

# A Fluency Error Categorization Scheme to Guide Automated Machine Translation Evaluation

Debbie Elliott, Anthony Hartley, and Eric Atwell

School of Computing and Centre for Translation Studies, University of Leeds, LS2 9JT, UK  
{debe,eric}@comp.leeds.ac.uk, a.hartley@leeds.ac.uk,

**Abstract.** Existing automated MT evaluation methods often require expert human translations. These are produced for every language pair evaluated and, due to this expense, subsequent evaluations tend to rely on the same texts, which do not necessarily reflect real MT use. In contrast, we are designing an automated MT evaluation system, intended for use by post-editors, purchasers and developers, that requires nothing but the raw MT output. Furthermore, our research is based on texts that reflect corporate use of MT. This paper describes our first step in system design: a hierarchical classification scheme of fluency errors in English MT output, to enable us to identify error types and frequencies, and guide the selection of errors for automated detection. We present results from the statistical analysis of 20,000 words of MT output, manually annotated using our classification scheme, and describe correlations between error frequencies and human scores for fluency and adequacy.

## 1 Introduction

Automated machine translation evaluation is quicker and cheaper than obtaining human judgments on translation quality. However, automated methods are ultimately validated by the establishment of correlations with human scores. Overviews of both human and automated methods for MT evaluation can be found in [1] and on the FEMTI<sup>1</sup> website [2]. Although existing automated methods such as BLEU [3] and RED [4] can produce scores that correlate with human quality judgments, these methods still require human translations, which are expensive to produce. BLEU requires up to four human ‘reference’ translations against which MT output is automatically compared and scored according to modified  $n$ -gram precision. The test corpus used for this research comprised 500 sentences from general news stories, with four human translations of each. RED, on the other hand, automatically ranks MT output based on edit distances to multiple reference translations. In [5], 16 human reference translations of 345 sentences in two language directions were used from the Basic Travel Expression Corpus [6].

To eliminate the expense of producing human translations, and to investigate the potential of a more portable method, our aim is to design an automated MT evaluation system, initially for language pairs in which the target language is English, which does not require human reference translations. The system will detect fluency errors

---

<sup>1</sup> A Framework for the Evaluation of Machine Translation in ISLE.

characteristic of MT output, and will be designed to meet the needs of post-editors, purchasers and developers.

## 2 Texts for Evaluation Research

Many published MT evaluation projects, such as BLEU [3] and the DARPA evaluation series [7] have based their research entirely on newspaper texts. Many subsequent MT evaluation experiments have also made use of the DARPA corpus, such as [8], [9], [10], [11], [12] and [13]. Consequently, we conducted a survey of MT users in 2003 to find out which text types were most frequently translated by MT systems. Responses showed a great difference between the use of MT by companies/organizations and by individuals who machine translated documents for personal use [14]. It was found that companies most frequently machine translated user manuals and technical documents on a large scale. As a result, the decision was taken to collect such texts for our evaluation research, along with a smaller number of legislative and medical documents, which also figured highly among survey responses. The resulting multilingual parallel corpus is TECMATE, (a TEchnical Corpus for MACHine Translation Evaluation), comprising source texts, human and machine translations, and human scores for fluency and adequacy for an increasing number of texts [15].

## 3 Designing an Error Categorization Scheme

The decision to devise a classification scheme of fluency errors stemmed from the need to identify error types in MT output to guide automated evaluation. Statistics from the human annotation of MT output using such a scheme would provide information on the frequency of error types in texts produced by different MT systems and would help us select errors for automated detection. Statistics would also enable us to compare error type frequency with human judgments for fluency and adequacy, enabling us to focus on the detection of those error types whose frequency correlated with lower human scores for one or both of those attributes.

Fine-grained error classification schemes are not practical for the black-box evaluation of large numbers of machine translations; such a method is even more time-consuming than, for instance, the evaluation of fluency or fidelity at segment level. Consequently, few MT error classification schemes have been devised, and most have been designed with a particular purpose in mind. The SAE J2450 Quality Metric, developed by the Society of Automotive Engineers [16], and the Framework for Standard Error Marking devised by the American Translators Association [17] were both designed for the evaluation of human translations and are insufficiently fine-grained for our purpose. Correa's typology of errors commonly found in automatic translation [18] was also unsuited to our needs, largely because it was designed for glass-box evaluations during system development. Flanagan's Error Classification for MT Evaluation [19] to allow end-users to compare translations by competing systems, Loffler-Laurian's typology of errors for MT, based on linguistic problems for post-editors [20] and classifications by Roudaud et al. [21], Chaumier and Green in [22] provide a more useful starting point for our work. However, these are still insuf-

ficiently fine-grained for our purpose, all rely on access to the source text, and most are based on errors found in translations out of English. As our intention is to design an automated error detection system that does not require access to the original or to any human translation for comparison, it was essential to devise categories based on the analysis of MT output in isolation.

Our classification of errors was progressively developed during the analysis and manual annotation of approximately 20,000 words of MT output, translated from French into English by four systems (Systran, Reverso Prompt, Comprehium and SDL's online FreeTranslation<sup>2</sup>). The four machine translations of twelve texts (each of approximately 400 words) from the TECMATE corpus were annotated with error types. The texts comprised three extracts from software user manuals, three FAQs (frequently asked questions) on software applications, three press releases on technical topics and three extracts from technical reports taken from the BAF corpus<sup>3</sup>. All texts were chosen on the basis that they would be understandable to regular users of computer applications.

Annotations were made according to items that a post-editor would need to amend if he/she were revising the texts to publishable quality. Although the source text was not made available, knowledge of the source language was necessary, as the scheme requires untranslated words to be annotated with parts-of-speech. Furthermore, it was important for the annotator to be familiar with the named entities and acronyms (eg. names of software applications) in the texts, to better represent the end-user and code these terms appropriately.

Errors were annotated using the Systemic Coder<sup>4</sup>, a tool that supports hierarchical linguistic coding schemes and enables subsequent statistical analyses. Error types were divided according to parts-of-speech, as this would provide more detailed information for analysis and would enable us to make more informed decisions when selecting and weighting errors for our automated system. As the Coder supports the insertion of new nodes into the hierarchy at any time, this facilitated the progressive data-driven refinement of the coding scheme. For example, after annotating around 1,000 words, a decision was taken to sub-divide 'inappropriate' items (see Figure 1) into 'meaning clear', 'meaning unclear' and 'outrageous' (words with an extremely low probability of appearing in a particular text type and subject area). This refinement would enable us to make better comparisons between MT systems, and isolate those errors that have a greater effect on intelligibility.

During these initial stages of analysis, it became clear that, having set out to annotate fluency errors, adequacy errors were also detectable as contributors to disfluency, despite the absence of the source text. Words or phrases that were obviously incorrect in the given context were marked as 'meaning unclear' and can be seen as both fluency and adequacy errors. For this research, we can, therefore, define each annotated error as a unit of language that surprises the reader because its usage does not seem natural in the context in which it appears.

---

<sup>2</sup> <http://www.freetranslation.com/>

<sup>3</sup> <http://www-rali.iro.umontreal.ca/arc-a2/BAF/Description.html>

<sup>4</sup> <http://www.wagsoft.com/Coder/index.html>

<i>Part of speech</i>	Noun: Compound	Noun string or Named entity	Inappropriate	Meaning clear/unclear
				Part meaning clear/unclear
			Untranslated	Untranslated
				Part untranslated
	Noun: Acronym	Incorrect		
	Noun: Pronoun	Inappropriate	Incorrect anaphor	
			Other	
		Untranslated		
		Unnecessary		
		Omitted	Direct object pronoun	
	Relative pronoun			
	Other			
	Noun: Common or Adjective or Adverb or Conjunction	Inappropriate	Meaning clear/unclear	
			Outrageous	
		Untranslated		
		Unnecessary		
	Preposition	Omitted		
		Inappropriate	With noun/verb/adjective	
		Untranslated		
		Unnecessary		
	Determiner	Inappropriate		
		Untranslated		
		Unnecessary	Definite article	
			Indefinite article	
Other				
Omitted		Definite article		
		Indefinite article		
	Other			
Verb	Inappropriate	Meaning clear/unclear		
		Outrageous		
		Multiword verb structure		
	Untranslated			
	Unnecessary			
Omitted				
<i>Tense or conjugation</i>	Tense or mood			
	Conjugation			
<i>Incorrect position</i>	Acronym / pronoun / common noun / adjective / adverb / conjunction / preposition / determiner / verb / negator / noun string appendage			
	Compound noun sequence	Noun string or Named entity	Word order	
			Arrangement	
<i>Other</i>	Part of speech incorrect / inelegant or inappropriate style / incomprehensible expression / spelling error / incorrect negation / ordinal number untranslated / qualifier unnecessary			
	Number	Singular should be plural / plural should be singular		
	Case	Upper case required / lower case required		

Fig. 1. Fluency Error Categorization Scheme

## 4 Organization of Error Categories

The current scheme contains all error types found in the French-English MT output. However, the organization of categories reflects the constraints of the tool to a certain extent. It was noticed during the annotation process that items often involved two and, in rare instances, three error types. For example, a noun could be ‘inappropriate’, its position within the phrase could be incorrect and it could lack a required capital letter, or a verb could be ‘inappropriate’ and the tense also incorrect. The scheme was, therefore, organized in such a way that the tool would allow all of these combinations of categories to be assigned to the same word or group of words where necessary.

**Table 1.** Some definitions of categories and examples of annotation

<b>Error category</b>	<b>Definitions and examples</b>
Outrageous	The item has an extremely low probability of appearing in this text type and subject area. Eg. <i>beach</i> rather than <i>time slot</i> , <i>shelterers</i> rather than ( <i>web</i> ) <i>hosts</i> .
Multi-word verb structure	A verb comprising multiple words (in addition to prepositions) is incorrect. Eg. <i>are more priority than</i> as opposed to <i>take priority over</i> .
Noun string / named-entity word order	The constituent parts are ordered incorrectly. Eg. <i>Properties Internet Connection</i> rather than <i>Internet Connection Properties</i> .
Noun string / named-entity arrangement	The constituent parts are ordered incorrectly and additional words are included (common in translations from French into English). Eg. <i>window of definition of the filter</i> rather than <i>filter definition window</i> .
Noun string appendage position	Two noun strings are ‘combined’ so that when translated into English, the word order is incorrect. Eg. <i>tabs of options <u>and regulations</u></i> should be <i>options <u>and regulations</u> tabs</i> . Here <i>tabs of options</i> would be marked ‘noun string arrangement’ and the words underlined would be marked as ‘incorrect noun string appendage position’.
Noun inappropriate, meaning clear	Eg. ... <i>a cd-rom placed in the <u>reader</u> of the device</i>
Noun inappropriate, meaning unclear	Eg. ... <i>activating the <u>notch</u>, you will see the lunar globe ...</i>
Verb inappropriate, meaning clear	Eg. ... <i>if the open file was already <u>registered</u> in this format ...</i>
Verb inappropriate, meaning unclear	Eg. ... <i>the main window <u>behaves</u> the menu bar...</i>
Adverb position	Eg. <i>Francophone users avoid <u>systematically</u> using...</i>
Definite article omitted	Eg. <i>Among *** most frequent, ...</i> (Three asterisks are inserted to mark the omission of an item.)

An item can be annotated with up to four main categories: part-of-speech, verb tense or conjugation, incorrect position and ‘other’, as shown on the left-hand side of Figure 1. Sub-categories must then be selected, moving from left to right, until the final node is reached. The Systemic Coder allows categories to be added, moved or deleted at any time. This will be essential for the analysis of MT output from other source languages, in which we expect to find different error types. Table 1 provides definitions of some of the categories with examples of coded text.

## 5 Capturing the Essence of Fluency and Adequacy

Statistics from our annotations were compared with human evaluation scores to explore correlations between the number of errors annotated and intuitive judgments on fluency and adequacy. Each of the 48 machine translations was evaluated by three different judges for each attribute. Texts were evaluated at segment level on a scale of 1-5, using metrics based on the DARPA evaluations [7]. For fluency, evaluators had access only to the translation; for adequacy, judges compared candidate segments with an aligned human reference translation. A mean score was calculated per segment for each attribute. These scores were then used to generate a mean score per text and per system. Methods and results are described in [15].

Assuming that all error types in the classification scheme affect fluency, we initially compared the total number of errors per system with human fluency scores. We then removed all error categories that were considered unlikely to have an affect on adequacy (such as ‘inappropriate’ items with a clear meaning, unnecessary items, inappropriate prepositions and determiners, omitted determiners, incorrect positions of words, spelling errors, case errors and incorrect verb tense/mood or conjugation, the majority of these being an inappropriate present tense in English). The remaining classification of adequacy errors was then compared with the adequacy scores from our human evaluations, as shown in Table 2.

**Table 2.** Human scores and error counts for fluency and adequacy

System	Mean human fluency score and rank	Number of fluency errors and rank	Mean human adequacy score and rank	Number of adequacy errors and rank
Systran	3.519 (1)	1015 (1)	4.136 (2)	127 (1)
Reverso	3.466 (2)	1020 (2)	4.142 (1)	132 (2)
Comprehium	3.221 (3)	1195 (3)	4.013 (3)	161 (3)
FreeTranslation	2.827 (4)	1460 (4)	3.644 (4)	287 (4)

Human fluency scores and the number of annotated fluency errors rank all four systems in the same order. The picture is slightly different for adequacy, with Systran and Reverso competing for the top position. We calculated Pearson’s correlation coefficient  $r$  between the human scores and the number of errors per system for each attribute. A very strong negative correlation was found between values: for fluency the value of  $r = -0.998$  and for adequacy  $r = -0.997$ . Of course, only four pairs of variables are taken into consideration here. Nevertheless, results show that we have man-

aged to capture adequacy as well as fluency by annotating errors without reference to the source text.

## 6 Correlating Error Frequency and Human Scores by Text Type

We computed error frequencies for the four text types. For each of the four systems, user manuals were annotated with the largest number of errors, followed by FAQs, technical reports and finally, press releases. The number of fluency errors and the subset of adequacy errors were then compared with human scores for fluency and adequacy according to text type. No significant correlation was found. In fact, human scores for fluency were highest for user manuals for all systems, yet these texts contained the largest number of annotated errors. It is clear, therefore, that errors must be weighted to correlate with intuitive human judgements of translation quality. The two main reasons for the large number of errors annotated in the user manuals were (i) the high frequency of compound nouns (eg. computer interface items and names of software applications), which, in many cases, were coded with two error types (eg. inappropriate translations and word order) and (ii) the high number of inappropriately translated verbs, which although understandable in the majority of cases, were not correct in the context of software applications (eg. *leave* instead of *quit* or *exit*, *register* or *record* instead of *save* etc.) Furthermore, user manuals were annotated with the largest number of untranslated words, yet many of these were understandable to evaluators with no knowledge of French, having little or no adverse effect on adequacy scores. A further experiment showed that 58% of all untranslated words in this study were correctly guessed in context by three people with no knowledge of French. In fact, 44% of these words, presented in the form of a list, were correctly guessed out of context.

## 7 Selecting Error Types for Automated Detection

The eight most common error types (from a total of 58 main categories) were found to be the same for all four systems, although the order of frequency differed between systems and text types. The frequency of these eight errors represents on average 64% of the total error count per system.

Table 3 shows that only in the case of inappropriate verbs (2) and inappropriate prepositions (6) does the total number of errors correspond to the rank order of the four systems according to human scores for fluency. The number of inappropriate noun string content errors (7) corresponds to human rankings for adequacy. Furthermore, the frequency of very few error types in the entire scheme corresponds to human rankings of the four systems for either fluency or adequacy. It is also clear from Table 3 that the frequency of particular errors within a given text type does not represent system performance as a whole.

Findings show that, while the frequencies of the above eight error types are significant, detecting a small number of errors to predict scores for a particular text type or system is not sufficient. Quality involves a whole range of factors – many of which must be represented in our automated system. Furthermore, our intention is to build a

tool that will provide information on error types to help users and developers, rather than merely a mechanism for producing a raw system score. It is clear, therefore, that a number of different error categories should be selected for detection, based on their combined frequencies, and on their computational tractability; we still need to determine which error types could be detected more successfully.

**Table 3.** Top eight error types by system

Error type	Number of errors annotated			
	Systran	Reverso	Comprend	FreeTrans
(1) Incorrect compound nn sequence	130	145	151	148
Manuals/FAQs/Press/Reports	56/21/39/14	58/30/42/15	64/28/41/18	62/30/40/16
(2) Inappropriate verb	121	126	135	141
Manuals/FAQs/Press/Reports	51/31/20/19	38/42/23/23	47/39/22/27	48/46/22/25
(3) Unnecessary determiner	105	102	137	121
Manuals/FAQs/Press/Reports	31/14/18/42	28/15/18/41	39/20/24/54	31/17/19/54
(4) Inappropriate noun	77	82	79	105
Manuals/FAQs/Press/Reports	17/20/22/18	16/14/24/28	19/18/24/18	18/24/29/34
(5) Incorrect verb tense or mood	76	56	103	90
Manuals/FAQs/Press/Reports	29/30/11/6	12/24/13/7	40/42/14/7	25/35/13/17
(6) Inappropriate preposition	73	77	84	89
Manuals/FAQs/Press/Reports	27/21/8/17	23/19/12/23	24/22/18/20	28/28/15/18
(7) Inappropriate.nn string content	48	38	69	82
Manuals/FAQs/Press/Reports	23/9/10/6	11/8/17/2	26/18/19/6	35/13/27/7
(8) Inappropriate adjective	48	37	59	42
Manuals/FAQs/Press/Reports	7/8/13/20	7/6/8/16	8/8/17/26	5/11/7/19

## 8 Conclusions and Future Work

We have devised an adaptable fluency error categorization scheme for French-English MT output, which also enables the detection of adequacy errors, without access to the source text. Preliminary analyses show that the number of errors annotated per system correlates with human judgments for fluency, and that a sub-set of error categories correlates with human judgments for adequacy. The annotated MT output has provided us with valuable information for the design of an automated MT evaluation system to help users and developers. Statistics are enabling us to identify the weak points of participating systems, and findings show that we must aim to automatically detect a good number of these to represent system performance.

Future work will involve:

- investigating inter-annotator agreement, as error annotation is subjective (for some categories more than others);
- the subsequent investigation and evaluation of methods for automating the detection of selected errors using machine-learning techniques on the annotated and part-of-speech tagged corpus;
- investigating correlations between human judgments and error type/frequency within texts;
- research into error weighting;

- the classification of errors by relative difficulty of correction during post-editing and/or by the possibility of correction by updating user dictionaries (although not appropriate for online MT systems);
- expanding the scheme to accommodate additional source languages translated into English;
- producing detailed documentation on the error categorization scheme, to include a full tag-set with examples;

## References

1. White, J.S.: How to evaluate machine translation. In Somers, H. (ed.): *Computers and translation: a translator's guide*. J. Benjamins, Amsterdam Philadelphia (2003) 211-244
2. FEMTI: A Framework for the Evaluation of Machine Translation in ISLE: <http://www.issco.unige.ch/projects/isle/femti/> (2004)
3. Papineni, K., Roukos, S., Ward, T., Zhu, W.: *Bleu: a Method for Automatic Evaluation of Machine Translation*. IBM Research Report RC22176. IBM: Yorktown Heights, NY (2001)
4. Akiba, Y., Imamura, K., Sumita, E.: Using multiple edit distances to automatically rank machine translation output. In: *Proceedings of MT Summit VIII, Santiago de Compostela, Spain (2001)*
5. Akiba, Y., Sumita, E., Nakaiwa, H., Yamamoto, S., Okuno, H.G.: Experimental Comparison of MT Evaluation Methods: RED vs. BLEU. In: *Proceedings of MT Summit IX, New Orleans, Louisiana (2003)*
6. Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H., Yamamoto, S.: Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In: *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC), Las Palmas, Canary Islands, Spain (2002)*
7. White, J., O'Connell, T., O'Mara, F.: The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. In: *Proceedings of the 1994 Conference, Association for Machine Translation in the Americas, Columbia, Maryland (1994)*
8. Rajman, M., Hartley, A.: Automatically predicting MT systems rankings compatible with Fluency, Adequacy or Informativeness scores. In: *Proceedings of the Fourth ISLE Evaluation Workshop, MT Summit VIII, Santiago de Compostela, Spain (2001)*
9. Rajman, M., Hartley, A.: Automatic Ranking of MT Systems. In: *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC), Las Palmas, Canary Islands, Spain (2002)*
10. Vanni, M., Miller, K.: Scaling the ISLE Framework: Validating Tests of Machine Translation Quality for Multi-Dimensional Measurement. In: *Proceedings of the Fourth ISLE Evaluation Workshop, MT Summit VIII, Santiago de Compostela, Spain (2001)*
11. Vanni, M., Miller, K.: Scaling the ISLE Framework: Use of Existing Corpus Resources for Validation of MT Evaluation Metrics across Languages. In: *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC), Las Palmas, Canary Islands, Spain (2002)*
12. White, J., Forner, M.: Predicting MT fidelity from noun-compound handling. In: *Proceedings of the Fourth ISLE Evaluation Workshop, MT Summit VIII, Santiago de Compostela, Spain (2001)*
13. Reeder, F., Miller, K., Doyon, K., White, J.: The Naming of Things and the Confusion of Tongues. In: *Proceedings of the Fourth ISLE Evaluation Workshop, MT Summit VIII, Santiago de Compostela, Spain (2001)*

14. Elliott, D., Hartley, A., Atwell, E.: Rationale for a multilingual corpus for machine translation evaluation. In: Proceedings of CL2003: International Conference on Corpus Linguistics, Lancaster University, UK (2003)
15. Elliott, D., Atwell, E., Hartley, A.: Compiling and Using a Shareable Parallel Corpus for Machine Translation Evaluation. In: Proceedings of the Workshop on The Amazing Utility of Parallel and Comparable Corpora, Fourth International Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal (2004)
16. SAE J2450: Translation Quality Metric, Society of Automotive Engineers, Warrendale, USA (2001)
17. American Translators Association, Framework for Standard Error Marking, ATA Accreditation Program, <http://www.atanet.org/bin/view/fpl/12438.html> (2002)
18. Correa, N.: A Fine-grained Evaluation Framework for Machine Translation System Development. In: Proceedings of MT Summit IX, New Orleans, Louisiana (2003)
19. Flanagan, M.: Error Classification for MT Evaluation. In: Technology Partnerships for Crossing the Language Barrier, Proceedings of the First Conference of the Association for Machine Translation in the Americas, Columbia, Maryland (1994)
20. Loffler-Laurian, A-M.: Typologie des erreurs. In: La Traduction Automatique. Presses Universitaires Septentrion, Lille (1996)
21. Roudaud, B., Puerta, M.C., Gamrat, O.: A Procedure for the Evaluation and Improvement of an MT System by the End-User. In: Arnold D., Humphreys R.L., Sadler L. (eds.): Special Issue on Evaluation of MT Systems. Machine Translation vol. 8 (1993)
22. Van Slype, G.: Critical Methods for Evaluating the Quality of Machine Translation. Prepared for the European Commission Directorate General Scientific and Technical Information and Information Management. Report BR 19142. Bureau Marcel van Dijk (1979)