Cognitive Evaluation Approach for a Controlled Language Post-Editing Experiment

Irina Temnikova
Research Institute in Information and Language Processing
University of Wolverhampton, UK
E-mail: I.Temnikova2@wlv.ac.uk

Abstract

In emergency situations it is crucial that instructions are straightforward to understand. For this reason a controlled language for crisis management (CLCM), based on psycholinguistic studies of human comprehension under stress, was developed. In order to test the impact of CLCM machine translatability of this particular kind of sub-language text, a previous experiment involving machine translation and human post-editing has been conducted. Employing two automatic evaluation metrics, a previous evaluation of the experiment has proved that instructions written according to this CL can improve machine translation (MT) performance. This paper presents a new cognitive evaluation approach for MT post-editing, which is tested on the previous controlled and uncontrolled textual data. The presented evaluation approach allows a deeper look into the post-editing process and specifically how much effort post-editors put into correcting the different kinds of MT errors. The method is based on existing MT error classification, which is enriched with a new error ranking motivated by the cognitive effort involved in the detection and correction of these MT errors. The preliminary results of applying this approach to a subset of the original data confirmed once again the positive impact of CLCM on emergency instructions' machine translatability and thus the validity of the approach.

1. Introduction

In emergency situations such as flooding, fires and earthquakes it is crucial that instructions transmitted by the government to the general population are clear and straightforward to understand. It is known that human comprehension functions differently in stress situations than under normal conditions. For this reason a specially designed controlled language CLCM, which is based on psycholinguistic studies of human comprehension under stress, was developed on the basis of text analysis of a corpus of emergency instructions. A previous experiment was carried out to test CLCM on both human and machine translation tasks. The results proved that instructions written according to CLCM rules have a positive impact on both human and machine translation. This was proved by two automatic evaluation metrics (times to post-edit and edit distance). This paper describes a new evaluation approach which was applied to a subset of the previous experiment data and presents its results. The new approach allows a deeper look into the post-editing process (PE) of the MT output in relation to the cognitive effort involved in the detection and correction of the MT errors. The remainder of the paper is structured as follows: Section 2 gives an overview of the state-of-the-art in controlled languages and in measuring PE effort, Section 3 describes the previous experiment, Section 4 presents the new evaluation approach, including a discussion of the shortcomings of the existing cognitive evaluation approaches, Section 5 discusses the evaluation results, and Section 6 provides the conclusions and some future work.

2. Controlled Languages and the State-of-the-Art in Measuring PE Effort

We define controlled languages (CLs) 'specially designed natural language subsets that obey syntactic and lexical restrictions' (Temnikova & Margova, 2009). Different types of CLs exist, depending on the purpose of the restrictions aiming to reduce text ambiguity and complexity. Generally, CLs for humans improve human comprehension (e.g. Plain English Campaign[1] and Basic English (O'Brien, 2006)). CLs designed for computer applications (for MT, text summarization and other) aim to adapt the text input for particular applications in order to improve their performance. This paper focuses on evaluating the impact of pre-editing instructions according to the CLCM rules (which are designed for humans) on machine translation. An example of CLs specifically designed for MT is the one by Nyberg et al. (2003). Controlled language pre-editing of MT input texts has been done also before with examples like the TITUS system, which restricted the input syntax and vocabulary for machine translations of abstracts in textile industry (Streiff, 1985) and the machine translation at XEROX (Ruffino, 1982), who pre-edited their texts with their own controlled language rules. There are several approaches to evaluating the impact of CL pre-editing on MT performance. Only a few of them involve a combination of MT quality evaluation, such as BLEU scores (Papineni et al., 2002; Aikawa et al., 2007), and PE evaluation metrics. The existing PE evaluation metrics measure the PE effort from three different points of view: temporal,

[1] http://www.plainenglish.co.uk.

technical and cognitive. Measuring temporal effort (Krings, 2001; O'Brien, 2005; O'Brien, 2006) requires measuring the time necessary to post-edit MT output. Technical metrics involve computing the number of additions, deletions and cuts-and-pastes (Krings, 2001; O'Brien, 2005; O'Brien, 2006) or measuring character-based edit distance (Aikawa et al., 2007) between MT output and post-edited text. Cognitive approaches involve standard cognitive metrics such as think-aloud protocols (TAP) (Krings, 2001) or Choice Network Analysis (CNA) (O'Brien, 2005; O'Brien, 2006). The present paper proposes a new cognitive evaluation approach, which will help to overcome the limitations of the existing cognitive metrics, while offering a valid way to evaluate PE effort.

3. Description of the Previous PE Experiment

As mentioned before, the evaluation approach described in the present paper is tested on a subset of the data of a previously conducted and evaluated experiment. The experiment aimed to test whether a CL designed for improving human comprehension would improve MT performance. CLCM covers the crisis management domain which is distinctive in the fact that its language encompasses several different sub-languages. CLCM was developed in the context of the MESSAGE Project[2], an EU-funded project which produced CL rules for texts in the emergency domain for four EU languages (English, French, Polish and Spanish) and prototypes for two more (Bulgarian and Greek). The experiment involved two different texts of comparable length (around 140 words) and similar complexity, which was ensured by the fact that both texts were parts of the same source document "Individual Preparedness and Response to Chemical, Radiological, Nuclear, and Biological Terrorist Attacks"[3]. One text consisted of instructions for the general public about what to do in case of a chemical attack and the other one consisted of instructions in case of a nuclear attack. The two texts are given in the Appendix of our previous paper (Temnikova & Orasan, 2009). The "nuclear attack" instructions are left as they are from the original source, while the "chemical attack" ones have been re-written according to the CLCM rules. The two texts were translated with a freely available statistical MT engine (Google Translate). Although MESSAGE Project developed CL rules for six EU languages, the CL rules were language-dependent and the experiment was conducted only on English texts. The pre-editing was based on an initial prototype of CLCM and involved the following set of rules:

- Use only literal meaning.
- Avoid idiomatic expressions.
- Use concrete (instead of abstract) concepts.
- Write short sentences.
- Write only 1 piece of information (condition, instruction, item) per line.
- Use the allowed structure 'How to ..' for writing titles.
- Divide the specific situations into separate blocks.
- Write a title for every specific situation.
- Remove unimportant information.
- If an adjective modifies more than 1 entity repeat the adjective next to every modified entity.
- Write conditions before the instructions.
- Use less ambiguous expressions.
- Avoid technical terms.
- If possible, use finite verb instead of an '-ing' form.

Twenty-five translation specialists were asked to post-edit the MT output translated from English into seven European languages (Bulgarian, Dutch, Maltese, Modern Greek, Russian, Slovenian and Spanish). The number of specialists per language varied from 3 to 5 and depended on the availability of translators. Post-editors were given instructions regarding the purpose of the post-editing (the edited document will not be published). Also, post-editors were given the instruction to ignore stylistic errors if they didn't affect sentence meaning.

In order to evaluate the impact of CLCM on MT performance, two completely automatic evaluation metrics were used: time employed by the translation specialists to post-edit the machine translation of both texts and Levenshtein edit distance (Levenshtein, 1966) between the MT outputs and post-edited texts for both texts. This allowed evaluation from both temporal and technical points of view. Edit distance was calculated in order to evaluate the amount of changes the post-editors needed to apply to the MT output in order to re-write it in correct language. Table 1 and Table 2 show the post-editing times and edit-distance average values for both original and re-written texts (a simple (not weighted) average is calculated), together with their difference in percentage. Positive values indicate improvement in the performance for the controlled language text, while negative ones – a decrease in performance for the controlled language text.

[2]This paper was produced in the context of MESSAGE Project (full title: Alert Messages and Protocols), project financed by the European Union (JLS/2007/CIPS/022).

With the support of the Prevention, Preparedness and Consequence Management of Terrorism and other Security-related Risks Programme European Commission - Directorate-General Justice, Freedom and Security.

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein

[3]The document is available at:
http://www.rand.org/pubs/monograph_reports/MR1731/

Target language	Not-rewritten text	CL re-written text	% difference
BG	300.2 s	237.8 s	21%
SL	302.7 s	218.0 s	28%
RU	501.3 s	438.3 s	13%
ES	351.0 s	195.3 s	45%
NL	394.0 s	366.0 s	7%
MT	249.7 s	245.0 s	2%
EL	318.3 s	252.0 s	21%
Average improvement			20%

Table 1. Post-editing times

Language pair	Original	Simplified	% difference
English – Bulgarian	0.375	0.235	37%
English – Russian	0.388	0.328	16%
English – Slovenian	0.469	0.417	11%
English – Dutch	0.401	0.403	-1%
English – Maltese	0.370	0.405	-9%
English – Greek	0.349	0.194	45%
English – Spanish	0.272	0.195	28%
Average improvement			18%

Table 2. Average number of changed characters relative to the whole number of characters per text

Although the overall results showed improvement, they presented a big variation. For all language pairs the results showed that less time was employed to post-edit the MT output for the CLCM pre-edited text, but the improvement varied strongly - between 2% (for Maltese) and 45% (for Spanish) with an average of 20%. The edit distance per character showed that fewer changes were needed in the case of the pre-edited text for all languages except for Maltese and Dutch, where the results showed a negative impact of the CL on the MT performance. The average improvement was 18%. One of the explanations of such great variation could be that the experiment was not conducted in a controlled environment, so some of the post-editors could be distracted by external factors while performing the task. Other explanations include language-specific reasons which will be explored with the new evaluation approach in future work. There are also shortcomings in using edit distance as a measure for this kind of task. For example it would not distinguish between different types of changes.

The limitations of the employed evaluation techniques motivated the need for a third evaluation method and a deeper examination of the PE of the MT output.

4. Cognitive Effort Evaluation Approach

As mentioned in Section 2, there are two existing evaluation metrics for measuring the cognitive effort involved in post-editing: the think-aloud protocols (TAP) and the choice-network analysis (CNA). The TAP consisted of post-editors commenting on their decisions out loud (Krings, 2001). TAP's shortcoming is that it cannot be easily formalized and reused. CNA (O'Brien, 2005; O'Brien, 2006) focuses on the number of different changes the post-editors apply to MT output words. It was considered that the bigger the number of different changes of the same word, the more cognitive effort a post-editor needs to undertake in order to choose an option. The shortcoming of this evaluation technique lies in the fact that it is not certain that all options are equally available to all post-editors.

On the basis of the shortcomings of these two existing cognitive evaluation techniques and given the need for triangulation, we propose a new evaluation approach. The new method enriches an existing error classification of the MT output (Vilar et al., 2006) with an interpretation of the errors in light of the cognitive effort post-editors need to apply in order to correct different MT errors (orthographic, semantic/syntactic).

Manual evaluation of the data of three out of the seven languages involved in the previous experiment (Bulgarian, Russian and Spanish) was conducted. The choice of languages was motivated by both the need to test the evaluation approach on different language types and the availability of evaluators. The languages chosen were from two different language families. Bulgarian is considered to be distinctive in the fact that it exhibits characteristics of both a Slavic and a Balkan language, Russian is known to be a highly inflected Slavic language, while Spanish is an analytic language from the Romance family. The evaluation involved three independent evaluators (one per language) who are native speakers of the respective languages and are not co-authors of this paper. The evaluators were given instructions based on a modified version of the MT error classification presented in (Vilar et al., 2006), which classifies errors in four main categories: missing words (1), word order (2), incorrect words (3) and punctuation errors (4). Some of the main categories are further divided into a few sub-categories. The complete error classification is given in Table 3:

Error	Correction
(1.) Missing word	Error correction requires adding the missing word
(2.1.) Word order error	Error correction requires moving

	single words
(2.2.) Word order error	Error correction requires moving whole phrases
(3.1.) An incorrect word	Error correction requires replacing with a completely different lexical item
(3.2.) Correct word with an incorrect ending (e.g. number or case)	Error correction requires replacing with a different ending
(3.3.) An incorrect word	Error correction requires a different style synonym
(3.4.) Extra word	Error correction requires deleting the extra word
(3.5.) Error due to incorrectly recognised idiomatic expressions	Error correction requires replacing with the correct translation of the idiomatic expression
(4.1.) Missing punctuation sign	Error correction requires adding the missing punctuation sign(s)
(4.2.) Incorrect punctuation sign	Error correction requires replacing with the correct punctuation sign(s)

Table 3. MT error classification

The original MT output error types are enriched with information regarding the cognitive effort involved in detecting and correcting those errors. This allows errors to be ranked according to how difficult it is for post-editors to correct them, based on the error type and processing sentence span. An attempt for a relative ranking of MT errors, with (1) being the easiest one and (10) the hardest one, is given in Table 4:

Morphol. level	1. Correct word, incorrect form (CInF)
Lexical level	2. Incorrect style synonym (Styl)
	3. Incorrect word (InW)
	4. Extra word (ExW)
	5. Missing word (MissW)
	6. Idiomatic expression (Idiom)
Syntactic level	7. Wrong Punctuation (InP)
	8. Missing Punctuation (MissP)
	9. Word Order at Word level (WoW)
	10. Word Order at Phrase level (WoPh)

Table 4. Cognitive MT Error Ranking

The proposed here error ranking is based on the cognitive model of reading (Harley, 2008), Baddeley's working memory theory (Baddeley & Hitch, 1974) and written error detection studies (Larigauderie, 1998). Working memory plays an important role in reading. According to (Harley, 2008), working memory is composed of a central executive (which plays the role of a supervisory system), a phonological loop and a visuo-spatial temporal store. Written language comprehension is performed unconsciously on several levels (Baddeley & Hitch, 1974; Harley, 2008) with different working memory components involved in the process. First grapheme recognition is performed, then lexical access and then syntactic and semantic processing. The first two levels are less cognitively costly because they require activation in memory of previous representations and mental vocabulary look-up. The phonological loop is activated for any processing above the word level and experiments (Larigauderie, 1998) showed that the syntactic and semantic levels challenge the central executive much more, as they involve understanding the whole text passage and relating its meaning to the meanings of the previous text passages. Task difficulty also depends on processing span with the following ranking of difficulty: word level, clause level, and sentence level. We assume that the post-editing task is very similar to appositely performed error detection and correction task. According to this point of view, the less cognitively costly errors are those at word level, i.e. words with wrong endings which require only a grammar rule representation look-up and the most cognitively expensive ones are those involving syntactic and semantic processing of the whole sentence. In this paper we do not enter into details about the difference in the costs of the errors in relation to the processing span involved in their detection and correction.

5. Discussion of the Evaluation Results

The results of applying this error analysis on the three languages are given respectively in Figure 1, Figure 2 and Figure 3 for Bulgarian, Russian and Spanish. For each error type, the darker column represents the number of errors found in the complex text ('complex' meaning the text not subjected to the CL), while the lighter one - the number of errors of the same type found in the simplified text (i.e. the text subjected to the CL pre-editing). It is clearly seen that the generated MT errors of the complex text and those of the simplified text have different distributions. In fact, the biggest problem for the complex text is 'incorrect words', and the smallest – stylistic errors. This can be explained by the fact that post-editors were explicitly given the instruction to avoid considering stylistic errors if there is no drastic change in the meaning. In contrast, the biggest problem of the simplified text is 'correct word, incorrect form' (i.e. 'morphological') errors, while there are no errors of the types 'incorrect punctuation', 'missing punctuation', 'word order at word level' and 'word order at phrase level' (i.e. semantic and syntactic kinds of errors). This means that while the controlled language pre-editing of MT input texts generates a higher number of errors which are easy to detect and correct, it still generates a smaller number of

errors which are cognitively difficult to detect and correct. Thus, the new evaluation approach also confirms the improvement, which is shown by the first two evaluation methods.

Additionally, there are almost no errors (and no changes in the number of errors) for 'extra words', and 'missing words'. The increased number of 'missing and extra words' in Russian and Spanish controlled texts can be explained by the fact that less context generates more syntactical ambiguity and thus more syntactic errors. The nature of the missing and extra words is not taken into consideration in this paper and for this reason no conclusions can be drawn regarding whether these errors affect the syntax or the semantics of the text. There are no 'idiomatic expressions' errors in Spanish and Russian, while errors of this kind are present in the Bulgarian complex text and disappear in the Bulgarian simplified text. We explain their disappearance with the instruction given to post-editors to specifically avoid using idiomatic expressions. It can be also noticed that the three graphics have different distributions, which can be explained by the different nature of the analyzed languages.

There are no word-order-related MT errors in Russian in both texts. This can be explained by the fact that since Russian is considered to be a relatively free-word-order language, unusual word order is not considered to be a mistake by either the evaluators or the post-editors.

In this way the results of our study, even if it has been conducted on very little statistical data, show that the proposed error ranking should be different depending on the language.

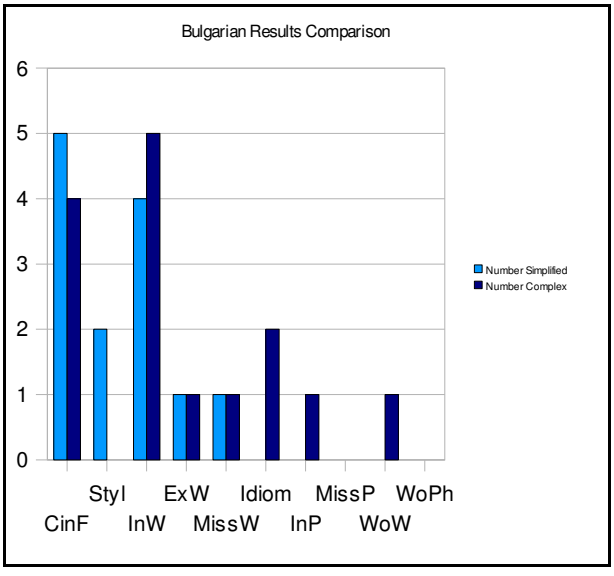

Figure 1. Comparison of the results for Bulgarian

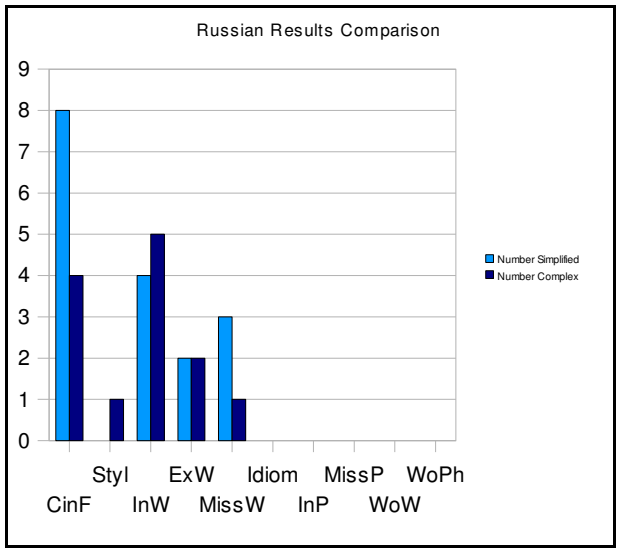

Figure 2. Comparison of the Results for Russian

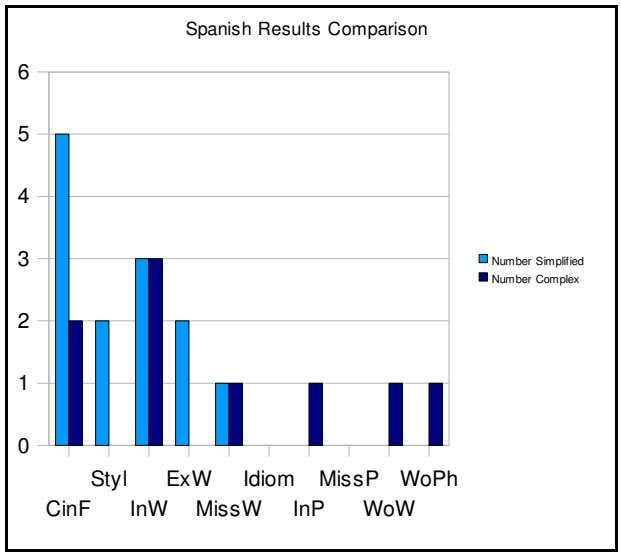

Figure 3. Comparison of the Results for Spanish

6. Conclusions and Future Work

This paper has presented a new cognitive evaluation approach, which allows in-depth examination of the post-editing process in order to spot any differences between the text which has not been subjected and the text, which has been subjected to a specifically designed emergency controlled language pre-editing rules. The approach is based on a combination of semi-automatic MT error classification and error ranking based on the human cognitive effort required to detect and correct different types of errors generated by an MT engine. The results of applying this approach to a subset of evaluation data taken from a previous experiment have been presented. The evaluation data consists of the machine translations

of two short emergency instructions texts, one subjected to controlled language rules and one left as it was originally. The results show the same conclusions as the previously applied two other evaluation metrics. In fact, there are less errors which are cognitively difficult to detect and correct and more errors which are cognitively easy to detect and correct for the simplified text in comparison with in the complex text. Future work will include testing the evaluation approach on larger data sets, more languages, a statistical significance analysis and more fine-grained error ranking according to the processing span involved in post-editing error detection and correction.

7. Acknowledgements

The author is thankful to Prof. R. Mitkov, Dr. K.B. Cohen, Dr. C. Orasan, Dr. L. Specia, Dr. L.A. Ha, R. Evans and A. Carminke for all their useful comments and discussions of the paper.

8. References

Aikawa, T. et al. (2007). "Impact of Controlled Language on Translation Quality and Post-Editing in a Statistical Machine Translation Environment". In Proceedings of the MT Summit XI. Copenhagen, Denmark, 10-14 September. 1-7.

Allen, J.: 2001, 'Post-Editing: An Integrated Part of a Translation Software Program'. Language International, April 2001, pp. 26-29.

Allen, J. (2002). Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes, book review, Multilingual Computing&Technology, 13.2, March 2002, 27-29.

Arnold, D. et al. Machine Translation. An Introductory Guide. Blackwells-NCC, London, 1993.

Baddeley, A.D., & Hitch, G. (1974). Working memory. In G.H. Bower (Ed.), The psychology of learning and motivation: Advances in research and theory (Vol. 8, pp. 47--89). New York: Academic Press.

Harley, T.A., (2008). The Psychology of the Language: from data to theory, Psychology Press, Hove and New York.

Larigauderie P, Gaonac'h D, Lacroix N. Working memory and error detection in texts: what are the roles of the central executive and the phonological loop? Applied Cognitive Psychology, 1998, 12: 505~527.

Levenshtein, V.I., (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10 :707–710.

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

Marslen-Wilson,W. (1990). Activation, competition, and frequency in lexical access. In G. Altmann (Ed.), Cognitive Models of speech processing, pp. 148-172. Cambridge: MIT Press.

Mitamura, T. and Nyberg, E. (2002) Automatic rewriting for controlled language translation. NLPRS 2002 Workshop on Automatic Paraphrasing: Theories and Applications.

Muegge, U. (2006), "Fully automatic high quality machine translation of restricted text: A case study", Proceedings of the twenty-eighth international conference on translating and the computer, 16-17 November 2006, London: Aslib, pp. 16–17.

Murphy, D., (2000). Keeping translation technology under control. In: Machine Translation Review, issue 11: December 2000; pp. 11-13.

Nyberg E., Mitamura T. & Huijsen, W.O., (2003). Controlled language for authoring and translation. In: Harold Somers (ed.) Computers and translation: a translator's guide (Amsterdam/Philadelphia: John Benjamins Publishing Company, 2003); pp.245-281.

O'Brien, S. (2005). 'Methodologies for Measuring the Correlations between Post-Editing Effort and Machine Text Translatability'. Machine Translation. Vol. 19, No 1. pp. 37-58.

O'Brien, S. (2006). Controlled Language and Post-Editing. Multilingual, Issue 83, pp. 17-19.

Ogden, Ch. K. (1930) Basic English: a general introduction with rules and grammar, London,Kegan Paul, Trench, Trubner.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). "BLEU: a method for automatic evaluation of machine translation" in ACL-2002: 40th Annual meeting of the Association for Computational Linguistics pp. 311–318.

Ruffino, J.R. (1982). Coping with machine translation. In: Lawson (1982), 57-60.

Schäfner, F. (2003), "MT post-editing: How to shed light on the "unknown task". Experiences made at SAP", Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop, 15-17 May 2003, Dublin City University.

Snover, M., Madnani, N., Dorr B. and Schwartz, R. "Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric," in Proceedings of the Fourth Workshop on Statistical Machine Translation at the 12th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2009), 2009.

Streiff, A.A. (1985). New developments in TITUS 4. In: Lawson (1985), 185-192.

Temnikova, I. and Margova, R. (2009). Towards a Controlled Language in Crisis Management: The Case of Bulgarian. Proceedings of the International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages (ISMTCL), Besancon, France, July 1-3, 2009.

Temnikova, I. and Orasan, C. (2009). Post-editing Experiments with MT for a Controlled Language. Proceedings of the International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages (ISMTCL), Besancon, France, July 1-3.

Vilar, D., Xu, J., D'Haro, L., Ney, H. (2006). Error analysis of statistical machine translation output. 5th International Conference on Language Resources and Evaluation, LREC'06, pp. 697–702.

Wagner, E. "Post-editing Systran – A Challenge for Commission Translators", Terminologie & Traduction, 1985-3.