Source Text Characteristics and Technical and Temporal Post-Editing Effort: What is Their Relationship?

Midori Tatsumi
School of Applied Language and Intercultural Studies
Dublin City University
Glasnevin, Dublin 9
Ireland
midori.tatsumi2@mail.dcu.ie

Johann Roturier
SES EMEA Research
Symantec Ltd.
Ballycoolin Industrial Estate
Blanchardstown, Dublin 15
Ireland
johann_roturier@symantec.com

Abstract

This paper focuses on the relationship between source text characteristics (ambiguity, complexity and style compliance) and machine-translation post-editing effort (both temporal and technical). Post-editing data is collected in a traditional translation environment and subsequently plotted against textual scores produced by a range of systems. Our findings show some strong correlation between ambiguity and complexity scores and technical post-editing effort, as well as moderate correlation between one of the style guide compliance scores and temporal post-editing effort.

1 Introduction

In the last few years, Machine-Translation post-editing has clearly become mainstream with more and more language service providers offering this type of activity as part of their range of services. However, the Post-Editing (PE) work to be performed is not yet fully understood, as shown by the recent creation of a dedicated Post-Editing Service Level user group.[1] Besides, production-ready post-editing environments are currently not optimized.[2] They tend to fall into two categories: recycled translation editors or native Machine Translation (MT) application clients. At best, the former connect to MT system(s) and retrieve raw translated strings, while the latter display system-specific information without necessarily interfacing with other systems.

In the first type of post-editing environment, there is a clear lack of source-related knowledge for post-editors apart from a traditional fuzzy match metric, the presence of inline tags, and potential terminology hits. Everything else has to be "computed" implicitly by the post-editor on-the-fly when glancing at or reading the text; for example, is this a long sentence? Does it contain multiple clauses? Does it contain spell-checking errors?

While adding extra information about a given translation unit may clutter the second type of post-editing environment (especially if it is not configurable), Blanchon et al. (2009) report that finding "good ways to compute scores reflecting the usefulness for post-edition of individual pre-translations of the text to translate" is an open research issue. Such scores may indeed help post-editors prioritize their work especially when working under severe time constraints. Rather than simply sorting on the status of a given translation unit (for-review, raw, verified, etc…), it may be useful to present scores (possibly using source text characteristics) to estimate how much time would have to be spent working on a particular segment (as long as a correlation exists between these source characteristics and the time spent post-editing). Depending on their level of experience, post-editors may also prefer working on certain types of segments (short segments or segments that are not too complex), rather than working from a complete document from start to finish (which is less fre-

[1] http://www.linkedin.com/groups?home=&gid=3056423
[2] Some experimental research prototypes also exist (such as Caitra (http://tool.statmt.org/) or SECTra_w (http://eolss.imag.fr/xwiki/bin/view/Main/)

Ventsislav Zhechev (ed.): *Proceedings of the Second Joint EM+/CNGL Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry" (JEC '10)*, pp. 43–51. Denver, CO, 4 November 2010.

quent nowadays with the fragmentation and parallelisation of translation tasks).

Several reasons exist for the lack of data on source text characteristics in post-editing environment:

• Lack of tool interoperability: some values generated by one system may not be exportable, importable or visualizable.

• Lack of system openness: it may be difficult to re-generate some of the values produced by specific systems (some of them may be proprietary and therefore inaccessible to users; some of them may not have APIs).

• Lack of transparency: it may be decided that certain values generated by one system should not be presented to other stakeholders.

Some of these gaps, which have been described in Lewis et al. (2009), suggest that more research work is required in identifying source text characteristics that can be linked to post-editing activity in order to make the post-editing task more efficient (and possibly enjoyable).

In this paper, we report the results of an analysis which aimed at exploring the relationship between source text characteristics and post-editing effort. We report findings on whether characteristics such as ambiguity, complexity and style guide compliance correlate with a traditional MT evaluation metric as well as post-editing time.

The rest of the paper is organized as follows: we briefly introduce related research in Section 2, and describe the methodology of our user study in Section 3. We present and discuss the results in Sections 4 and 5 respectively. Section 6 concludes and points out avenues for future research.

2 Related Work

Several strands of research are related to the present work. The first one concerns the identification of translatability indicators (or negative translatability indicators) and their impact on post-editing activity.

In Underwood and Jongejan's implementation of a translatability index (2001), two sets of translatability indicators are used: a set of phenomena identified in others' work on translatability, "including a) structural ambiguity caused by PP-attachment, relative and other sub-clause attachment and multiple coordination b) compounds comprising 3 or more nouns, c) "sentences" without (finite) verbs, d) lexical ambiguity and e) sentence length (both very long and very short sentences)", as well as a set of MT system-specific indicators. However, no empirical results have been published to indicate a potential correlation between the score computed from these phenomena and the subsequent post-editing effort.

This contrasts with O'Brien's study, who used an IT user guide translated into German by the IBM WebSphere MT engine (O'Brien, 2006) to measure the effect of CL rules on temporal, technical and cognitive post-editing effort (cf. Krings, 2001), using professional translators. Her findings were that post-editing effort can be decreased by suppressing Negative Translatability Indicators (NTIs) from the source text. This study also found that the removal of some NTIs had a greater impact on post-editing effort than the removal of others.

Another strand of research concerns the modeling of translation recommendations, such as the approach proposed in (He et al., 2010a). This method is based on a "Support Vector Machine classifier using features from the SMT system, the TM and additional linguistic features to estimate whether the SMT output is better than the hit from the TM" (and therefore easier to post-edit). Evaluation results for the English–French language pair will be presented in (He et al., 2010b). This work differs from the present work because the linguistic features used are limited to "source-side language model score and perplexity and a pseudo-source fuzzy match score".

A final strand of research concerns the design of MT confidence estimation measures that should be useful in a TM environment, such as (Specia et al., 2009a), by improving confidence measures for MT by training regression models to perform confidence estimation on scores assigned by post-editors. While the method described in this work has not been directly tested using post-editors, it has shown that its predicted quality estimate correlates better with human scores than reference-based MT evaluation metrics (Specia et al., 2009b).

The present study, which focuses on the English–Japanese language pair, does not try to predict whether some sentences are going to take less time to post-edit. Rather it tries to analyze the relationship between the post-editing effort and source text characteristics.

3 Methods

3.1 Test set

The source text chosen for this study was extracted from a user manual of a software publisher (Symantec), consisting of 3,916 English words in 269 sentences, which was machine translated into Japanese.

The English source text was in XML format, and written according to the controlled language rules used at Symantec, though the possibility of having uncontrolled sentences cannot be ruled out. Machine translation was performed in three steps: 1) pre-processing by using pre-processing scripts, 2) translating using Systran version 6, and 3) post-processing by using post-processing scripts. The pre-processing scripts included commands to make the source text more amenable to machine translation, such as protection of XML tags. The post-processing scripts included mainly commands that perform repetitive editing in the target text including the deletion of unnecessary spaces and personal pronouns, correction of style and expressions, such as inappropriate endings and misuse of polite and non-polite forms, and replacement of punctuations, counters, and other lexical items that are constantly inappropriately translated and difficult to be controlled by user dictionaries. In using Systran, general dictionaries and Symantec's product-specific user dictionaries were activated to ensure customised translation.

3.2 Post-editing

The MT output was post-edited by means of SDL Trados Translator's Workbench and TagEditor by nine Japanese professional translators; seven of them had experience in post-editing IT-related documentation, one in non-IT-related documentation, and one had no experience in post-editing. Participant post-editors were provided with brief PE guidelines that emphasized that PE should only be performed to make the MT output convey the correct meaning of the source text, and conform to Japanese grammar.

3.3 Scoring systems

The scoring of source text characteristics was performed using the following three software programs.

Systran: complexity and ambiguity

Systran version 6 offers a function that measures the syntactic complexity and lexical ambiguity of the source sentences. These metrics are provided to help the authors of the documentation to produce source text well-suited for translation by Systran.

The complexity metric takes into account a number of aspects of the source text, including "the number of clauses, conjunctions, phrases in parentheses, prepositional phrases, sentence length, sentence type (question or declarative sentence) as well as multiple additional language-specific criteria" (SYSTRAN: p.141), and calculates the scores for each sentence; the lowest score is 1, and the higher it becomes the more complex the sentence is.

The ambiguity score is given based on the number of ambiguous words in a sentence. A word is considered ambiguous if it has a) multiple meanings, or b) multiple parts of speech, and the latter criteria has higher significance in Systran's scoring system. According to the user guide, a high ambiguity score "reflects poor User Dictionary coverage" and adding user dictionary entries help to reduce the ambiguity in most cases. (SYSTRAN: p.141).

After the Deadline: style

After the Deadline is an open-source technology offered by Automattic Inc. We used the version of its API[3] available on August 10th 2010 with no customization. It offers three language checking functions: spelling, style, and grammar.

The spell checker finds misspellings as well as the words whose spelling is correct but possibly inappropriate in the context. The style checker reviews the document against Plain English[4] and detects complex phrases, passive voice, nominalisations, phrasal redundancy, etc. to help the author write clearly and concisely. The grammar checker spots repetition, disagreement of auxiliary verbs, disagreement of determiners, etc. to prevent common grammatical errors.[5]

In the present study, only the style checker was employed for the following reasons. 1) The test set did not include any true spelling errors; the spell checker detected XML placeholders, legitimate IT

[3] http://service.afterthedeadline.com/
[4] http://www.plainlanguage.gov/
[5] http://www.afterthedeadline.com/features.slp

terms, for example, "Ctrl", and 'misused' words that were appropriate in the specific context of our test set. 2) No grammatical errors were detected.

acrolinx IQ: grammar and style

acrolinx IQ supports controlled authoring by checking the document against a defined set of terms and rules to minimize ambiguity and promote consistency in English source content. We used the full set of controlled language rules used at Symantec. acrolinx IQ reviews the source text in terms of grammar and stylistic appropriateness (Bredenkamp et al., 2000), and assigns *flags* to indicate the absolute number of detected problems, and *scores*, which is the normalized values of flags in relation to the sentence length. We employed the flag count for two reasons: 1) ease of analysis, since scores were distributed in a heavily skewed manner, as 84% of the sentences were problem free (score 0), and the rest was scattered in the range from 250 to 5,000, while even though the distribution of flags was also skewed, it was milder compared to the scores as flag counts fell in the range from only zero to four, 2) suitability for sentence level analysis, since the scoring mechanism is designed to be more appropriate for document level analysis.[6]

3.4 Analysis method

We examined the relationship between these scores and the amount of PE effort from two aspects: *technical* and *temporal*, following Krings' three aspects of PE effort: technical, temporal, and cognitive (Krings, 2001). We employed the textual difference between MT output and the post-edited product as a proxy for technical PE effort, and measured it using GTM (General Text Matcher) (Melamed et al., 2003, Turian et al., 2003) version 1.3 with exponent set at 1.2, which mildly penalizes the word order difference (Callison-Burch et al., 2007).[7] GTM was chosen among other automatic evaluation metrics as it proved to have higher correlation with Japanese PE speed than BLEU, NIST, and TER in a related study (Tatsumi, 2009). As the Japanese writing system does not insert spaces to mark the boundary of words, the text was tokenised by means of MeCab.[8] Temporal PE effort is represented by the PE speed (words/minute); the word count was provided by acrolinx IQ and the time data was obtained by means of SDL Trados Translator's Workbench with a custom macro. GTM and PE speed data were obtained for each sentence.

In analysing the results, we took into account the difference in sentence structures. All sentences in the test set were classified into three categories: *simple* sentence, *complex/compound* sentence, and *incomplete* sentence. Simple and complex/compound sentences were identified according to Leech's definition: a simple sentence contains only one clause, a compound sentence contains two or more clauses linked by coordination, such as 'and' and 'but', while a complex sentence contains one or more subordinate clauses (Leech, 2006). Additionally, an incomplete sentence was defined for the purpose of this study: textual fragments consisting of words and phrases that cannot stand alone as a complete sentence. Examples of each category taken from the test corpus are shown below.

Simple sentence:
- *Delete the item from the vault.*
- *An envelope with a paperclip indicates an email with one or more attachments.*

Compound sentence:
- *The shortcut is a direct link to the archived item, and it has the following icon.*

Complex sentence:
- *Select the items that XXX is processing.*
- *Put the item in the Restored Items folder in the mailbox that is specified in the Settings dialog box.*

Incomplete sentence:
- *File size*
- *For a file system vault:*
- *If there is more than one page of search results:*

[6] http://www.acrolinx.com/uploads/documents/doc-center/acrolinxIQSuite1.0/Plug-inUserGuides/EN/acrocheck%20for%20Word%20Plug-in%20User%20Guide.pdf

[7] While it has been reported that the smaller exponent results in a better correlation with human evaluation in terms of adequacy and the larger exponent results in a better correlation with human evaluation in terms of with fluency (Lin and Och, 2004), as a result of testing with different settings in the present study, it was found that exponent 1.2 had the highest correlation with PE speed.

[8] Developed by Kyoto University and NTT. Accessible from: http://mecab.sourceforge.net/

4 Results

The box plots in Figures 1 to 8 represent the distribution of average GTM scores or PE speed of nine post-editors by score categories. Although GTM scores can range from 0 to 1, average scores for individual post-editors all fell within the range of 0.4 to 1, thus the y-axes for GTM scores show only the applicable range. For PE speed, the average speed for post-editors fell within the range of 0 to 80 words/min.

The white line in each box shows the median value among nine post-editors, and the box represents the range of distribution in the interquartile range (IQR, the range between 25th and 75th percentile), which shows approximately the middle 50% of the data. The horizontal lines above and below each box show the highest and the lowest values within the range of 1.5 IQR above and below the IQR respectively. Any values outside this range are shown by dots. The number in parentheses under each category indicates the number of observations found in the category. We excluded from the analysis the categories that have only one observation for statistical validity reasons. The Spearman correlation coefficient (Woods et al., 1986) is shown in the upper right corner in each figure.

4.1 Systran: complexity

Figures 1 and 2 show the distribution of average GTM scores and PE speed of nine post-editors by Systran complexity score categories.

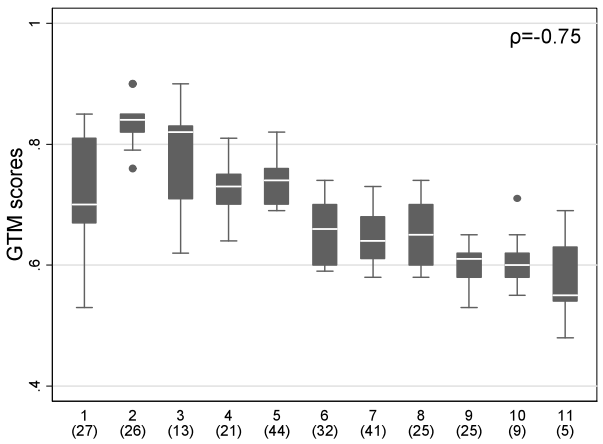

Figure 1. Systran complexity scores and GTM scores

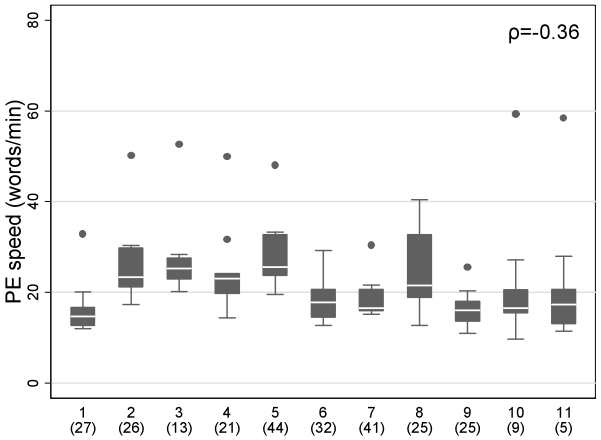

Figure 2. Systran complexity scores and PE speed

The Systran complexity scores have a clear negative relationship with the average GTM scores except for the score 1 category, while they have an indefinite relationship with the PE speed. The PE speed can be divided into three groups: the sentences in the score 1 category are slowest to post-edit, those in the score 2 to 5 categories are fastest, and the sentences in the score 6 to 11 categories are in the middle. This may partly be explained by the proportion of sentence structures in each score category; all 27 sentences in the score 1 category are incomplete sentences, and as the score increases, the proportion of simple sentences increases, and as the score increases further, the proportion of complex/compound sentences increases. Table 1 shows the overall average GTM and PE speed for all post-edited MT sentences by sentence structure. As can be seen, average GTM score is highest for simple sentences, and lowest for complex/compound sentences, while average PE speed is fastest for simple sentences and slowest for incomplete sentences.

	Average GTM score	Average PE speed (words/min)
Incomplete	0.73	19.02
Simple	0.74	25.92
Complex/Compound	0.65	22.42

Table 1. The effect of sentence structures on GTM and PE speed

4.2 Systran: ambiguity

Figures 3 and 4 show the relationship between Systran ambiguity scores and the GTM scores and the PE speed, respectively.

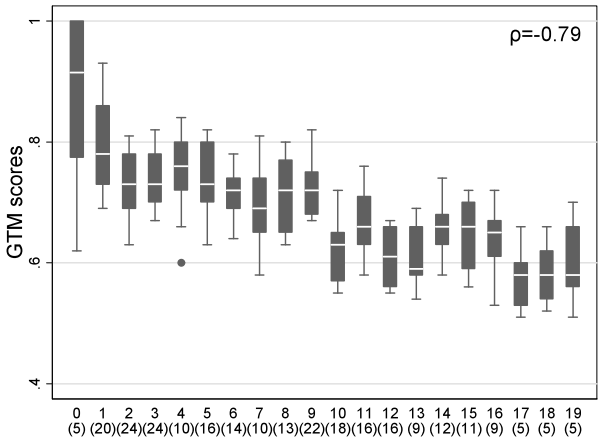

Figure 3. Systran ambiguity scores and GTM scores

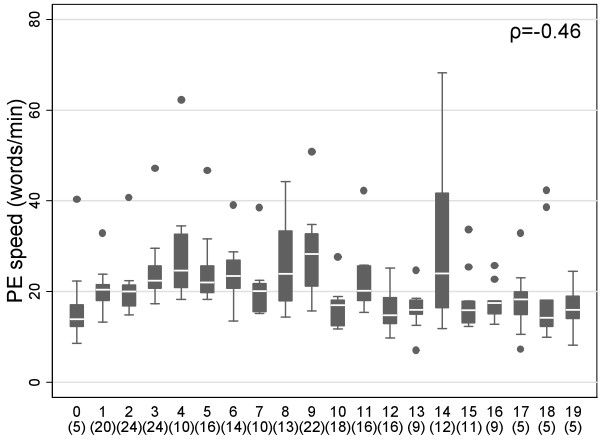

Figure 4. Systran ambiguity scores and PE speed

The Systran ambiguity scores, similar to the complexity scores, have a clear negative relationship with the GTM scores, and somewhat quadratic relationship with the PE speed. This, similar to the complexity scores, may have a relationship with the sentence structures; 92% of score 0 and 1 items are incomplete sentences, as the score becomes higher, the proportion of simple sentences gradually increases, and 96% of the items with score 10 or higher are complex/compound sentences.

4.3 After the Deadline: style

Figures 5 and 6 show the relationship between the After the Deadline style flag scores and the GTM scores and the PE speed, respectively.

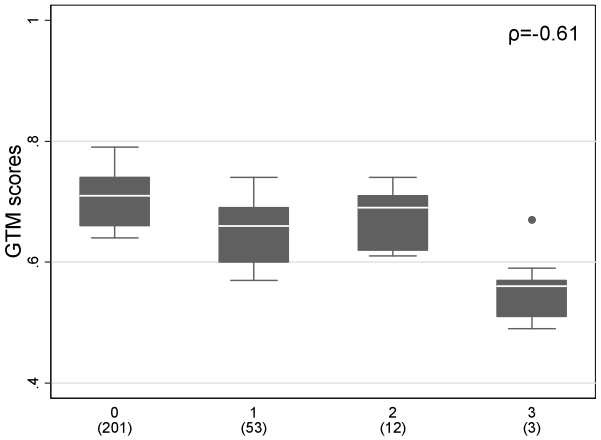

Figure 5. After the Deadline styles flag and GTM scores

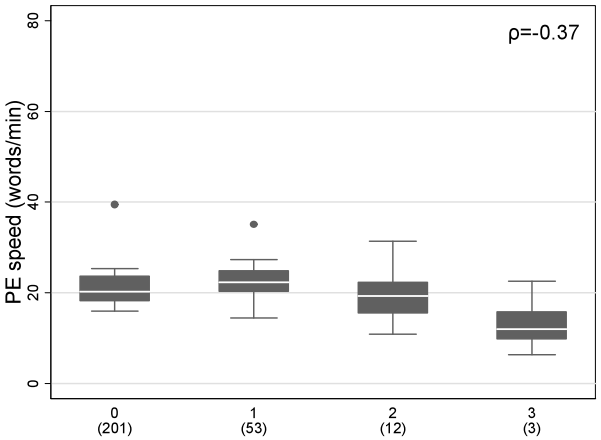

Figure 6. After the Deadline styles flag and PE speed

The After the Deadline style flag scores correlate somewhat negatively with both GTM scores and PE speed.

One of the reasons for the relatively slow PE speed for the score zero category may again be the ratio of sentence structures in each category. Table 2 shows the proportion of sentences in each structure in each score category. As can be seen, the ratio of incomplete sentences is exceptionally high for the score 0 items, which might have slowed down PE for the sentences in this category.

	0	1	2	3
Incomplete	49	4	0	0
Simple	55	5	0	0
Complex/Compound	97	44	12	3

Table 2. Distribution of sentence structure by score categories

4.4 acrolinx IQ: grammar and style

Figures 7 and 8 show the relationship between the acrolinx IQ flag scores and the GTM scores and the PE speed, respectively

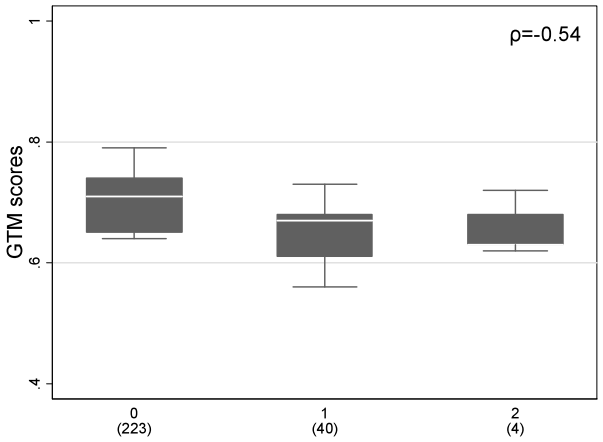

Figure 7. acrolinx IQ flag and GTM scores

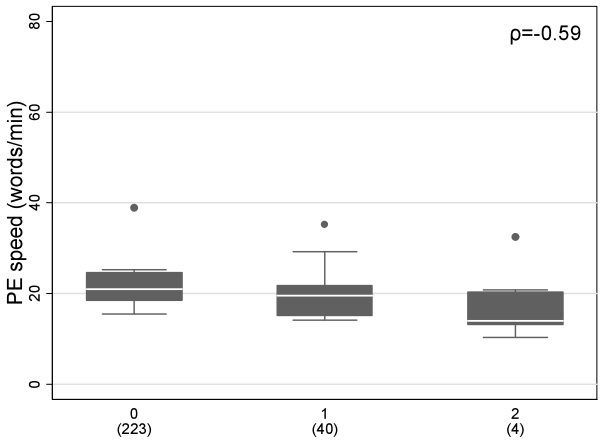

Figure 8. acrolinx IQ flag and PE speed

The biggest reason that over 80% of the sentences fall in the score zero category is the fact that the controlled authoring at Symantec is aided by acrolinx IQ, and in theory, all sentences should have been checked beforehand. Both the GTM scores and the PE speed have somewhat negative relationships with acrolinx IQ flag scores, though the distribution is small and the differences between the categories are rather small.

5 Discussion

Among all relationships examined, the strongest correlation can be observed between Systran complexity and ambiguity scores and the GTM scores, both of which are negative (Spearman correlation coefficient: ϱ=-0.75 and ϱ=-0.79 respectively). This relationship, however, may be related to the sentence lengths. Both Systran complexity and ambiguity scores have high correlations with the source sentence length (ϱ=0.90 and ϱ=0.87 respectively). Figure 9 shows the distribution of the average GTM scores for nine post-editors by sentence length categories. The sentences were categorised into groups according to the number of words contained: 1-5, 6-10, 11-15, 16-20, 21-25, and over 25 words. As can be seen, the source sentence length and GTM scores have a clear negative correlation (ϱ =-0.75).

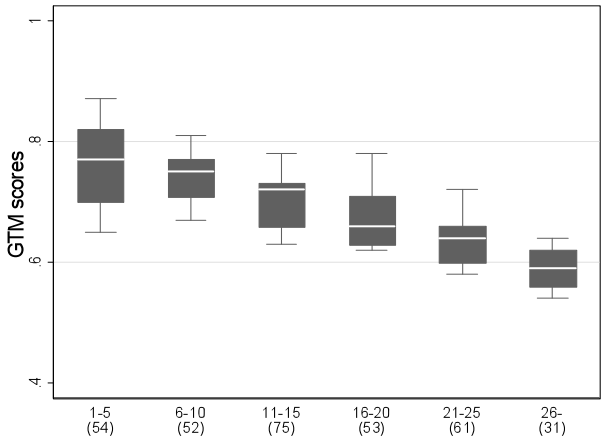

Figure 9. Source sentence length and GTM scores

The After the Deadline and acrolinx IQ scores also show negative relationships with GTM scores (ϱ=-0.61 and ϱ=-0.54, respectively), though the evidence is moderate with the values for most categories close to the overall average. This may partly be because most of the sentences are categorised as 'error free' by these scoring systems, and a few are given a small number of flags, thus the difference in average GTM scores or PE speed were not large enough to be clearly seen.

While the GTM scores had moderate to strong correlation with these scores, the PE speed did not show direct linear relationships with any of the tested scores except with acrolinx IQ (ϱ=-0.59). This may suggest that, though we are more interested in predicting the amount of temporal PE effort, these scores are more capable in predicting the amount of technical PE effort. However, the GTM scores and the PE speed have been proven to have moderate correlation (Pearson correlation coefficient r=0.56) (Tatsumi, 2009). A detailed an-

alysis revealed that some of the variance between the two can be explained by taking into consideration other source text characteristics, including the aforementioned sentence structures, the document component parts (for example. procedural sentences are faster to post-edit than other types of sentences), and the presence or absence of user interface terms. In addition, we cannot ignore the post-editors' individual differences. We found that both within and between post-editor variance is much higher in terms of PE speed compared to the amount of technical PE effort. This means that the amount of textual changes made during PE is more or less similar within and between post-editors, while the time taken to make the changes varies greatly both within and between post-editors.

6 Conclusions and Future Work

This work investigated the relationship between source text characteristics and technical and temporal post-editing effort for the English–Japanese language pair. Despite being limited to a small number of segments and one language pair, strong correlation was found between SYSTRAN's complexity and ambiguity scores and technical post-editing effort (using GTM scores), as well as moderate correlation between acrolinx IQ scores and temporal post-editing effort. This work could be extended by looking at larger data sets, more varied types of sentences (from a controlled language compliance perspective), additional source text characteristics (such as those described in Section 5), additional language pairs and possibly other types of systems.

In terms of future work, we suggest conducting studies to investigate how post-editors would interact with these scores if they were presented in their post-editing environment—would they find them useful? Would their usage vary based on the post-editor's experience? As discussed in the previous section, PE speed varies from one post-editor to the next, so a specific category of post-editors may benefit more from these scores (for example, post-editors with little experience).

We feel it would also be worthwhile to investigate whether these characteristics could be included as features in a translation recommendation system, such as the one mentioned in Section 2.

Finally we would like to make some recommendations for developers: tools or systems generating such scores should be designed in such a way that these values are leveraged by other systems. Existing standards (such as XLIFF) could easily accommodate such values using extension points to include application-specific information. Besides, future user interfaces should be flexible enough to allow users to display such scores in an intuitive manner, possibly using bookmarklets or extensions if working in a Web-based environment, or allowing for custom plug-ins to be easily created or extended if working in a desktop application scenario.

Acknowledgments

This work was made possible thanks to funding received from Enterprise Ireland and Symantec Corporation. The authors would also like to acknowledge the comments made by Sharon O'Brien on an earlier version of this paper.

References

Blanchon, Hervé, Christian Boitet and Cong-Phap Huynh. 2009. A Web Service Enabling Gradable Post-edition of Pre-translations Produced by Existing Translation Tools: Practical Use to Provide High-quality Translation of an Online Encyclopedia. In *Proceedings of MT Summit XII 2009, Beyond Translation Memories: New Tools for Translators Workshop*. pp. 20–27. Ottawa, Ontario, Canada.

Bredenkamp, Andrew, Berthold Crysmann, and Mirela Petrea. 2000. Looking for Errors: A Declarative Formalism for Resource-Adaptive Language Checker. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pp. 667–673. Athens, Greece.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) Evaluation of Machine Translation. In *Proceedings of The Second Workshop on Statistical Machine Translation*, pp. 136–158. Prague, 2007. Association for Computational Linguistics.

He, Yifan, Yanjun Ma, Josef van Genabith and Andy Way. 2010. Bridging SMT and TM with Translation Recommendation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10)*, pp. 622–630. Uppsala, Sweden.

Krings, Hans P. 2001. *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*. The Kent State University Press, Kent, OH.

Leech, Geoffrey. 2006. *A Glossary of English Grammar*. Edinburgh University Press Ltd., Edinburgh.

Lewis, David, Stephen Curran, Gavin Doherty, Kevin Feeney, Nikiforos Karamanis and Saturnino Luz. 2009. Supporting Flexibility and Awareness in Localisation Workflows. In *LRC XIV "Localisation in The Cloud": The 14th Annual Internationalisation and Localisation Conference*, Limerick, Ireland.

Lin, Chin-Yew and Franz Josef Och. 2004. ORANGE: A Method for Evaluating Automatic Evaluation Metrics for Machine Translation. In *Proceedings of CoLing 2004: 20th International Conference on Computational Linguistics*, University of Geneva, Switzerland, pp. 501–507.

Melamed, I. Dan, Ryan Green, and Joseph P. Turian. 2003. Precision and Recall of Machine Translation. In *Proceedings of HLT-NAACL 2003: Conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series*, Edmonton, Canada, pp. 61–63.

O'Brien, Sharon. 2006. *Machine-Translatability and Post-Editing Effort: An Empirical Study Using Translog and Choice Network Analysis*. PhD Dissertation. Dublin City University.

Specia, Lucia, Craig Saunders, Marco Turchi, Zhuoran Wang and John Shawe-Taylor. 2009a. Improving the Confidence of Machine Translation Quality Estimates. In *Proceedings of MT Summit XII 2009,* Ottawa, Ontario, Canada, pp. 136–143.

Specia, Lucia, Nicola Cancedda, Marc Dymetman, Marco Turchi, and Nello Cristianini. 2009b. Estimating the Sentence-Level Quality of Machine Translation Systems. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT '09)*, Barcelona, Spain, pp. 28–35.

SYSTRAN *SYSTRAN 6 Desktop User Guide*. SYSTRAN.

Tatsumi, Midori. 2009. Correlation between Automatic Evaluation Metric Scores, Post-Editing Speed, and some other Factors. In *Proceedings of MT Summit XII 2009,* Ottawa, Ontario, Canada, pp. 332–339.

Turian, Joseph P., Luke Shen, and I. Dan Melamed. 2003. Evaluation of Machine Translation and its Evaluation. *In Proceedings of MT Summit IX*, New Orleans, USA, pp. 386–393.

Underwood, Nancy L. and Bart Jongejan. 2001. Translatability Checker: A Tool to Help Decide Whether to Use MT. In *Proceedings of MT Summit VII: Machine Translation in the Information Age*, ed. Bente Maegaard, Santiago de Compostela, Spain, pp. 363–368.

Woods, Anthony, Paul Fletcher, and Arthur Hughes. 1986. *Statistics in Language Studies.* Cambridge University Press, Cambridge.