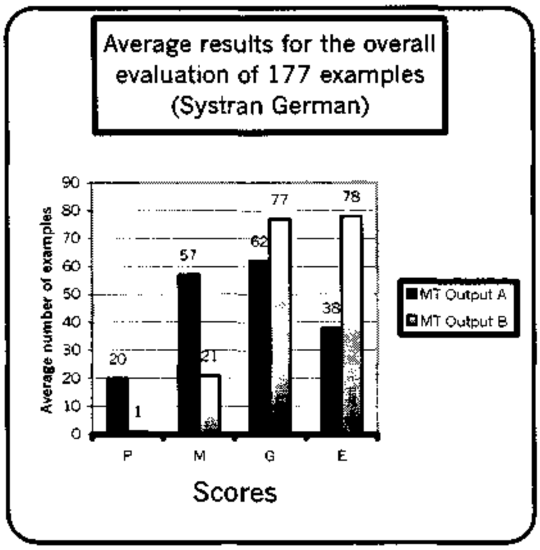


Figure 4: Average results for the overall evaluation of 177 examples (Systran German)

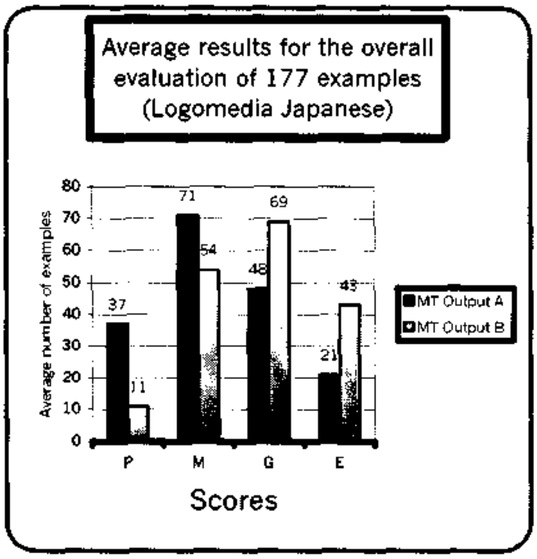


Figure 5: Average results for the overall evaluation of 177 examples (Logomedia Japanese)

These three charts clearly indicate that the CL rule set at least doubled the overall number of excellent examples in every language. However, the rules appear to have had a greater effect on French, and to a lesser extent on German, than on Japanese. The difference between the German scores and the French scores may be partly accounted by the absence of support for the German spelling reform in Systran 4.0, something that penalised several quasi-excellent sentences. Another recurrent linguistic issue concerned the handling of the adj+noun agreement rule for coded terms. Once again, this minor problem affected the score given to some examples.

The difference between the German and French scores and the Japanese scores may partly be explained by the fact that a different MT system was used. Besides, several recurrent linguistic issues appeared in the Japanese MT output B, such as the handling of modal verbs ('may' and 'will'). Another frequent issue concerned the translation of the determiner 'some' with the unusual '若干'. This last example shows that CL rules can introduce words that will be rendered curiously and redundantly in the MT output (Japanese does not need determiners). Overall, these results still suggest a clear gain in terms of post-editing effort, since in the worst scenario, excellent examples account for 25% of the overall number (43 out of 177). However, it should be noted that these excellent examples still need to be read and compared with the original during the post-editing process.

Another interesting result concerns the number of medium quality examples that almost totally disappeared for French, decreased significantly for German, and was reduced by 24% for Japanese. The risk that is associated with this type of output is that the post-editors could spend more time trying to figure out a way to fix the text than it would have taken them to translate it from scratch. This issue has been suggested in a previous study (Krings, 2001), and stresses the need to focus on CL rules that can rid MT output of these medium quality scores.

5.2.2 *Micro-level evaluation*

In order to assess the efficiency of a CL rule, the difference between the scores attributed to the three examples B and the three examples A was calculated regardless of the language pair. To do so, the aforementioned metrics were replaced by values (E by 4, G by 3, M by 2 and P by 1). The following formula indicates how the relative scores were obtained:

CL Rule Score = (sum of the scores of examples B1+B2+B3 in 3 languages) - (sum of the scores of examples A1+A2+ A3 in 3 languages)

The highest score (50) was obtained by the lexical rule ‘Only use the approved grammatical category/transitivity of a given word’, stressing the need for an extensive standard and normative terminological repository. The CL rule with the second highest score (46) indicates that ‘a particle should always follows the verb it modifies’. A micro-level evaluation showed as well that most of the 17 guidelines present in Symantec Editorial Styleguide improved the MT output significantly (with an average score of (20.75). The next step would therefore be to turn these guidelines into rules to ensure that they can be applied.

However, two of the rules that obtained very high scores in the test suite could lead to a potential clash with existing Symantec Guidelines. For instance, one of these guidelines advocates the use of parallel constructions and bulleted lists. Such a recommendation goes against the CL rule stating that ‘each segment should be able to stand alone’ (O'Brien, 2003).

Further discussions with the editors will therefore be required to check whether a compromise can be reached.

5.3 Selection of a subset of rules

After obtaining these preliminary results, the CL rules were assessed with regard to the initial comments that had been received from a team of editors. It should be stressed that their comments are invaluable to check whether some of these rules could be realistically implemented in the future. For instance, the rule 'do not use periods in proper names' conflicts with Symantec's naming conventions- the names of viruses and worms often include two or three periods, as in *Trojan.Ducky.B.* Even though this CL rule improved the MT output in all languages, it should be abandoned and replaced by a pre-processing task (proper names can be protected in Systran’s UD to prevent the engine from parsing them). Another example of a rule that was dropped concerned the use of personal pronouns. The editors had expressed some concern about this rule, due to the repetitive style that may result from its implementation. It actually turned out that the systems handled personal pronouns relatively well in most test examples, even when the pronouns did not refer to the noun they immediately followed.

After taking note of the editors' initial comments, the following CL rules have been selected to be applied on the whole corpus due to their high scores (given in brackets):

Always write a verb next to its particle. (46)

Do not use slashes to list lexical items (except for product names). (44)

Repeat the head noun with conjoined articles and prepositions. (37)

Avoid footnotes in the middle of a segment. Turn footnotes into independent segments. (37)

Do not omit words within lexical items, even when the term has already been used in the sentence. (29)

Make sure that every segment can stand alone syntactically. (27)

Only use the modal ‘could’ when the sentence contains ‘if’, otherwise use ‘can’. (26)

Be very careful with the -ing words: If it is a gerund, use an article in front of it. (14). If it is introducing a new clause, use 'by' in front of it (23). If it is modifying a noun in a non-finite clause, replace it with a relative clause. (16)

6. Conclusion

6.1. Guidelines for CL rules selection

The present study showed that editors should be consulted in the interactive process of CL rule selection, since they are the ones who will be implementing them. Even though a CL checker should be used during the authoring process, the editors may decide to discard the prescribed changes if they do not grasp the issues that may arise during the translation process.

In addition, the number of CL rules that should be chosen for a specific type of document authoring seems to be affected by a certain number of factors:

- Existing style guides
- Availability of a CL checker
- Platform used for authoring documents
- Time allocated to authoring documents
- Technical background of the editors (influence of programming languages on their authoring style)
- Linguistic knowledge of the editors
- Need for an exhaustive terminology extraction and UD customization stage
- Need for a Post-Editing process

Once all of these factors have been assessed, a subset of successful CL rules may be submitted to the editors for their approval and subsequent training. When the impact of a rule is not obvious, there seems to be no point in jeopardizing the readability of the source text (and that of the target text) by introducing redundant lexical or syntactic clues. In this light, it could be argued that the rule concerning the compounding of terms (no more than three nouns) could be replaced by the UD coding stage to deal with a term such 'Task Manager process list'. However, the terms 'Task Manager' and 'process list' may already have been present in the UD, and a simple preposition would ensure their proper usage (the process list *of* the Task Manager).

It emerges that the selection of CL rules therefore depends on the amount of time spent in the following activities:

- Authoring
- Terminology extraction
- Dictionary coding
- Post-editing
- Reviewing (if PE is outsourced)

6.2 Future directions of the study

The next step of this study will concern the application of the subset of chosen rules to the whole corpus. The objective will be to check whether the post-editing effort has sufficiently decreased so as to reduce the translation costs to 30% of the traditional method. The results so far are encouraging, since it took between 1h 30 and 2h for the translators to post-edit the MT output B (2,660 words, i.e. roughly a translator's daily output). Besides, Systran 5.0 will provide linguistic improvements (such as the implementation of the German spelling reform), which should facilitate the task of the post-editor. The test suite that was designed will be used to record these improvements. It is also hoped that the test suite could be employed to assess other MT systems in the near future.

This study will also try to isolate the persistent problematic linguistic patterns of the MT outputs. These linguistic problems may possibly be remedied with the use of macros or a checker. A batch pre-post-editing process could not only save time, but also reduce the possible frustration of post-editors when confronted with the same recurrent errors. Such an approach presents MT as a help rather than a threat. This study hopefully attempted to indicate that commercial MT systems showed great promise when used in conjunction with a set of stringent CL rules.

7. Acknowledgements

The author would like to thank Professor Jenny Williams, Dr. Minako O'Hagan and Sharon O'Brien for their valuable comments on an earlier draft of this paper.

8. References

Allen, J. 2003. “Post-Editing”. In *Computers and Translation: A translator's guide.* Somers, H. (ed). Amsterdam: John Benjamins, pp. 297-317.

Bennett, S. and L. Gerber. 2003. “Inside Commercial Machine Translation” In *Computers and Translation: A translator's guide.* Somers, H (ed). Amsterdam: John Benjamins, pp. 175-190.

Bernth, A. 1998. “Easy English: Preprocessing for MT”. *Proceedings of the Controlled Language Application Workshop (CLAW),* Language Technologies Institute, Pittsburgh, USA. pp. 30-41.

Bemth, A. and C. Gdaniec. 2001. “Mtranslatability”, in *Machine Translation 16.* pp. 175-218.

Coughlin, D. 2003. “Correlating Automated and Human Assessments of Machine Translation Quality”. *Proceedings of MT Summit X,* New Orleans, USA. pp. 63-70.

Doddington, G. 2002. “Automatic Evaluation of Machine Translation Quality Using N-Gram Co-Occurrence Statistics.” In *Proceedings of the Second International Conference on Human Language Technology.* San Diego, CA. pp. 138-145.

Farrington, G. 1996. “Simplified English: an Overview of the International Aircraft Maintenance Language”. *Proceedings of CLAW-96,* Centre for Computational Linguistics, Leuven, Belgium, pp. 1-21.

Fuchs, N.E. & R. Schwitter. 1996. “Attempto Controlled English (ACE)”. *Proceedings of CLAW-96,* Centre for Computational Linguistics, Leuven, Belgium, pp. 124-136.

Hajič, J. et al. 2003. “A simple multilingual machine translation system”. *MT Summit X,* New Orleans, USA. pp. 157-164.

Huijsen, W.O. 1998. “Controlled Language - An Introduction”. *Proceedings of the Second Controlled Language Application Workshop (CLAW),* Language Technologies Institute, Pittsburgh, USA. pp. 1-15.

Kamprath, C. et al. 1998. “Controlled Language for Multilingual Document Production: Experience with Caterpillar Technical English”. *Proceedings of the Second Controlled Language Application Workshop,* Language Technologies Institute, Pittsburgh, USA. pp. 51-61.

King, M. & K. Falkedal. 1990. “Using Test Suites in Evaluation of Machine Translation Systems”. *Proceedings of the 18th COLING Conference* vol. 2. Helsinki, Finland.

Kittredge, R. 1987. “The significance of sublanguage for machine translation”. In *Machine translation - Theoretical and methodological issues.* S. Nirenburg (ed). Cambridge: Cambridge University Press pp. 59-67.

Krings, H.P. 2001. *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes.* Koby, G.S (ed). Kent, Ohio: The Kent State University Press.

O’Brien, S. 2003. “Controlling controlled English: An Analysis of Several Controlled Language Rule Sets”. *Proceedings of EAMT-CLAW-03,* Dublin

City University, Dublin, Ireland, 15-17 May 2003. pp. 105-114.

Papineni, K., S. Roukos, T. Ward, W.-J. Zhu. 2001. "BLEU: A Method for Automatic Evaluation of Machine Translation". In *Proceedings of the 40th Annual Meeting of ACL,* Philadelphia, PA.

Pearson, J. 1998. *Terms in Context.* Amsterdam: John Benjamins.

Rychtyckyj, N. 2002 "An Assessment of Machine Translation for Vehicle Assembly Process Planning at Ford Motor Company". *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas, AMTA 2002, Tiburon, CA, USA.* pp. 207-215.

Warburton, K. 2001. "Globalization and Terminology Management." In *Handbook of Terminology Management Vol.2.* S.E Wright and G. Budhin (eds), Philadelphia: John Benjamins. pp. 677-696.