[*Translating and the Computer 27*, November 2005 [London: Aslib, 2005]

Controlled Language and the Implementation of Machine Translation for Technical Documentation

Laura RAMIREZ POLO and Johann HALLER
laura_ramirez@gmx.de
hans@iai.uni-sb.de
Universität des Saarlandes

Abstract

We aim at carrying out an empirical study to clarify if texts checked with a controlled language (CL) checker are indeed more translatable than other texts which are not compliant with the CL rule set, evaluating thus the degree of success of the application of such a restricted language with regard to its machine oriented features. For such an evaluation we adapt the FEMTI-Framework (Hovy et al., 2002) to our needs and divide our evaluation in two parts: selection of resources for the evaluation of a CL, and evaluation of the CL. In this article only the findings and results of the first part are presented. In order to simulate a real context of work, we use the system MULTILINT, a sophisticated language checker developed by the Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung e.V. an der Universität des Saarlandes (IAI). Furthermore, we use automatic translations made with a Machine Translation System.

1 Introduction

In the past few decades many efforts have been made in order to establish some guidelines for writing technical communication intended for an international audience. Due to its inherent complexity and ambiguity, natural language represents very often difficulties for both readers and translators. Controlled Languages (CL) aim at tackling this problem by restricting vocabulary and setting rules in a definite domain used to write specialized text, usually technical documentation. The application of controlled language has been a common practice in industry.

CLs were first used in the aeronautic industry to cope with the increasing complexity of technical documentation for aircraft (Farrington, 1996). CLs have since been applied in different industrial domains, such as heavy machinery (Kamprath et al., 1998). engineering (Adams et al., 1999) or automotive (Means and Godden (1996); Haller (2001); Bernardi et al. (2005); Rychtyckyj (2000) and Almqvist and Hein (1996)) among others.

It is commonly accepted that texts written according to the rules of a CL become easier to read and to understand (Nyberg et al.. 2003), since the consistency and language quality of the documentation are enhanced. This improves the efficiency and accuracy of all tasks related to the production of technical documentation. Furthermore, the formalization of a language helps to smooth the human-machine interaction in applications such as Translation Memories or Machine Translation (Controlled Translation). Much work can be saved by investing in pre-edition processes, where only the source language is affected, rather than in post-edition, where many languages have to be revised (Bernth, 1998).

All these statements are based on intuition and on some empirical studies (Mitamura and Nyberg (1995); Adams et al. (1999) and Barthe et al. (1999)), though results cannot be generally applied for all domains and languages. Differences in the structure of different languages and complexity of domains signal that CLs are not always appropriate (Janowski, 1998).

The goal of our work is to develop a method to assess if texts written according to the rules of a controlled language are more translatable than others. We do this by applying the principles of context-based evaluation and placing a hypothetical situation in an industrial context where MULTILINT is used as a CL language checker and Machine Translation (MT) comes into question as a technology. After a short theoretical introduction on MULTILINT (Section 2), the problems of evaluating controlled languages (Section 3). the principles of the FEMTI-Framework (Section 4) and an outline of evaluation metrics (Section 4.1), we present the methodology of our evaluation . This is based in two phases, the selection of resources and the evaluation of the CL. The focus of both phases is different and can be summarized as follows:

1. Selection of resources (Section 5)
   (a) Selection of the most suitable text type
   (b) Selection of the most suitable MT system
2. Evaluation:
   (a) Analysis of MULTILINT translatability features for MT

The findings and results gained from the realisation of the first phase are also presented in section 5. Finally, a conclusion and an outline of the second phase of the evaluation round off the article (Section 6).

2 Controlled German and CL Checkers

Although most controlled languages are for English, there have been attempts to define restricted rule sets for other languages. Some examples are GIFAS for French (Barthe, 1998), ScaniaSwedish for Swedish (Almqvist and Hein, 1996), or Controlled Siemens Documentary German (Schachtl, 1996). Anne Lehrndorfer (1996) deals with theoretical and methodological issues to design a controlled German. She defines syntactical and lexical guidelines, taking into account linguistic and psychological aspects of text understanding as well as the characteristics of technical documentation. However, this controlled German has never been deployed in a real work context.

In 1995, the German Federal Ministry of Economy fostered the project MULTILINT. BMW AG and the Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung e.V. an der Universität des Saarlandes (IAI) were, among others, the main partners in this project. Its goal was to develop an intelligent linguistic system for the production and administration of multilingual technical documentation (Haller, 2001). The subsequent project, TETRIS (starting in 1999 and lasting until 2002), resulted in the development of the tool MULTILINT (Figure 1), a sophisticated language checker.

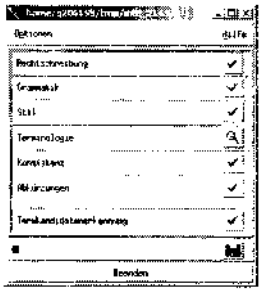

Figure 1: MULTILINT Front end

The approach of MULTILINT deviates slightly from the traditional approach and definition of a controlled language, since there is no previously defined controlled language. Rather, MULTILINT aims at "controlling" the language by helping authors to write technical documentation according to a definite set of style, spelling and grammar rules (general language correctness). These rules belong to the core of the system. The style rules represent an exception. These are given by the system, but the author or linguistic resources manager can add new rules or adapt them to the style of the company where the checker is being deployed. Besides, authors are required to use a controlled vocabulary and a controlled terminology (corporate language correctness). The latter is defined by the user (Reuther, 1998).

In 2002, MULTILINT was upgraded by CLAT (Controlled Language Authoring Tool). Though the linguistic intelligence behind MULTILINT and CLAT is the same, both systems present some differences. These include, among others, the front end, which is implemented in Java in CLAT (Figure 2), in contrast to the tcl tk implementation of MULTILINT, the interaction of the different modules, and an editor where the author can correct and test in CLAT the suggestions of the system. MULTILINT and CLAT are in use by important industrial companies in Germany, such as Heidelberger Druckmaschinen, Sun Mycrosystems Inc. (for English) and BMW AG (IAI, 2005).

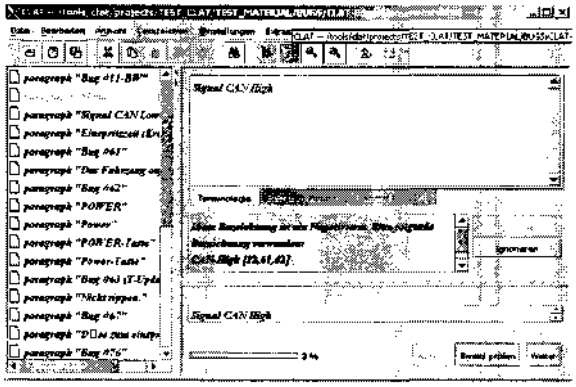

Figure 2: CLAT Front end

3 Evaluating CL Checkers

The aim of CL Checkers is to check the correctness of texts, being able to handle correct input, but also detecting the errors and correcting them or at least making suggestions. They are complex tools containing many different modules, such as parsers, grammars, sets of rules and terminological databases, which must interact with each other to produce the desired results. If we want to test a CL Checker regarding its intrinsic features, we will have to take all these factors into account. It is, however, very complex to obtain reliable results with this kind of testing, since data can be biased by subjective factors and full coverage of precision and recall results is not always possible. Besides, the success of this type of testing does not indicate that the application of a controlled language is indeed results in any of the effects pointed before (better understandability and readability as well as translatability).

In the case of MULTILINT, one chapter in the TETRIS project documentation (IAI, 2005) dealt with its evaluation. This was divided in two parts: "Proof-Reading" and "Hit Rate in Translation Memory Systems". The goal of the first evaluation scenario was to determine the average cost saving potential gained by using MULTILINT in contrast to

human proofreading. The tests included a statistical macro evaluation, where factors such as different scenarios for creation of content, usability of the system and general program behavior were tested. A dynamic micro evaluation was also carried out, focusing on texts verified with MULTILINT. In this case, the results had to be evaluated regarding the information retrieval measures precision and recall, that is, how many mistakes were recalled and, from them, how many of them were indeed correctly recalled (precision). The conclusion of this first evaluation scenario was that MULTILINT, although it assists the technical writer to an important degree, could not completely substitute an experienced and specialized human proof-reading.

The second evaluation scenario, "Hit Rate in Translation Memory Systems", intended to prove that the use of MULTILINT could increase the hit rate in translation memory systems by assuring more consistency in the source texts. Though this scenario was repeated twice, the results were not meaningful enough due to subjective factors such as the learn effect on MULTILINT and the differences on the writing skills of the different authors.

All in all, it was not possible to assess and prove the quality of MULTILINT in a meaningful way. Therefore, a new evaluation approach is needed to test the extrinsic features of a CL Checker, that is, whether the application of a CL Checker carries the advantages pointed out in the first section, and, if so, under which conditions. For this purpose, we use MT technology in order to assess the effectiveness of MULTILINT to make texts translatable. MT technology is the most "objective" evaluator, since translation quality is always the same, with no subjective factors such as translation experience, subject knowledge or rnood playing a role.

4 Evaluating MT: The FEMTI-Framework

Our evaluation methodology is based on the FEMTI-Framework[1] (Hovy et al. (2002); Popescu-Belis et al. (2001)), which offers a base for designing an evaluation procedure of MT systems. The FEMTI-Framework bases on the principles of context-based evaluation (Arnold et al. (1994) and Klein et al. (1998)). This methodology postulates that, before the evaluation starts, it is important to define the context in which it is going to take place. This description contributes to the subsequent choice of the appropriate features to be evaluated and the appropriate metrics to evaluate these features. Indeed, experience has shown that only context-based evaluations in a well-defined domain offer relevant data that fulfil the needs of the evaluator or end-user (King and Falkedal, 1990).

The FEMTI Framework is divided into two sections: the first section contributes to the definition and description of a context in which the evaluation is going to take place. Features such as the purpose of the evaluation, the input characteristics or the role of the MT system within a translation work flow are taken into account. The second section concentrates on the MT internal and external characteristics, meaning the software architecture and the quality of the output. Here features such as MT-system specific characteristics, the functionality of the translation, reliability or usability have to be evaluated.

For the measurement of these features, however, FEMTI only offers a listing of different metrics from the literature, without assessing any standard. The user must decide, according to the context defined, which metrics from the literature are most appropriate to measure the features chosen or he must develop new metrics according to his needs.

4.1 Human versus Automatic Evaluation: Metrics and Measures

Since evaluation has been an issue in MT research and development, human evaluation has been the classical method to assess the quality of a system. This is usually done by means of scales, where the evaluator grades a translation from best to worst, or with questionnaires about the text to check if he understood it. However, this type of evaluation has three main pitfalls: it is costly and time consuming, since usually external evaluators have to be hired to do the job and it takes a while and many evaluators to obtain statistical significant results. Besides, the results of such an evaluation are hardly reusable, since every time an evaluation takes place, the whole procedure has to be repeated. Finally, the results of a human evaluation are subjective, since two evaluators can assess a sentence in a different way depending on many factors such as their education, experience, background information etc.

In the past years new ngram-based intrinsic metrics have been developed to automatically score system-outputs against human-produced reference documents. One of these is BLEU, a corpus-based metric based on the assumption that "the closer a machine translation is to a professional human translation, the better it is" (Papineni et al., 2002). Thus, to assess the quality of a machine translation, the numeric closeness between two translations (a candidate machine translation and one or more reference translations) is calculated, though overgeneration of correct word forms is penalised in order to avoid erroneous results. Also included is a brevity penalty that penalises test sentences found to be much shorter than the reference sentences. NIST was the following important measure to appear (Doddington, 2002), also using ngram co-occurrence statistics.

Automatic evaluation represents a cost-effective method to carry out quick and frequent evaluations. These methods are also useful for contrasting the relative frequency of different MT outputs. However, the results are not always reliable and it is difficult

[1] http://www.issco.unige.ch/projects/isle/femti

to make any statements about the real quality of the system. What does, for instance, a BLEU score of 0,326 mean? Therefore, it is always recommendable to cross-check the results with human evaluation results.

The notion of quality in translation, and more especially in MT, is complex, and it is extremely difficult, not to say impracticable, to find a generally accepted definition. This is because it is impossible to define a "golden standard" to refer to when evaluating a translation, since there are always different translations for a single source text. Translation quality generally depends on the final users and what these expect from it.

5 Selection of resources

The first phase of this study consisted in the selection of the resources needed for the later evaluation. For this purpose, a hypothetical industrial context was first outlined. We considered an automotive company producing a highly technical product, where CL is applied for the creation of technical documentation in German and where, due to high internationalisation and localisation costs, MT could be considered as a complementary solution to other types of translation automation and human translation, especially for the language pair German-English.

In the next sections we present how different technical documents were analysed and how two text types were chosen to carry out the tests. Furthermore, we explain how we proceeded to build a text corpus and, parallelly, to choose three MT commercial systems, with which we translated the whole corpus. Next, the design of a reduced test corpus to carry out the human evaluation, as well as the methods used to carry out automatic evaluations, are exposed. Finally, the results of automatic and human evaluations are described.

5.1 Text Type

It is generally accepted that certain types of text are more appropriate for MT than others. One example is technical documentation. Numerous references (Lehrberger and Bourbeau (1998, p. 192) and Bernth and Gdaniec (2001, p. 175)) highlight this view. Since our aim was to build a corpus of texts most appropriate for MT, we analysed different technical documents from the automobile domain, including repair instructions, technical data and training documentation. These had to fulfil different requirements such as a middle text length or the presence of translatability indicators, which should render the degree of translatability of the texts. In this respect, translatability criteria were studied from different authors (Reuther (2003); Grasse (2001); Bernth and Gdaniec (2001) and Underwood and Jongejan (2001)). After detailed study, we grouped these criteria into four main groups:

- Formal Rules: This group includes criteria regarding punctuation, formatting, layout and orthography. This is an essential category for MT since output quality can suffer enormously if segmentation is not carried out properly.
- Grammar: This group includes syntactic indicators such as ambiguous or too complex structures, subordinate and coordinate clauses, order of elements, use of pronouns, prepositions and articles and sentence length. Other aspects refer to the use of certain verbal forms and tenses, the structure of noun phrases and the presence of ungrammatical constructions.
- Terminology: The restricted use of variants (spelling variants, compound variants, synonyms), abbreviations and acronyms, as well as the usage of a consistent and standardised terminology constitutes the main focus of this group.
- Style: This group concentrates on elliptical and passive constructions, the use of metaphors, slang or dialect variants and application of negation.

All these criteria were cross checked with MULTILINT rules, with the discovery that many of the rules to cover most aspects of translatability. Therefore, we assumed that texts checked with MULTILINT were highly translatable, so only such texts were included in the corpus.

According to the requirements exposed above, repair instructions was the most appropriate text type for the first phase.

5.2 The Text Corpus

As Elliot et al. (2003) point out, there are two ways of assessing the quality of a MT system: A test suite and a text corpus. A test suite is usually artificially created and is designed to test specific linguistic phenomena. This kind of resource is especially used by MT developers to check where the system fails and where it can be improved (a glass-box evaluation approach). Besides, a text corpus is composed by real texts and is therefore more useful for a potential end-user of MT, such as the language department in a company. The corpus typically comprises an original version of the source text, different MT translations (especially if the goal of the evaluation is to compare different MT systems for acquisition) and, possibly, a human reference translation. This depends on the features evaluated and the metrics applied.

Since the goal of the first phase was to choose the MT system that best matches with texts written according to the rules of MULTILINT, a text corpus of real texts verified with the CL checker was built, resulting in over 3000 different segments. Besides, a reduced corpus for human evaluation was designed. It included 250 segments divided into two parts that had to be evaluated with respect to comprehensibility and post-editability. In order to make this reduced corpus as representative as possible,

we analysed the entirety of the corpus to find common grammatical patterns, such as infinitive constructions, imperatives, pre-modifying participial attributes etc. This reduced corpus reflexes statistically the content of the bigger corpus, that is, the sentences chosen represent proportionally the sentences in the larger corpus.

The reduced test corpus was built containing following parts:

- A questionnaire before the evaluation that should give us general information about the situation of the evaluator, his experience with the types of text evaluated etc. This information should explain occasional statistical deviations in the results.
- Test 1 that contained 125 segments that should be evaluated according to a scale of comprehensibility.
- Test 2 that contained another 125 segments that should be first evaluated according to their post-edit ability
- A final questionnaire to check the impressions of the evaluator and his disposition to do post-editing work instead of pure translation.

5.3 MT System

For the realisation of our tests we considered different types of MT Systems, ranging from statistical to rule-based. Although we think that statistical or example-based MT systems could deliver very good results after training them with a representative corpus, two main reasons put us off making use of them: Firstly, the nearly absolute commercial absence of these systems, since most of them are designed in universities and research institutions. Secondly, the large effort needed to build a corpus and train the system so that it could translate texts of a given domain. These are the reasons why, after long consideration, we decided to use for the evaluation only commercial rule-based systems, which are to be used most probably in an industrial context by a language department in a big company or by a translation agency.

After an Internet and literature inquiry, following criteria were considered (Hutchins (2004) and Bernth and Gdaniec (2001)):

- Language pairs: This is one of the key factors when using a MT system commercially. Since we check documents written in German and checked with MULTILINT, the language pairs selection metric was based on the greatest number of language-pairs from German and into German.
- Terminology: Another key feature when selecting a MT system is dictionary coverage. Two aspects were considered for this characteristic: Specialized dictionaries and the possibility to create user dictionaries for corporate terminology.
- Status of Vendor: As mentioned before, literature and Internet research have confirmed that the selected systems, each to a different degree, have successfully carried out projects with important clients. “Buying an MT system is a considerable investment, and the stability and future solvency of the vendor is an important consideration” (Arnold et al., 1994, p. 158).
- Evaluation studies: All of these systems have been evaluated in other studies and have obtained the best general results or were preselected for the evaluation on the basis of favorable characteristics.

The analysis of these factors resulted in the preselection of three commercial systems. In order to choose one of them, it was necessary to determine which of the systems rendered the best output quality for the type of text we had selected.

FEMTI distinguishes two modes in which quality of a translation can be evaluated: without and with adjustment. In the first case, the system is evaluated before the dictionary and/or grammar is adjusted. In the second case, dictionary and/or grammar are adjusted, in order to obtain the best possible results. Of course, the more adjustments are realised, the more severely the evaluation has to be made. Since we were interested in achieving the best possible translation quality and there is no scenario we can imagine in which it could be interesting to apply MT without adjustment, we opted for the second option.

This selected type of text (Section 5.1)included mostly infinitive sentences with imperative meaning, in the form “Trennschloss entriegeln”. In English, this kind of construction has to be translated as pure imperatives, placing the verb at the beginning: “Release belt lock”. We realised, however, that one of the systems was not translating this construction correctly, placing the verb at the end of the sentence. We decided to keep this system in the evaluation because the results of the internal characteristics were quite promising and because the quality of the translation of other constructions was satisfactory. Thus in order to balance the results with the other systems, we added other technical documents to the corpus that contained other types of constructions which cpiod ne better dealt with by the system.

AUTOMATIC EVALUATION

For the automatic evaluation we used the NIST MT evaluation kit, provided by the National Institute of Standards and Technology (NIST) and freely downloadable from their web page[2]. With this kit it is possible to evaluate a corpus using the BLEU and NIST metrics. Different options allow us to consider lower and upper case differences as well as to

[2] http:// www.nist.gov/speech/tests/mt/resources/scoring.htm

change the level of the evaluation from corpus-based and document-based to segment-based.

Both the whole corpus and the reduced corpus were evaluated and are shown in diagrams 3 to 6.

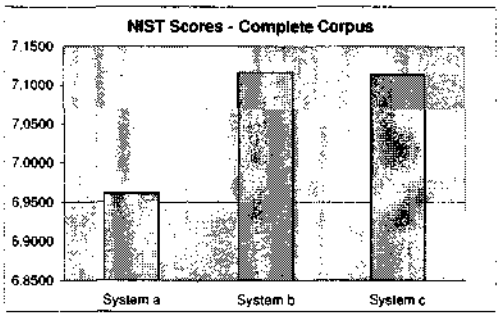

Figure 3: NIST Results on Whole Corpus

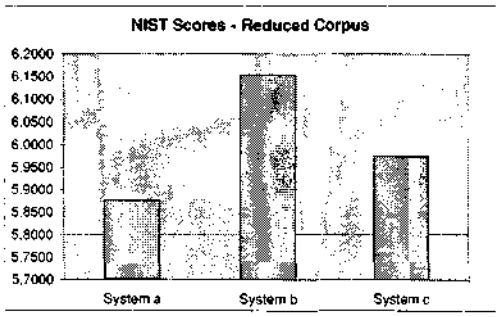

Figure 4: NIST Results on Test Suite

The NIST score declares, on the complete corpus, system B as a winner, closely followed by system C. However, in the evaluation of the reduced corpus, the difference between system B and C grows, and B seems to deliver the best results. In both evaluations, system A falls clearly behind.

The BLEU scores confirm this trend, with system B leading the results, both in the evaluation of the whole corpus and the reduced corpus. However, the distance between system A and C is not so substantial.

As we have seen, the results of both evaluation metrics were contradictory, especially as to the decision between system B and system C. Besides, automatic evaluation presents the problem that it is difficult to interpret these results in terms of their real application when assessing if MT is a technology that can come into question or not. Therefore, we decided to carry out a human evaluation in order to assess if the results of the automatic evaluation were reliable and what steps were necessary to take further on.

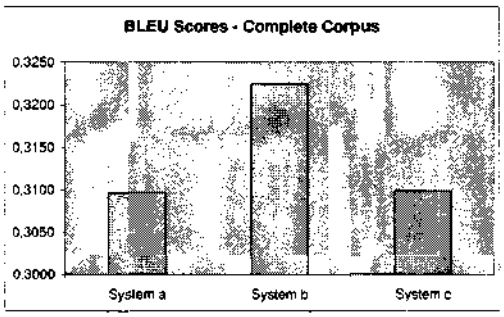

Figure 5: BLEU Results on Whole Corpus

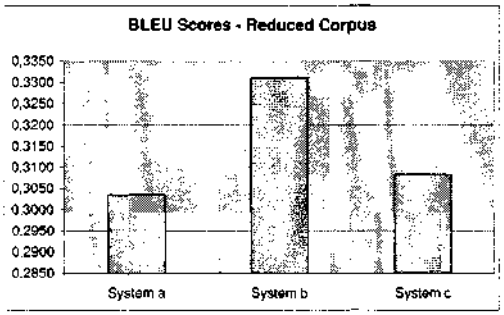

Figure 6: BLEU Results on Test Suite

HUMAN EVALUATION

The evaluation team was composed of 8 professional translators with English as a mother tongue who had at least 3 years experience translating complex technical texts. In this way. we wanted to obtain as h results as homogeneous as possible. The amount of time available for the experiment was one week, since, due to performance questions, we considered that no more than 4 hours a day should be dedicated to evaluate the segments. All in all, translators needed an average of 16 hours to evaluate the whole reduced corpus.

The 250 segments were evaluated regarding following criteria: comprehensibility and post-editability. Based on Rodrigo and Braun-Chen (2001), we assigned two main properties to the criteria:

- K4IN: Key for Information Purposes. MT output as an information source. This property belongs to the measure comprehensibility.
- K4TR: Key for Translation Purposes. MT output regarded as an aid for producing translations of publication quality. This property belongs to the measure post-editability.

As the authors point out as a result of their investigation, the following correlation is usually established: the more informative and/or intelligible the MT output, the more usable it is for information purposes; the less post-editing is needed, the more suitable the MT output is for translation purposes.

The evaluators had to assess the quality of the segments using scales adapted from the literature (see FEMTI). We kept these scales as compact as possible, since too fine grained scales make it difficult to draw definite conclusions.

Comprehensibility (Figure 7) measures the degree to which MT output can be understood by the user. This is especially important for technical documentation, especially for instructional texts, since an accurate understanding of the text is essential to carry out a task. The scale of Comprehensibility included following grading:

1. Totally intelligible: The meaning of the segment is perfectly clear. It is grammatical and reads like ordinary text.
2. Very intelligible: The segment has minor mistakes, but is generally clear and intelligible. It is possible to understand (almost) immediately what it means.
3. Intelligible: Sense can only be understood after repeated reading.
4. Non-intelligible: Segment is unintelligible.

Post-Editability (Figure 8) evaluates how "useful" (usability aspect) the translations produced by the MT systems were in the case that these had to be improved later for publication. This index was intended to indicate the real effort that would be needed to transform machine translated segments into publishable ones. The scale of post-editability included the following grading: No post-edition needed, minimal post-edition needed (option a: only "superficial" modifications such as morphological dependencies, punctuation, accents or articles must be modified), minimal post-edition needed (option b: text must be slightly modified for publication due to ellipsis, over generation or a false sense) and total post-edition needed (the text must be modified for publication, but the source text is needed to make sense of it).

The results of both tests clearly showed the drawback of system A, with the lowest number of "totally intelligible" segments and the highest number of "totally unintelligible" and "total post-edition" segments. This responds with the results of the automatic evaluation.

With respect to the results of systems B and C, these also reflected the results of the automatic evaluation (especially those of the whole corpus), since it was difficult to say which one offered the best "output quality". However, thanks to the human evaluation, we could state which of the systems was better for a special task. In this respect, system B offers

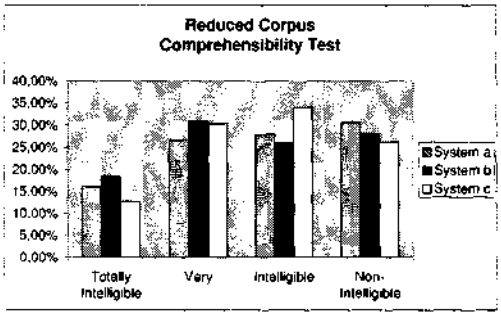

Figure 7: Comprehensibility Test

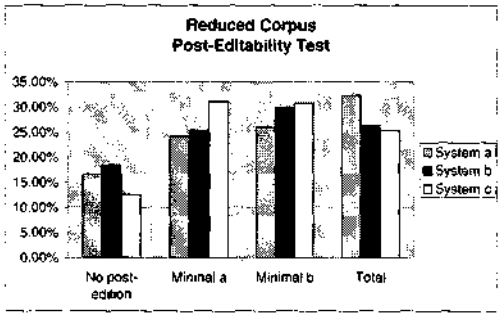

Figure 8: Post-Editability Test

the best Comprehensibility results, with the highest scores in "totally and very intelligible" and a middle score in "unintelligible". This system would be good for its deployment as a system for information gisting or rapid translation of e-mails, company reports etc. On the other side, system C offers the best post-editability results. This was the system that had difficulties when translating imperative constructions in German (see section 6.1). However, the rates in "minimal post-edition a and b" as well as the low "total post-edition" rate show that this system is more appropriate for translation. We think that, with the right implementation of a rule to translate correctly imperative sentences in German, the rate of "not post-edition" would increase dramatically, so that the decision between B and C would be more clear.

When correlating automatic and human results, the automatic evaluation correlates better with "Comprehensibility" measurements, rather than with "post-editability" measures. For measuring post-editability, other automatic metrics such as WER (Word Error Rate)(Tomas et al., 2003) would probably be more adequate.

The purpose of our evaluation was to choose one of these systems to carry out further tests in order to assess if the implementation of CL brings any advan-

tages with respect to the translatability. That is, the key of our evaluation was the translation purpose, regarding MT output as an aid for producing translations of publication quality (property "K4TR" for post-editability). Therefore, we decided to choose system C for the second phase of our study.

6 Conclusions and Outlook

In this paper we have described a methodology to select resources for an evaluation of a controlled language and, more specifically, how to carry out an evaluation of MT systems with respect to their output quality. For this purpose, we have adapted the FEMTI-Framework to an industrial context and have used both human and automatic metrics. The results of both metrics have offered similar results, though the complementary information of the human evaluation and the previous detailed analysis of the context has permitted us to select one of the systems.

In the second phase of our work, we will analyse which rules of the controlled language rule set have a real effect on the quality of the machine translations for the language pair German-English. We will also check if these deviate from human translatability rules and, if so, to which extent. This could lead to a prioritisation of the rules for certain contexts (where MT is going to be applied) or even to the discovery of new rules to improve machine translatability.

7 Acknowledgments

We would like to thank Nicole Brosig from BMW AG for her advice and support as well as for her helpful comments in the revision of this article.

References

Ann H. Adams, Gail W. Austin, and Melissa Taylor. 1999. Developing a Resource for Multinational Writing at Xerox Corporation. *Technical Communication,* 46(2):249-254.

Ingrid Almqvist and Anna Sågvall Hein. 1996. Defining ScaniaSwedish: A Controlled Language for Truck Maintenance. In *Proceedings of the First International Workshop on Controlled Language Applications,* pages 159-165.

D. Arnold, L. Balkan, S. Meijer, R. Humphreys, and L. Sadler. 1994. *Machine Translation: An Introductory Guide.* NEC Blackwell, Manchester, UK.

K. Barthe, C. Juaneda, D. Leseigneur. J.C. Loquet, C. Morin, J. Escande, and A. Vayrette. 1999. GIFAS Rationalized French: A Controlled Language for Aerospace Documentation in French. *Technical Communication,* 46(2):220-229.

Kathy Barthe. 1998. GIFAS Rationalised French Designing one Controlled Language to Match Another. In *Proceedings of the Second International Workshop on Controlled Language Applications,* pages 98-101.

Ulrike Bernardi, Andras Bocsak, and Jorg Porsiel. 2005. Are We Making Ourselves Clear? Terminology Management and Machine Translation at Volkswagen. In *Proceedings of the 10th Annual Conference of the European Association for Machine Translation.*

Arendse Bernth and Claudia Gdaniec. 2001. MTranslatability. *Machine Translation,* (16):175-218.

Arendse Bernth. 1998. EasyEnglish: Preprocessing for MT. In *Proceedings of the Second International Workshop on Controlled Language Applications,* pages 130-41.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Proceedings of the second international conference on human language technology research,* pages 138-145.

Gordon Farrington. 1996. AECMA Simplified English: An Overview of the International Aircraft Maintenance Language. In *Proceedings of the First International Workshop on Controlled Language Applications,* pages 1-23.

Nadine Grasse. 2001. Qualitätskontrolle des MÜ-Systems DCINTRANS in der Anwendung des Sprachendienstes der DaimlerChrysler AG.

Johann Haller. 2001. MULTIDOC: Authoring Aids for Multilingual Technical Documentation. In J. Chabas et al., editor, *Proceedings of the First International Conference on Specialized Translation,* pages 143-147.

Debbie Elliott; Anthony Hartley and Eric Atwell. 2003. Rationale for a multilingual corpus for machine translation evaluation. In *Proceedings of CL2003: International Conference on Corpus Linguistics,* pages 191-200.

Eduard Hovy, Margaret King, and Andrei Popescu. 2002. Principles of Context-Based Machine Translation Evaluation. *Machine Translation,* 17(l):43-75.

John Hutchins. 2004. Compendium of Translation Software: directory of commercial machine translation systems and computer-aided translation support tools. The European Association for Machine Translation (EAMT).

2005. Institut für Angewandte Informationsverarbeitung. http://www.iai-sb.de.

Wladyslaw Janowski. 1998. CL 14: Controlled Language - Risks and Side Effects. *TCForum,* (2):4-5.

C. Kamprath. E. Adolphson, T. Mitamura, and E. Nyberg. 1998. Controlled Language for Multilingual Document Production: Experience with Caterpillar Technical English. In *Proceedings of the 13th COLING Conference on Computational Linguistics,* pages 1-12.

Margaret King and Kirsten Falkedal. 1990. Using Test Suites in Evaluation of Machine Translation Systems. In *Proceedings of the 13th COLING Conference on Computational Linguistics,* pages 433-439.

Judith Klein, Sabine Lehmann, Klaus Netter, and Wegst Tillmann. 1998. DiET in the context of MT evaluation. In Harold Somers, editor. *Computers and Translation: A Handbook,* pages 71-110.

John Lehrberger and Laurent Bourbeau. 1998. *Machine Translation: Linguistic characteristics of MT systems and general methodology of evaluation.* John Benjamins Publishing Company, Asterdam/Philadelfia.

Anne Lehrndorfer. 1996. *Kontrolliertes Deutsch: Linguistische und sprachpsychologische Leitlinien für eine (maschinell) kontrollierte Sprache in der Technischen Dokumentation.* Ph.D. thesis, Universität München.

Linda Means and Kurt Godden. 1996. The Controlled Automotive Service Language (CASL) Project. In *Proceedings of the First International Workshop on Controlled Language Applications,* pages 106-114.

Teruko Mitamura and Eric Nyberg. 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,* pages 158-172.

Eric Nyberg, Teruko Mitamura, and Willem-Olaf Hujisen. 2003. Controlled Language for Authoring and Translation. In Uta Nübel. Rita Seewald-Heeg, editor, *Proceedings of the Konvens '98 in Bonn: Evaluation of the Linguistic Performance of Machine Translation Systems,* pages 107-126. Gardez! Verlag St. Augustin.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Test suites for Controlled Language Checkers. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics,* pages 311-318.

Andrei Popescu-Belis, Sandra Manzi, and Maghi King. 2001. Towards a Two-stage Taxonomy for Machine Translation Evaluation. In *Proceedings of the MT Summit VIII Workshop,* pages 1-8.

Ursula Reuther. 1998. Controlling Language in an Industrial Application. In *Proceedings of the Second International Workshop on Controlled Language Applications,* pages 174-183.

Ursula Reuther. 2003. Two in one: Can it work? Readability and Translatability by means of Controlled Language. In *Proceedings of the Joint Conference of the 8th International Workshop of the European Association for Machine Transla- . tion and the 4th Controlled Language Applications Workshop,* pages 124-132.

Elia Yuste Rodrigo and Francine Braun-Chen. 2001. Comparative Evaluation of the Linguistic Output of MT Systems for Translation and Information Purposes. In *Proceedings of the MT Summit VIII Workshop.*

Nestor Rychtyckyj. 2000. An assessment of Machine Translation for Vehicle Assembly Process Planning at Ford Motor Company. In John S. White, editor, *Proceedings of the 4th Conference of the Association for Machine Translation in the Americas,* pages 107-126. Springer.

Stefanie Schachtl. 1996. Defining ScaniaSwedish: A Controlled Language for Truck Maintenance. In *Proceedings of the First International Workshop on Controlled Language Applications,* pages 143-149.

Jesús Tomas, Josep Angel Mas, and Francisco Casacuberta. 2003. A quantitative method for machine translation evaluation. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics,* pages 12-17.

Nancy Underwood and Bart Jongejan. 2001. Translatability Checker: A Tool to Help Decide Whether to Use MT. In B. Maegaard., editor, *Proceedings of MT Summit VIII.* pages 125-138, Santiago de Compostela.