Machine translation errors and the translation process: a study across different languages

Michael Carl, Kent State University
M. Cristina Toledo Báez, University of Córdoba

ABSTRACT

The paper describes an experiment in which two groups of translators annotate Spanish and simplified Chinese MT output of the same English source texts (ST) using an MQM-derived annotation schema. Annotators first fragmented the ST and MT output (i.e. the target text TT) into alignment groups (AGs) and then labelled the AGs with an error code. We investigate the inter-annotator agreement of the AGs and their error annotations. Then, we correlate the average error agreement (i.e. the MT error evidence) with translation process data that we collected during the translation production of the same English texts in previous studies. We find that MT accuracy errors with higher error-evidence scores have an effect on the production and reading durations during post-editing. We also find that that from-scratch translation is more difficult for ST words which have more evident MT accuracy errors. Surprisingly, Spanish MT accuracy errors also correlate with total ST reading time for translations (post-editing and from-scratch translation) into very different languages. We conclude that expressions with MT accuracy issues into one language (English-to-Spanish) are likely to be difficult to translate also into other languages for humans and for computers – while this does not hold for MT fluency errors.

KEYWORDS

Translation quality assessment, machine error annotation, inter-rater agreement, post-editing, from-scratch translation, translation accuracy, translation effort, translation modes.

1. Introduction

The experiment described in this paper consists of two parts: in the first part, we conducted translation experiments with 48 Chinese and 32 Spanish translation students, each of whom translated 6 short English texts into simplified Chinese and Spanish respectively, in different translation modes (i.e. from-scratch translation, post-editing, monolingual post-editing, and sight translation). Translation process data (keystroke and gaze data with their production times) was recorded during the translation sessions and post-processed with the CRITT TPR-Data Base (Carl *et al.* 2016). In the second part of the experiment, the Google machine translation (MT) output into simplified Chinese and Spanish[1] of the 6 English texts was error-annotated by 16 Chinese translation students, who also participated in the translation experiment, and by 8 professional Spanish translators.

Unlike previous experiments in translation quality assessment (TQA), our annotators were asked to find minimal alignment groups (AGs) between the words in the source text (ST) and the MT output, the target text TT, and then assign pre-defined translation errors to the AGs, if applicable. Annotators proceeded in the following steps:

1. Align the ST and MT output (TT) in the most compositional and complete manner.
2. Mark the AGs obtained in step 1 with an error, if applicable.
3. Leave words un-aligned only if content was un-translated or added.

Annotators were asked to align ST and TT (i.e. MT output) words even if the TT contained wrong, missing or additional items, which would then be marked in step 2 as a translation error. For instance, the English word "sentence" has two meanings (a linguistic sentence and a punishment), which has two different realizations in Chinese and which was wrongly produced in the MT output. As it was obvious for the Chinese annotators to identify the ST token which had produced the wrong Chinese translation, they would first align the wrong lexeme with the English ST word (sentence) and then mark it, e.g. as "mistranslation." Error codes were also given to AGs if they contained fluency errors. ST words or TT words were only left unaligned when they were either unintelligible, omitted or content added.

In this way, it is possible to assess not only to what extent annotators agree on error codes (i.e. to identify erroneous words in the TT) but also whether they agree on ST-TT alignment groupings, i.e. to identify whether annotators agree on which (groups of) ST words are linked to which (groups of) TT words. All ST words in an AG would inherit the annotated MT error code, which allows for new possibilities to investigate translation quality across annotators and languages:

- As the simplified Chinese and Spanish MT output of the same English ST was annotated with the same error taxonomy, it is possible to assess whether the same ST words produce similar translation errors across different languages.
- As translation process data (e.g. translation production duration, fixation time, number of revisions, etc.) is available for a large number of post-edited and from-scratch translated versions of the same ST into various languages, it becomes possible to correlate types of MT errors with process data.

Consequently, we will try to answer the following research questions:

1. How strong is the inter-annotator agreement of MT error annotation for simplified Chinese and Spanish?
2. What is the effect of the MT errors on the post-editing duration and gazing time?
3. Are ST words with evident MT errors also more difficult to translate from scratch?

Previous work on a similar data set (Carl forthcoming; Carl and Schaeffer 2017) suggests that ST words or phrases with a larger number of possible

translations are harder to translate in post-editing as well as 'from-scratch' translation. Carl and Schaeffer (2017) hypothesized that translation problems may be due to the decision process which occurs when selecting 'the best' among several possible translation options. They show that the number of translation options within an SMT system correlates with the variation observed in human translations. Carl (forthcoming) shows that the number of different translations also correlates across very different languages. This suggests that MT systems and humans face similar decision-making problems for the same ST words across different languages, despite the fact that human translators have at their disposal a large repository of strategies which MT systems do not have, as Aragonés and Way (2017) rightly point out.

As an elaboration of these findings, this paper investigates how MT errors relate to post-editing behaviour and whether ST words that trigger erroneous MT output are also difficult to translate from scratch, across different languages.

2. Previous research on TQA

Evaluating MT output is essential for the development, improvement and fine-tuning of MT systems. MT researchers distinguish between reference-based MT evaluation and reference-less MT evaluation (Lavie 2011). In the former, the quality is measured by comparing the MT output against reference human translations. The most commonly used reference-based metrics, including BLEU (Papineni *et al.* 2002) and Meteor (Lavie and Agarwal 2007), measure the overlap of n-grams between the system and one or more reference translations. As Daems *et al*. (2017) point out, while from-scratch translations have traditionally been used as reference translations, MT researchers have recently deployed post-edited sentences as reference translations. Accordingly, novel metrics, such as HTER (Snover *et al.* 2006), measure the amount of required human post-editing as a proxy to assess MT quality.

Reference-less MT evaluation, also referred to as quality estimation (QE), implies evaluating a translation system's quality without access to reference translations (Bojar *et al*. 2015). Research in QE is focused on the design of features and the selection of learning schemes to predict translation quality, using source sentences, MT outputs, internal MT system information and source and target language corpora (Fomicheva *et al*. 2016). QE can be performed at different levels of granularity: sentence level and phrase levels (Bojar *et al*. 2016; Bojar *et al*. 2015; Shah *et al*. 2015) and document level (Scarton *et al*. 2016; Scarton and Specia 2014). Recently, there have been efforts to apply neural networks to QE (Bojar *et al*. 2016; Shah *et al*. 2015).

Human evaluation of MT errors has also become a focus of research. Error taxonomies for human evaluation are used on a word and phrase level and

on a sentence level. In sentence level assessments, two main methods are used: adequacy judgments and ranking judgments (Denkowski and Lavie 2010). With regards to word- or phrase-based assessment, several error taxonomies have been suggested (Vilar *et al*. 2006; Costa *et al*. 2015). Other widely used taxonomies include the Multilingual eLearning in Language Engineering project (MeLLANGE) (2006), the American Translators Association (ATA) grading rubric (Koby and Champe 2013) and the Multidimensional Quality Metrics (MQM) (Lommel *et al*. 2014a), which will be used in this study and are detailed below (see section 4.1.).

Apart from considering MT as a final product and evaluating its quality, MT quality assessment can also be used to investigate whether MT output is fit for post-editing (Denkowski and Lavie 2012). This perspective is becoming more attractive to both the translation industry and academia. An important concept in this context is post-editing effort, and how it can be minimized. Krings (2001) established the now recurrent three-fold definition of post-editing effort which involves a combination of cognitive, technical and temporal effort. Given that effort can be defined in many different ways, different measures have been used to measure it: time and effort (Koponen *et al*. 2012), fixation duration and number of fixations (Doherty and O'Brien 2009), pauses in post-editing (O'Brien 2006), number of editing events (Lacruz and Shreve 2014), etc.

Post-editing effort and MT quality are connected when it comes to establishing whether MT is fit for post-editing. MT quality has an impact on post-editing effort and, consequently, it is relevant to know the MT errors that have the highest impact on this effort. As Daems *et al*. (2017) explain, the estimation of MT quality has been approached by using human quality ratings ranging from 'good' to 'bad' (Doherty and O'Brien 2009; Koponen 2012) and error typologies (Koponen *et al*. 2012).

However, the comparative assessment of MT errors arising from the MT of one set of texts, and their relation to post-editing and from-scratch translation effort into different target languages has, to our knowledge, not been investigated. In this study we show that the cross-lingual analysis of MT errors may uncover hidden translation processes that are common to different modes of translation production.

3. Translation process data

This study makes use of the *multiLing* data set which is already available through the CRITT TPR-DB. This data set consists of six short English STs each of which has between 110 and 160 words. These texts were translated by multiple translators into six different target languages (TL) so as to obtain alternative translations for the same texts. Translation activity data was collected using Translog-II (Carl 2012) which allows the logging of keystrokes and gaze data from a connected eye tracker. Translations were performed in different translation modes including from-

scratch translation, post-editing, monolingual post-editing and translation dictation. Four texts (Texts 1-4) were news texts, two texts (Texts 5 and 6) were adapted from a sociological encyclopaedia with a total of 41 segments and 847 words for all six texts. The texts were displayed in large font (17 point Tahoma) and double spacing. Table 1 shows summary information for the six English source texts including the number of ST segments and words, and number of alternative translations in the data per text, target language and translation mode: from-scratch translation (T) and post-editing (P). The columns under the label "T-Words" show the number of ST words that have been translated into the different target languages, per translation mode and in total. The entire data set consists of 107867 ST words which were translated into six different languages. All translations were manually word-aligned and post-processed in the CRITT TPR-DB (cf. Carl *et al*. 2016). The data is freely available under a GPL license and can be downloaded via the CRITT TPR-DB web page[2]. Various versions of this data have been used in several studies (e.g. Carl forthcoming; Carl and Schaeffer 2017) and are presented there in more detail.

This paper focuses on English-to-Spanish and English-to-simplified Chinese post-editing and from-scratch translation. For the Spanish sessions, gaze data was collected with a Tobii T60 remote eye-tracker and for most of the Chinese translations with a Tobii 300TX. The average viewing distance aimed at was 50-60 cm from the screen, but no head or chin rest was used.

S-Text	1		2		3		4		5		6		**Total**			
Segs	11		7		5		5		6		7		**41**			
Words	160		153		146		110		139		139		**847**	T-Words		
Task	P	T	P	T	P	T	P	T	P	T	P	T	**Texts**	P	T	**Total**
da		24		23		22							**69**	0	10571	**10571**
de	8	7	7	8	8	8	8	8	7	8	8	8	**93**	6484	6616	**13100**
es	10	11	12	9	10	11	12	10	8	11	12	8	**124**	8996	8484	**17480**
hi	8	7	12	7	8	6	10	7	12	6	11	6	**100**	8581	5505	**14086**
ja	13	13	12	13	13	12	12	13	13	12	12	13	**151**	10609	10726	**21335**
zh	17	18	20	20	22	17	21	18	17	19	15	18	**222**	15750	15545	**31295**
Total	**56**	**80**	**63**	**80**	**61**	**76**	**63**	**56**	**57**	**56**	**58**	**53**	759	**50420**	**57447**	**107867**

Table 1. Information on the 6 English STs and their translations.

For the English-to-Spanish study (es), the first two texts were from-scratch translated, the next two texts were post-edited (MT output generated with Google translate in 2012), and the last two texts were monolingually post-edited (i.e. post-edited without access to the source text). Between 30 and 32 alternative English-to-Spanish translations were collected in 2012 in the framework of the CASMACAT project[3] for each of the six source texts – however, in this study we only consider the translation (T) and post-editing (P) sessions. The participants were part of

a translation class at the Universidad Autónoma de Barcelona (UAB), all of whom were Spanish native speakers.

For the English-to-simplified Chinese (zh), we merged data from three studies (MS12, RUC17 and STC17). As in the English-to-Spanish study, the texts were rotated, but the first two texts were always translated from-scratch, the next two texts were post-edited and the last two texts were monolingually post-edited (for MS12) or sight translated (translation was spoken) in the RUC17 and STC17 studies. The study MS12 was conducted in 2012 with translation students from Macau University. The MT output for the (monolingual) post-editing sessions was generated by a PBMT system. The studies RUC17 and STC17 were conducted in 2017 with NMT based on Google Translate's output from April 2017. The RUC17 data was collected in spring 2017 with 21 first-year Master and MTI students. The STC17 data was collected in autumn 2017 by 16 first-year Master and MTI students. From these studies we will only consider the from-scratch translation (T) and post-editing (P) tasks in this paper. All Chinese texts are produced in simplified Chinese, and all students but one were native speakers of Chinese.

4. Translation errors

This section first introduces the error taxonomy that we used for both annotation tasks, and then describes the annotation procedure. The Spanish and simplified Chinese MT output of the six texts was annotated by the 16 Chinese translation students in the STC17 study and by 8 professional Spanish translators respectively.

4.1. Error taxonomy

Chinese and the Spanish error annotations were based on the same MQM-derived error taxonomy. MQM (Lommel *et al*. 2014a) is a flexible hierarchical translation error taxonomy that can be tailored for different languages, texts and purposes. At its top level it makes a distinction – among other categories - between fluency errors and accuracy errors. Lommel *et al*. (2014a) explain that *fluency errors* are "related to the language of the translation, regardless of its status as a translation" while *accuracy errors* are "related to how well the content of the target text represents the content of the source." The MQM taxonomy[4] suggests how these error categories can be expanded into a finer grained hierarchical network and allows the use of only a subset of the entire error taxonomy. Errors can also be marked as "minor" or "critical", where "critical" would be appropriate for errors with a severe impact on understanding and/or fluent reading. All other errors are minor. Four of the errors in our taxonomy, highlighted in bold in Table 2, could be marked as "minor" or "critical" and four other error types have only one severity option. The adopted error taxonomy in Table 2 was developed and discussed amongst the 16 Chinese students in the classroom. They then also annotated the

simplified Chinese MT output as part of their homework. The same taxonomy was then used by the 8 Spanish translators for the Spanish MT output.

Accuracy	Fluency
Mistranslation	**Cohesion**
Omission/Addition	**Word Form**
Unintelligible	**Word Order**
	Punctuation
	Spelling

Table 2. Error types.

4.2. Error marking

The error annotation schema in our experiment was implemented in YAWAT (Germann 2008) as shown in Figures 1 and 2.

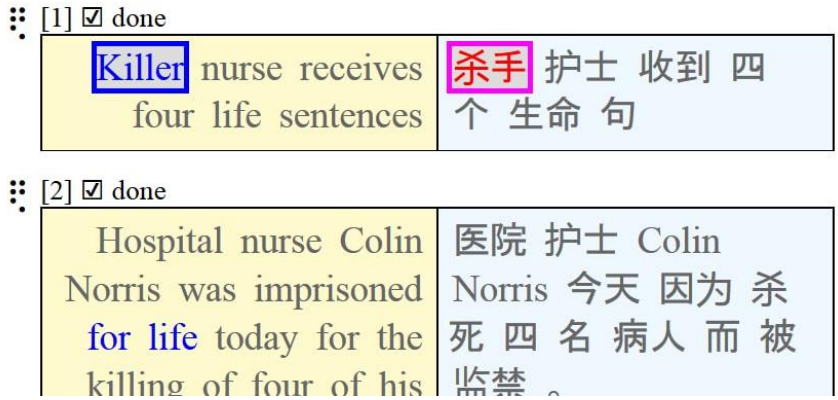

Figure 1. Error annotation in YAWAT.

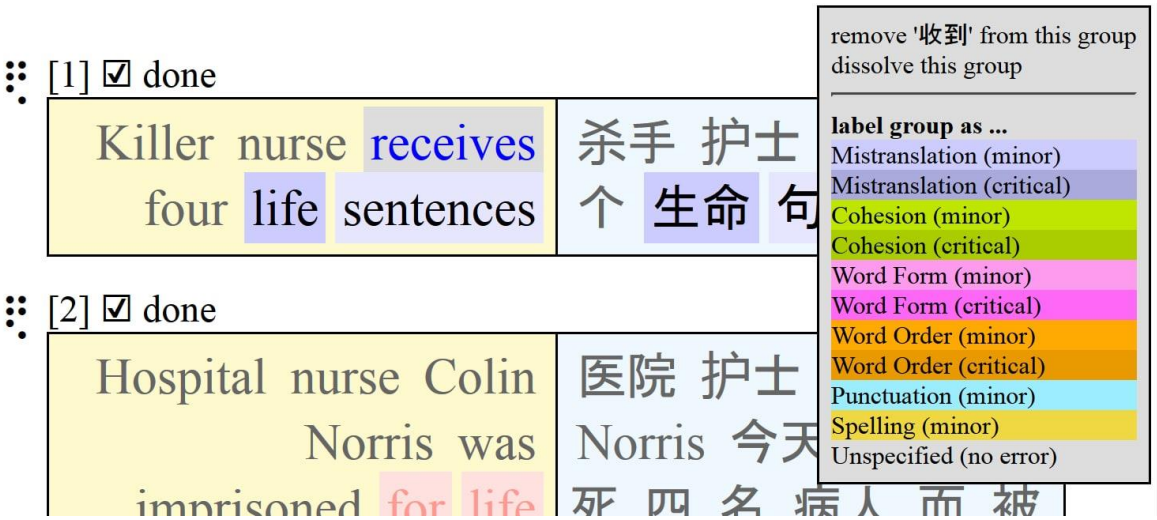

Figure 2. Error annotation in YAWAT.

YAWAT is a tool that can be used to manually align translation equivalences[5]. Every annotator was asked to align the six English STs described in section 3 with their respective Spanish or simplified Chinese MT output. Annotators were asked to fragment the translations into alignment groups (AGs), and then assign an error category to that AG, if applicable. In case words in the ST or in the TT could not be aligned, they could be marked as "Omission/Addition," if content from the ST was missing or added in the TT, or as "Unintelligible." The annotation schema allows for 13 annotation categories: 4 critical errors, 8 minor errors and the default case (no error). Annotators were introduced to the MQM and a discussion on how to produce the annotations was conducted.

5. Annotation agreement

From the 847 English ST words, 568 and 508 words were *not* error-annotated in the Spanish and simplified Chinese translation respectively. Table 3 (upper part) shows the distribution of the 12 error annotations across the 8 Spanish and 16 Chinese annotators for the remaining 279 and 339 words respectively. The columns are sorted by the number of total error annotations produced by the Spanish and Chinese annotators. There is a large variation in the number of annotated errors per annotator: between 40 to 156 errors per annotator for the Spanish professional translators and between 28 and 144 per annotator for the Chinese translation students. A total of 810 annotations were produced by the 8 annotators for the 279 error annotated words in the Spanish data, which amounts to an average of approximately 3 annotations per error-word. For the simplified Chinese data, 1155 total annotations were produced by 16 annotators, leading to an average of 3.4 annotations per error word.

Also, the choice of the error annotations is quite different: There are relatively more mistranslations (mistr) in the simplified Chinese MT output (35% minor and 36% critical error annotations) than in the Spanish data (18% minor and 23% critical mistranslations). In reverse, there are almost no annotations for punctuation (punct), spelling (spell) and word form (wform) errors in the Chinese data (3%), while these error categories make up 22% of the Spanish MT errors.

The lower part of Table 3 groups the error annotations in two different ways. It shows the number of critical (in **bold**) vs. minor errors and accuracy (*italics*) vs. fluency errors for each of the Spanish and Chinese annotators. There are slightly more minor errors than critical errors for both the Spanish (55%) and the simplified Chinese (59%) data. However, the distribution into accuracy and fluency errors is very different across both data sets. A total of 84% of the simplified Chinese MT errors are labeled as accuracy errors, while this is the case for only 47% of the Spanish error annotations.

	Spanish Annotators									Chinese Annotators															
Error	P02	P05	P07	P08	P01	P03	P06	P04	%	P03	P10	P04	P05	P13	P23	P19	P25	P01	P24	P11	P18	P09	P06	P02	P14	%
addom	5	3	1	8	7	7	7	5	0.05	1	7	4	10			23		25	10	18	16	10	22		6	0.13
cohes		1	1	2	8	9	26	22	0.09	3		1	1			3	6			3	13	3		12		0.04
cohesc	**1**	**1**		**5**	**8**	**5**	**20**	**4**	0.05							**2**	**10**		**6**	**5**	**3**	**4**		**4**		0.03
mistr	6	3	21	4	25	30	16	44	0.18	14	17	30	16	25	16	19	1	17	2	8	26	23	56	42	92	0.35
mistrc	**21**	**25**	**12**	**36**	**27**	**28**	**20**	**18**	0.23	**9**	**13**	**5**	**11**	**22**	**36**	**4**	**40**	**25**	**41**	**35**	**17**	**31**	**26**	**71**	**30**	0.36
punct			7		2		2		0.01											3						0.00
spell		2	2	2	7	7	3	8	0.04															3		0.00
unint					2		1		0.00		1									2						0.00
wform	1	13	11	15	13	15	10	17	0.12				3			2					2	11	4			0.02
wformc	**2**	**10**		**9**	**8**	**8**	**2**	**3**	0.05										**5**	**2**		**1**			**2**	0.01
word	2	1	9	2	1	6	7	17	0.06	1						1	3		4		6	5	11	6	14	0.04
wordc	**2**	**9**	**5**	**8**	**10**	**17**	**22**	**18**	0.11								**3**		**1**	**1**	**3**	**4**				0.01
Total	40	68	69	91	118	132	136	156	810	28	38	40	41	47	52	54	63	67	69	77	86	92	119	138	144	1155
Minor	14	23	52	33	65	74	72	113	0.55	19	25	35	30	25	16	48	10	42	16	34	63	52	93	63	112	0.59
Critical	26	45	17	58	53	58	64	43	0.45	9	13	5	11	22	36	6	53	25	53	43	23	40	26	75	32	0.41
Accurac	32	31	34	48	61	65	44	67	0.47	24	38	39	37	47	52	46	41	67	53	63	59	64	104	113	128	0.84
Fluency	8	37	35	43	57	67	92	89	0.53	4	0	1	4	0	0	8	22	0	16	14	27	28	15	25	16	0.16

Table 3. Summary of error annotation.

5.1. Error annotation

Table 4 shows the first 6 words (i.e. the title) of text 1 "Killer nurse receives four life sentences" and the output of the Google MT system into simplified Chinese "杀手 护士 收到 四 个 生命 句"[6]. In the first section under the header "Alignment groups", Table 4 shows two different ways in which the texts were aligned. The columns TT_1 and TT_2 show the aligned word(s) in the MT output while C_1 and C_2 indicate the respective number of annotators that have chosen this particular alignment. All 16 Chinese annotators aligned "Killer" and "nurse" with "杀手"and "护士" respectively. However, annotators disagreed on how to group minimal translation equivalence of the remaining words. From the 16 annotators, only participant P06 grouped together "receives four" with "收到 四 个". Accordingly, the two lines show this translation in the TT_2 column. The other 15 annotators aligned "receives" with "收到" and "four" with "四 个". Four annotators (among them also P06) aligned the compound "life sentences" with "生命 句" while the other 12 annotators chose to compositionally align "life" with "生命"and "sentences" with "句". It should be noted here that "句" is a severe lexical mistranslation; it is the translation of "sentence" in its linguistic reading, rather than a judgement. While annotators were advised to align also mistranslations with their likely source, and then tag the AG accordingly, is surprising that annotators chose to group together "生命 句," as it seems to make little sense. However, this has only been done by 4 out of 16 annotators[7].

Table 4 also shows the error annotations of the AGs for the first six Chinese annotators. Each column shows a translation error that is linked to an AG, an omission or an unintelligible segment. Three dashes "---" indicate the default case (i.e. no error). Some instances of inconsistency can be observed. For instance, while all annotators agree that there is a

mistranslation of "sentences", for some (e.g. participant P03) it is a minor error (mistr), while for the others it is a critical error (mistrc). More variation can be observed with respect to the translation of "life:" while most annotators agree that there is an issue with this translation, there is a larger variation as to whether annotators think it is a minor or a critical mistranslation ("mistr" and "mistrc" respectively) or whether it is a cohesion problem ("cohes"), while for annotator P03 there is no issue with "生命" as a translation for "life" in this context.

	Alignment groups				Error annotation						Translation Error Evidence (TEE)				
SToken	TT_1	C_1	TT_2	C_2	P01	P02	P03	P04	P05	P06	any	crit	min	acc	flu
Killer	杀手	16			---	---	---	---	---	---	0	0	0	0	0
nurse	护士	16			---	---	---	---	---	---	0	0	0	0	0
receives	收到	15	收到_四_个	1	---	---	---	---	---	mistr	0.1875	0.062	0.125	0.1875	0
four	四_个	15	收到_四_个	1	---	---	---	---	---	mistr	0.125	0	0.125	0.125	0
life	生命	12	生命_句	4	mistr	mistrc	---	cohes	mistrc	mistrc	0.75	0.5	0.25	0.625	0.125
sentences	句	12	生命_句	4	mistrc	mistrc	mistr	mistrc	mistrc	mistrc	1	0.875	0.125	0.937	0.062

Table 4. Annotation of the title of text 1 in simplified Chinese.

5.2. Error evidence

The last five columns in Table 4 show *translation error evidence* (TEE), representing the average annotator rating with respect to the different error categories introduced above. The error evidence was computed in two steps. First, the original error labels were mapped onto 0 or 1 as follows:

- *any*: all error annotations were mapped onto "1" and cells with no error annotation (i.e. "---") were mapped to "0."
- acc: accuracy errors mapped to "1", all other cells to "0"
- *flu*: fluency errors mapped to "1", all other cells to "0"
- *crit*: critical errors mapped to "1", all other cells to "0"
- *min*: minor errors mapped to "1", all other cells to "0"

The error evidence is then computed for each category in a second step as the average annotation error. Thus, in Table 4, the column *any* indicates the *any-error* evidence as the average over all 16 annotators for each word. Similarly, the columns *crit*, *min*, *acc*, and *flu* indicate the average of critical, minor, accuracy and fluency errors respectively, where it should hold that:

$$any = crit + min = acc + flu.$$

For example, all annotators agree that there is *no* issue with the translations of "killer" and "nurse." Accordingly, all columns have a value of 0. Conversely, all annotators agree that there is an issue with the

translation of "sentence". 14 out of the 16 annotators (87.5%) think of the translation as a critical error (*crit*: 0.875), while 12.5% are of the opinion that this is a minor error (*min*: 0.125). There is a high evidence score that this translation is an accuracy issue (*acc*: 0.9375) while only one of the 16 annotators is of the opinion that it is a fluency issue (*flu*: 0.0625). Similarly, 75% of the annotators think there is an issue in the translation of "life," 18.7% think the translation of "receives" has an issue and 12.5% think so for the translation of "four" and the average agreement as to whether it is a critical, minor, accuracy or fluency errors becomes accordingly lower.

	Chinese			Spanish		
	#Obs	TEE	Kappa	#Obs	TEE	Kappa
all	---	---	0.23	---	---	0.32
any	339	0.213	0.30	279	0.361	0.43
critical	173	0.170	0.27	126	0.358	0.45
minor	301	0.142	0.14	239	0.232	0.22
accuracy	281	0.216	0.31	172	0.275	0.35
fluency	131	0.087	0.04	173	0.308	0.35

Table 5. Scores of translation error evidence and fleiss kappa for error categories (described in section 5.4).

Roughly, 40% (339 words) and 32% (279) of the 847 ST words were annotated with at least one error annotation in the Chinese and Spanish translations respectively, i.e. with the value *any* > 0. Table 5 shows the error evidence scores for each error category. The column #Obs shows the number of words for which at least one annotator thinks that it contains an error of that category; column TEE shows the total average evidence scores of that error category > 0. The Table shows that the error evidence is stronger for the Spanish MT error annotations than for the simplified Chinese ones with respect to every error category. While for Spanish the average translation error evidence (i.e. the inter-annotator agreement that there is an MT error of a particular type) is roughly between 25% and 36%, it is much lower for simplified Chinese. The low score for Chinese fluency error annotations is particularly surprising: the TEE score of 0.087 indicates that for the 131 instances in which a Chinese fluency error was annotated, there were, on average, less than two out of the 16 annotators of the same opinion.

We suggest that the error evidence score can act as an indicator of the average error agreement. The more annotators rate a translation as erroneous, the more we can assume the annotated error to be evident (or obvious). For instance, the evidence of an *any-error* for the translation of "four" as in Table 4 may be considered inferior to that of the translation of "sentence", according to the ratio of annotators who rate it as such.

However, the error evidence is relative to the language pair, text and annotator group, and perhaps additional parameters. For instance, there is a higher agreement among the Spanish annotators than among the Chinese annotators. Despite the fact that the Chinese translation students developed and discussed the error taxonomy in some detail, their lower agreement might be due to them being less experienced than the Spanish annotators[8]. Whatever the reason is, we do not wish to conclude from this that the Spanish MT errors are more evident than the Chinese ones – however, we take it that they can be ranked as more or less evident within the same group.

5.3. Cross-lingual correlation of error evidence

	any	acc	flu	crit	min
any	0.14	***0.98***	0.42	***0.84***	***0.81***
acc	***0.72***	**0.22**	0.22	***0.84***	***0.77***
flu	***0.76***	0.10	-0.03	0.27	0.44
crit	***0.81***	**0.63**	**0.57**	0.14	0.36
min	***0.73***	0.47	**0.60**	0.18	0.11

Table 6. Correlation of error evidence.

The upper part above the diagonal in Table 6 shows the Pearson correlation of the English-to-simplified Chinese error evidence, below the diagonal shows the correlation of English-to-Spanish error evidence. As can be expected, there is a strong correlation between *any* and the various error sub-categories for both language translation pairs. It is also not surprising that the lowest correlation is between accuracy/fluency and critical/minor errors as they constitute orthogonal categories. However, the two error dichotomies, accuracy vs. fluency and critical vs. minor are more clearly separated in the Spanish data than in the simplified Chinese one. Thus, *crit* and *min* correlate 0.18 in Spanish versus 0.36 in Chinese; *acc* and *flu* correlate 0.10 for Spanish versus 0.22 in Chinese, indicating a greater confusion amongst the Chinese annotators.

The red numbers on the diagonal axis show the Pearson correlation between the Chinese and the Spanish error evidence of the same error category: there is (almost) no correlation for most of the categories, but interestingly accuracy error evidence scores between Chinese and Spanish MT output is slightly higher (0.22). This suggests that at least some of the difficulties related to accuracy errors in the MT output occur across Spanish and Chinese.

5.4. Inter-annotator agreement with kappa

We also computed inter annotator agreement[9] using the *kappam.fleiss* function in R which is provided with the irr package[10].

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

Equation 1. Kappa fleiss function.

The kappam.fleiss implementation in R allows the comparison of several annotators and the assessment of their agreement above chance. The Kappa score[11] is shown in Equation 1. The factor Pr(e) amounts to the agreement by chance and Pr(a) is the observed agreement, which can range, theoretically, between 0 and 1. Pr(a) – Pr(e) is then the degree of agreement achieved above chance and 1 – Pr(e) the degree of agreement that can be maximally achieved above chance. Since Pr(a) can also be smaller than Pr(e), the kappa score can be negative. A score κ=1 indicates perfect agreement between all the annotators and a score κ<= 0 random choice.

We computed the kappa scores for the Spanish and Chinese annotators individually by clustering the errors in five different error categories, *any*, *crit*, *min*, *acc*, and *flu* as described in the first step of the error evidence calculation above (section 5.2.). In addition, we also added an error category which contains all original error labels:

- *all*: all of the 12 errors were kept plus one 'uncritical' default label resulting in 13 annotation classes

The annotations for each word and each annotator were represented as 16 x 847 and 8 x 847 matrices for all source 847 words and the 16 Chinese and 8 Spanish annotation data respectively. Apart from the *all* category – which has 13 class labels – all other annotation matrices had only two classes, "0" or "1"[12]. The resulting kappa scores for Chinese (zh) and Spanish (es) are provided in Table 5. As with the TEE score discussed above, the Chinese annotators agree less among themselves than the Spanish ones.

Landis and Koch (1977) label kappa scores as "poor" (<0) "slight" (0.0 – 0.2), "fair" (0.2 – 0.4), "moderate" (0.4-0.6), "substantial" (0.6-0.8) and "perfect" (0.8 – 1). According to this classification, most error categories in Table 5 show a "fair" agreement. The "minor" category (kappa scores of 0.14 and 0.22) and Chinese fluency errors (0.04) have the least inter-annotator agreement. The *all* category (which has 13 classes) has a kappa value of 0.23 and 0.32 for Chinese and Spanish respectively, and only two Spanish categories (*any* and *critical*) show "moderate" agreement. There is almost a perfect correlation (R = 0.94) between the error evidence

(TEE) and the kappa scores which are shown in Table 5 (i.e. where non-error agreement, or TEE=0, is not considered).

5.5. Inter-annotator AG agreement

This section assesses kappa scores for word AGs. In order to compute the kappa scores for AGs, we proceeded in a similar manner as for the calculation of error annotation, described in section 5.2. We filled the "Error annotation" cells as in Table 4 with those target word IDs (tid) that each annotator had linked the ST word with. Table 7 shows the excerpt of the English-to-simplified Chinese data segment for annotators P01 to P06 that was already discussed in the context of Table 4 in the light of error annotation. The columns "SToken" and "Alignment groups" are identical to those in Table 4 and show the aligned ST-TT words and the number of these AGs. Column "Alignment group encoding" in Table 7 shows the encoding of the AGs. For instance, all annotators grouped the English word "Killer" with the first Chinese word "杀手", which results in the label "1" for all AG encodings of "Killer". Participant P06 grouped together "receives four" with Chinese "收到 四 个" which happen to be words 3, 4 and 5 in the translation, and which thus results in the encoding "3+4+5".

	Alignment groups				Alignment group encoding					
SToken	TT_1	Cnt_1	TT_2	Cnt_2	P01	P02	P03	P04	P05	P06
Killer	杀手	16	杀手	16	1	1	1	1	1	1
nurse	护士	16	护士	16	2	2	2	2	2	2
receives	收到	15	收到_四_个	1	3	3	3	3	3	3+4+5
four	四_个	15	收到_四_个	1	4+5	4+5	4+5	4+5	4+5	3+4+5
life	生命	12	生命_句	4	6	6+7	6	6	6	6+7
sentences	句	12	生命_句	4	7	6+7	7	7	7	6+7

Table 7. English to Chinese data segment for annotators P01 to P06.

The kappa score was then computed based on the AG encoding in Table 7. Surprisingly, the kappam.fleiss function returned 0.653 for Chinese and 0.405 for Spanish. That is, our data show a higher agreement for AGs than for *all*-error annotation. Chinese, with a much lower error annotation agreement (Table 5), has a much higher agreement in AGs than Spanish.

This finding seems counter-intuitive to us and we do not have a good explanation for this. Low agreement scores for translation error rating can be expected. Lommel *et al*. (2014b), for instance, report kappa scores between 0.18 and 0.36. Similar values have also been reported in various WMT evaluation reports. One would expect a correlation between AGs and error coding, as a possible reason for lack of annotator agreement might be a confusion of AGs: if annotators agree how ST tokens align with the MT output, they might also agree whether and which error label should be

assigned to that AG. However, this does not seem to be the case in our data. In order to cross-check and quantify this observation, we also correlate the entropy of AGs and the entropy of error labels. Entropy indexes the distribution of observed configurations (i.e. the range of different classifications observed in the data). However, here too, there is a low correlation of the entropy values between AG encoding and error annotation, 0.04 and 0.18 for Spanish and Chinese respectively.

6. MT error evidence and translation effort

In this section, we assess the relation of MT errors to the duration of post-editing and total reading time of ST tokens.

Within the CRITT TPR-DB, each ST word is coded as a line and associated with (currently) 59 features in the ST tables. These features describe, among other things, properties of the ST words and their translations, including the word translation entropy (*Htra*) which indexes the number of observed translation alternatives in the corpus of translations, the number of keystrokes (insertions and deletions) used to produce the translation, typing duration (*Dur*) measured for typing the translation, the total reading times of the source word (*TrtS*) and the target word (*TrtT*), etc.

6.1. Evidence of MT errors and translation effort

In order to investigate whether evidence scores assigned to MT errors (according to our definition above) have an effect on post-editing and whether they are correlated with translation duration, we first merged the five average error annotation values into the CRITT TPR-DB tables, so that each line indexing the same ST word contained in addition to the 59 features also the five error evidence scores *any*, *acc*, *flu*, *crit*, and *min* for that word. This allowed us to run various regression analysis, where we used *Dur*, *TrtS* and *TrtT* as dependent and the error evidence scores as predictor variables. For all the analyses, we used R (R Development Core Team 2014) with the built-in linear regression function (lm).

The average *any-error* evidence had a significant effect in both languages, Chinese (zh) and Spanish (es), on *TrtS* (zh:p=0.045, es:p<0.0001), *TrtT* (zh:p=0.022, es:p<0.0001) but less so on *Dur* (zh:p=0.0396, es:p=0.063). The effect was also significant when adding the length of the ST word (*StokLen*) as a control variable[13].

	Post-editing (P)				Translation (T)			
	Spanish		Chinese		Spanish		Chinese	
	Dur	TrtS	Dur	TrtS	Dur	TrtS	Dur	TrtS
acc	***	***	***	***	*	***	---	---
flu	***	---	---	---	---	---	.	---
crit	***	***	***	---	.	*	---	---
min	***	---	*	***	---	---	---	---

Table 8. Significance levels of the regression analysis.

We further tested the effect of the other four error categories on total *TrtS* and *Dur* independently for post-editing (P) and from-scratch translation (T) and for the two target languages. The predictors were (*acc*+*flu*) and (*crit*+*min*). Table 8 shows the significance levels of the analysis[14]. Figure 3 shows the corresponding 8 graphs for each of the two target languages.

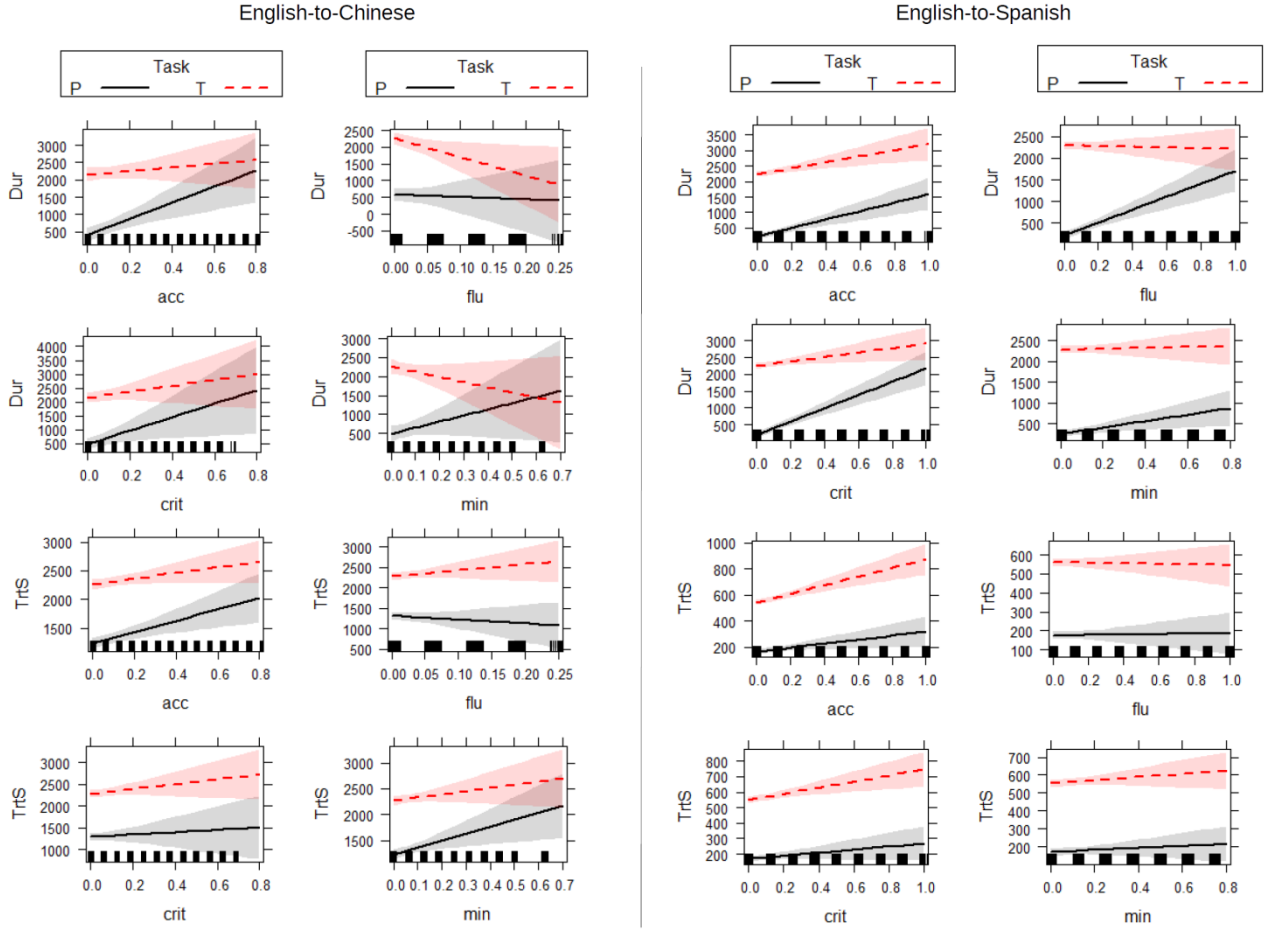

Figure 3. Effect of MT errors on production duration and ST reading times.

Most of the regression lines in Figure 3 show a positive trend between the independent variables, i.e. (*acc*+*flu*) and (*crit*+*min*), and the dependent ones, *Dur* and *TrtS*. However, as Table 8 shows, not all effects are significant. There are slight differences between the Spanish and the

Chinese data: all error categories have a highly significant effect on *Dur* in the Spanish data, which was not the case for *flu* in the Chinese data. Interestingly, and as one might expect, *acc* had also a significant effect on ST reading times (*TrtS*) in both languages, whereas *flu* did not. An explanation for this observation might be that, according to the definition, *acc-errors* are related to transfer problems, and thus may require cross-checking of the translation with the reference in the ST, while this is not the case for *flu-errors*, which can be solved in the translation without reference to the source.

6.2. MT errors and translation ambiguity

Another related observation is that Spanish *acc-errors* have also a significant correlation with ST reading time during from-scratch translation. This might indicate some common underlying problems in machine translation and from-scratch translation. In an earlier study, Carl and Schaeffer (2017) found that words (and sentences) which are difficult to translate for an MT system are also difficult in from-scratch translation. In line with Choice Network Analysis which "compares the renditions of a single string of translation by multiple translators in order to propose a network of choices that theoretically represents the cognitive model available to any translator for translating that string" (Campbell, 2000:215), they trace translation problems back, among other things, to word-translation ambiguities. Word-translation ambiguities reflect choices for rendering the TL that are measured as word translation entropy (*HTra*[15]) in the CRITT TPR-DB. Table 9 shows the correlation between *HTra* and the five error categories. The highest correlation is observed for *acc-errors* in the Spanish data; the lowest correlation for *flu-errors*. This observation confirms our previous finding in various ways. It suggests that a larger number of translation choices leads to increased (more evident) MT accuracy errors, but not so to more fluency errors. Provided this is true, it also shows that the Spanish annotators are more sensitive in distinguishing between accuracy and fluency issues during their annotations – which is less developed in our Chinese annotators. It is also surprising that, despite fair inter-annotator agreement of Spanish *acc-errors* (0.275 TEE agreement; kappa score of 0.35), an almost moderate correlation with *HTra* (0.48) can be measured.

	any	acc	flu	crit	min
Spanish	0.47	0.48	0.22	0.39	0.33
Chinese	0.35	0.33	0.18	0.27	0.31

Table 9. Pearson correlation between *HTra* and the five error evidence scores.

6.3. Cross-lingual effect of MT error evidence on translation effort

Carl (forthcoming) finds that *HTra* values correlate in translations across the six language pairs, English into Danish, Spanish, German, Hindi, Chinese and Japanese, shown in Table 1. He suggests that the reason might be that words that are difficult to translate into one language are likely also to be difficult to translate into another language. As shown above, higher *HTra* values relate to more evident accuracy errors, and to higher effort in from-scratch translation (i.e. translation production times, gaze duration). In Table 6 we also show that *acc-error* evidence scores from different MT systems into different languages correlate to some extent (R=0.22).

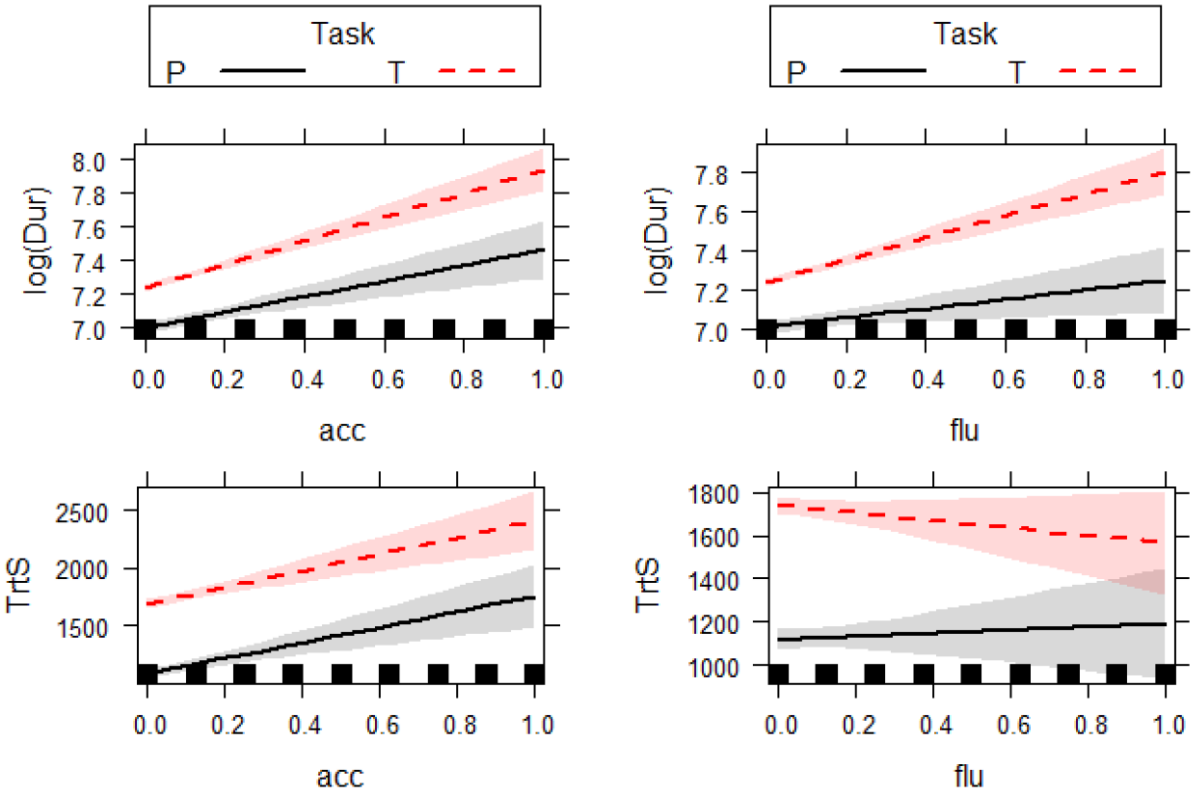

Figure 4. Effect of the Spanish *acc-errors* and *flu-errors* on the total reading times (TrtS) and log-transformed production duration (*log(Dur)*) for translating (T) and post-editing (P).

Here we correlate the Spanish *acc-error* and *flu-error* evidence scores with the total reading time (*TrtS)* and production time (*Dur*) for from-scratch translation (T) and post-editing (P) of the six language pairs discussed in Table 1. The evidence error scores were merged into all 107867 word translation records and several regression models were tested. Figure 4 shows some of the effect plots of the Spanish *acc-errors* and *flu-errors* on the *TrtS* and *log(Dur)*[16] for the two translation modes (P and T). There is a significant correlation of *acc-errors* and ST reading times and (log) translation production duration for post-editing and translation ($p<0.001$). There is also a significant correlation of *flu-errors* on *log(Dur)* ($p=0.0095$), but none on *TrtS* ($p=0.59$). For *TrtS* there is no interaction between the

two translation modes either for *acc-errors* (p=0.81) or for *flu-errors* (p=0.18). With respect to *TrtS* there is also no significant interaction between the six target languages; they all have a similar slope to the one shown in bottom left graph in Figure 4.

Dependent Variable	*log(Dur)*		*TrtS*	
Predictor Variable	acc	flu	acc	flu
Main effect	***	**	***	---
Task interaction	*	**	---	---
TL interaction	* - ***	* - ***	---	---

Table 10. Significance values for task and TL interaction models.

However, for *log(Dur)* there are various significant interaction effects between the two tasks and the six target languages, which are summarized in Table 10. The symbol "*-***' means that the significance level varies between 0.01<p<0.05 and p<0.001, depending of a particular TL. It thus appears that accuracy errors of MT output in one language (Spanish) can be used to predict the cognitive effort spent on ST reading during post-editing and during from-scratch translation of the same text into another language, while this seems to be less so the case for fluency errors.

7. Conclusions

The paper investigates inter-annotator agreement of Spanish and simplified Chinese MT errors and relates the error scores to post-editing and translation effort across several languages. Sixteen Chinese translation students and eight professional Spanish translators annotated MT output of the same English source texts into simplified Chinese and Spanish respectively, using the same MQM-derived error taxonomy. We compute an average error annotation score for several error classes and find that more evident MT errors lead to higher post-editing effort. In particular, critical and accuracy errors increase post-editing and ST fixation duration, as compared to minor and fluency errors. We suggest that accuracy errors are due to translation-ambiguities which are difficult to decide for both MT systems and human translators, and trigger longer from-scratch translation and post-editing times. With respect to the three initial research questions we conclude:

1a) we clustered the MT errors into five categories (any, accuracy, fluency, minor and critical errors) and computed the kappa scores and an average error evidence score. Both measures show i) fair agreement (most kappa scores between 0.20 and 0.40) and ii) that Chinese translation students agree less among themselves than professional Spanish translators. Despite the fact that the Chinese translation students developed and discussed the error taxonomy in some detail, their lower degree of

agreement might be due to a lack of translation experience as compared to the Spanish professional translators.

1b) surprisingly, the Chinese students agree more on the segmentation of the translations into alignment groups than the Spanish professional translators do. This finding contradicts Lommel *et al*. (2014b: 35) who find that "even though annotators largely agree on the existence of the problem, they often disagree on the location." However, in our data we found that there is a stronger agreement in alignment grouping (kappa score 0.653 for Chinese and 0.405 for Spanish) than in error labelling. That is, in our data annotators seem to agree more on the origin and the span of the error than on the nature of the error.

1c) we also examined to what extent the English ST produces similar MT errors in the Chinese and Spanish output. We found no correlation (r=-0.03) for fluency but a fair correlation (r=0.22) for accuracy errors into Chinese and Spanish. This suggests that MT accuracy errors may relate to ST difficulties, independent from the target language.

2) next, we investigated the effect of the MT errors on post-editing effort. We used production duration and ST gaze time as dependent and the MT error scores as predictor variables. Most of the error categories (accuracy, fluency, critical, minor) had a highly significant effect on post-editing duration, while only accuracy errors had a highly significant effect on ST reading times in both languages. An explanation for this observation might be that accuracy errors may require extended cross-checking of the ST reference, while this is not the case for fluency errors which can be solved in the translation without reference to the source.

3a) we also found that English ST words with evident Spanish accuracy errors require significantly longer ST reading times not only during post-editing but also during from-scratch translation. This indicates common problems in post-editing and from-scratch translation, which may be due to translation ambiguities (c.f. Carl and Schaeffer 2017). We therefore measured the translation ambiguity (*HTra*) of the English ST words, and found that accuracy errors correlate to a higher extent with *HTra* (r=0.48) than fluency errors (r=0.22).

3b) Maybe the most surprising finding in this study is the observation that the evidence scores of Spanish MT accuracy errors significantly correlate with patterns of translation behaviour of the same texts into other, very different languages (Danish, German, Japanese, and Hindi). This is consistent with other recent studies (Carl forthcoming; Carl and Schaeffer 2017) which find that ST words which are translation-ambiguous in one language (i.e. have many possible different translations) tend to be translation-ambiguous also in other languages. Given that ambiguous words are more difficult to translate for humans and machines alike than

less ambiguous words, it seems that translators and post-editors face similar translation problems – even for very different target languages.

References

- **Aragonés Lumeras, Maite and Andy Way** (2017). "On the complementarity between human translators and machine translation." *Hermes* 56, 21-42.

- **Artstein, Ron and Massimo Poesio** (2008). "Inter-coder agreement for computational linguistics." *Computational Linguistics* 34(4), 555-596.

- **Bojar, Ondřej, Chatterjee, Rajen, Federmann, Christian, Haddow, Barry, Huck, Matthias, Hokamp, Chris, Koehn, Philipp, Logacheva, Varvara, Monz, Christof, Negri, Matteo, Post, Matt, Scarton, Carolina, Specia, Lucia and Marco Turchi** (2015). "Findings of the 2015 Workshop on Statistical Machine Translation." *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 1-46.

- **Bojar, Ondřej, Chatterjee, Rajen, Federmann, Christian, Graham, Yvette, Haddow, Barry, Huck, Matthias, Jimeno Yepes, Antonio, Koehn, Philipp, Logacheva, Varvara, Monz, Christof, Negri, Matteo, Névéol, Aurélie, Neves, Mariana, Popel, Martin, Post, Matt, Rubino, Raphael, Scarton, Carolina, Specia, Lucia, Turchi, Marco, Verspoor, Karin and Marco Zampieri** (2016). "Findings of the 2016 Conference on Machine Translation (WMT16)." *Proceedings of the First Conference on Machine Translation. Volume 2: Shared Task Papers, Berlin 2016*, 131–198. http://www.aclweb.org/anthology/W16-2200 (consulted 20.11.2018).

- **Campbell, Stuart** (2000). "Critical structures in the evaluation of translations from Arabic into English as a second language." *The Translator* 6(2), 211-229.

- **Carl, Michael** (2012). "Translog-II: A Program for Recording User Activity Data for Empirical Translation Process Research." Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, Stelios Piperidis (eds) (2012). *Proceedings of the 8th International Conference on Language Resources and Evaluation, Istambul 2012*, 4108-4112. http://www.lrec-conf.org/proceedings/lrec2012/index.html (consulted 20.11.2018).

- **—** (forthcoming). "Literal Translation, Default Translation and the Similarity of Language Systems." To appear in *Target*.

- **Carl, Michael, Bangalore**, **Srinivas and Moritz Schaeffer** (2016). "The CRITT Translation Process Research Database." Michael Carl, Moritz Schaeffer and Srinivas Bangalore (eds) (2016). *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB*. Cham/Heidelberg/New York/Dordrecht/London: Springer, 13-56.

- **Carl, Michael and Moritz Schaeffer** (2017). "Why Translation Is Difficult: A Corpus-based Study of Non-literality in Post-editing and From-scratch Translation." *Hermes* 56, 43-57.

- **Carletta, Jean** (1996). "Assessing agreement on classification tasks: the Kappa statistic". *Computational Linguistics* 22(1), 249-254.

- **Costa, Ângela, Ling, Wang, Luís, Tiago, Correia, Rui and Luísa Coheur** (2015). "A linguistically motivated taxonomy for Machine Translation error analysis." *Machine Translation*, 29(2), 127-161.

- **Daems, Joke, Vandepitte, Sonia, Hartsuiker, Robert J. and Lieve Macken** (2017). "Identifying the Machine Translation Error Types with the Greatest Impact on Post-Editing Effort." *Frontiers in Psychology* 8, 1-15.

- **Denkowski, Michael and Alon Lavie** (2010). "Choosing the Right Evaluation for Machine Translation: An Examination of Annotator and Automatic Metric Performance on Human Judgment Tasks." *Proceedings of the Ninth Biennial Conference of the Association for Machine Translation in the Americas, Denver 2010*. https://www.cs.cmu.edu/~mdenkows/pdf/mteval-amta-2010.pdf (consulted 10.11.2018).

- **—** (2012). "Challenges in Predicting Machine Translation Utility for Human Post-Editors." *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA 2012), San Diego 2012*. http://www.cs.cmu.edu/~mdenkows/pdf/mt-post-amta2012.pdf (consulted 10.11.2018).

- **Di Eugenio, Barbara and Michael Glass** (2004). "The Kappa statistic: a second look". *Computational Linguistics* 30(1), 95-101.

- **Doherty, Stephen and Sharon O'Brien** (2009). "Can MT output be evaluated through eye tracking?" *Proceedings of the MT Summit XII, Ottwa 2009*, 214–221. http://www.mt-archive.info/MTS-2009-Doherty.pdf (consulted 10.11.2018).

- **Fomicheva, Marina, Bel, Nuria, Specia, Lucia and Iria Da Cunha** (2016). "CobaltF: A Fluent Metric for MT Evaluation." *Proceedings of the First Conference on Machine Translation. Volume 2: Shared Task Papers, Berlin 2016*, 483–490. http://www.aclweb.org/anthology/W16-2339 (consulted 20.11.2018).

- **Germann, Ulrich** (2008). "Yawat: Yet Another Word Alignment Tool." *ACL-08. Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. HLT Demo Session (Companion Volume), Columbus, Ohio 2008*. The Association for Computational Linguistics, 20-23. http://www.aclweb.org/anthology/P/P08/P08-40.pdf (consulted 20.11.2018).

- **Koby, Geoffrey S. and Gertrud G. Champe** (2013). "Welcome to the Real World: Professional-Level Translator Certification." *Translation & Interpreting* 5(1), 156-173.

- **Koponen, Maarit** (2012). "Comparing human perceptions of post-editing effort with post-editing operations." Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia (eds) (2012). *Proceedings of the 7th Workshop on Statistical Machine Translation, Montréal, June 7-8.* The Association for Computational Linguistics, 181–190. http://www.aclweb.org/anthology/W12-3100 (consulted 01.12.18).

- **Koponen, Maarit, Aziz, Wilker, Ramos, Luciana and Lucia Specia** (2012). "Post-editing time as a measure of cognitive effort." Sharon O'Brien, Michel Simard and Lucia Specia (eds) (2012). *AMTA 2012 Workshop on Post-editing Technology and Practice (WPTP), San Diego, United States, 28 October*. http://157.56.13.76/AMTA2012Files/html/13/13_paper.pdf (consulted 15.10.2018).

- **Krings, Hans P.** (2001). *Repairing Texts: Empirical Investigations of Machine Translation Post-editing Processes* (Geoffrey Koby, ed.). Kent: Kent State University Press.

- **Lacruz, Isabel and Gregory M. Shreve** (2014). "Pauses and Cognitive Effort in Post-Editing." Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard and Lucia Specia (eds) (2014). *Post-editing of Machine Translation: Processes and Applications*. Cambridge: Cambridge Scholars Publishing, 246-272.

- **Landis, Richard J. and Gary G. Koch** (1977). "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33(1), 159-174.

- **Lavie, Alon** (2011). "Evaluating the Output of Machine Translation Systems." *13th Machine Translation Summit* (Xiamen, 19 December 2011). http://www.cs.cmu.edu/~alavie/Presentations/MT-Evaluation-MT-Summit-Tutorial-19Sep11.pdf (consulted 10.11.2018).

- **Lavie, Alon and Abhaya Agarwal** (2007). "Meteor: An automatic metric for MT evaluation with high levels of correlation with human judgments." *StatMT '07 Proceedings of the Second Workshop on Statistical Machine Translation, Prague 2007*. The Association for Computational Linguistics, 228-231. http://www.statmt.org/wmt07/WMT-2007.pdf (consulted 10.11.2018).

- **Lommel, Arle, Uszkoreit, Hans and Aljoscha Burchardt** (2014a). "Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics." *Tradumàtica* 12, 455-463.

- **Lommel, Arle, Popović, Maja and Aljoscha Burchardt** (2014b). "Assessing Inter-Annotator Agreement for Translation Error Annotation." Keith J. Miller, Lucia Specia Kim Harris and Stacey Bailey (eds) (2014). *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, *Reykjavik 2014*, 31-37. http://mte2014.github.io/MTE2014-Workshop-Proceedings.pdf (consulted 10.11.2018).

- **Melamed, Dan I.** (2001). "Annotation style guide for the Blinker Project". Dan I. Melamed (ed.) (2001). *Empirical methods for exploiting parallel texts*. Cambridge, Massachusetts: MIT Press, 169-182.

- **MeLLANGE** (2006). Mellange WP4 Translation Error Typology. http://mellange.eila.jussieu.fr/Annotation_Schemes/current_translation_error_tree_29.jpeg (consulted 10.11.2017).

- **Merkel, Magnus** (1999). *Annotation Style Guide for the PLUG Link Annotator. Technical Report*. Department of Computer and Information Science, Linköping University.

- **O'Brien, Sharon** (2006). "Pauses as indicators of cognitive effort in post-editing machine translation output." *Across Languages and Cultures* 7(1), 1-21.

- **Papineni, Kishore, Roukos, Salim, Ward, Todd and Wei-Jing Zhu** (2002). "BLEU: A method for automatic evaluation of machine translation." *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (*ACL), Philadelphia, July 2002*, 311–318. https://www.aclweb.org/anthology/P02-1040.pdf (consulted 15.10.2018).

- **R Core Team** (2014). "R: A language and environment for statistical computing." *R Foundation for Statistical Computing*. Vienna. http://www.R-project.org/ (consulted 10.11.2018).

- **Scarton, Carolina and Lucia Specia** (2014). "Document-level translation quality estimation: exploring discourse and pseudo-references." Marko Tadić, Philipp Koehn,

Johann Roturier and Andy Way (eds) (2014). *Proceedings of the 17th Annual Conference of the European Association for Machine Translation. EAMT2014, Dubrovnik 2014*, 101-108. http://darhiv.ffzg.unizg.hr/id/eprint/5338/1/EAMT2014_proceedings.pdf (consulted 10.11.2018).

- **Scarton, Carolina, Beck, Daniel, Shah, Kashif, Smith, Karin Sim and Lucia Specia** (2016). "Word embeddings and discourse information for Machine Translation Quality Estimation." *Proceedings of the First Conference on Machine Translation. Volume 2: Shared Task Papers, Berlin 2016*, 831–837. http://www.aclweb.org/anthology/W16-2391 (consulted 10.11.2018).

- **Shah, Kashif, Logacheva, Varvara, Paetzold, Gustavo, Blain, Frédéric, Beck, Daniel, Bougares, Fethi and Lucia Specia** (2015). "SHEF-NN: Translation quality estimation with neural networks." *Proceedings of the Tenth Workshop on Statistical Machine Translation, Lisbon 2015*, 342–347. http://www.aclweb.org/anthology/W15-30 (consulted 10.11.2018).

- **Snover, Matthew, Dorr, Bonnie, Schwartz, Richard, Micciulla, Linnea and John Makhoul** (2006). "A Study of Translation Edit Rate with Targeted Human Annotation." *7th biennial conference of the Association for Machine Translation in the Americas. AMTA 2006, Cambridge, Massachusetts 2006*, 223-231. https://www.cs.umd.edu/~snover/pub/amta06/ter_amta.pdf (consulted 10.11.2018).

- **Vilar, David, Xu, Jia, Fernando D'Haro, Luis and Hermann Ney** (2006). "Error Analysis of Machine Translation Output." Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk and Daniel Tapias (eds) (2006). *Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genoa, Italy 2006*, 697–702. http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.pdf (consulted 10.11.2018).

Biographies

Michael Carl is a Professor at Kent State University (USA) and Director of the Center for Research and Innovation in Translation and Translation Technology (CRITT). He has studied Computational Linguistics and Communication Sciences in Berlin, Paris and Hong Kong and obtained his PhD degree in Computer Sciences from the Saarland University/Germany. His current research interest is related to the investigation of human translation processes and interactive machine translation.

E-mail: m.gummiball@gmail.com

M. Cristina Toledo Báez is an Assistant lecturer (with tenure) at the Department of Translation and Interpreting of the University of Córdoba (Spain). She holds a BA and a PhD in Translation and Interpreting from the University of Málaga. She has published extensively on translation technologies, translation assessment and specialised translation in peer-reviewed journals such as Spanish Journal of Applied Linguistics, Sendebar, Hikma and New Voices in Translation Studies, among others.

E-mail: cristina.toledo@uco.es

Notes

[1] English-simplified Chinese NMT output was obtained in May 2017 with Google NMT and the English-Spanish Google's Phrase-based Machine Translation (PBMT) was obtained in April, 2012 in the context of an earlier experiment within CRITT TPR-DB.

[2] https://sites.google.com/site/centretranslationinnovation/tpr-db

[3] http://www.casmacat.eu/

[4] http://www.qt21.eu/mqm-definition/definition-2014-06-06.html

[5] Researching word alignment is beyond the scope of our article. For further research into word alignment, see Melamed (2001) and Merkel (1999).

[6] Tokenization of the Chinese text was conducted with the Stanford segmenter (https://nlp.stanford.edu/software/segmenter.shtml). The example indicates token boundaries by blank spaces.

[7] The English-Spanish MT alignments data (MPM17) and the English-simplified Chinese MT alignments data (STCM17) are publicly available and can be downloaded from the CRITT TPR-DB via https://sites.google.com/site/centretranslationinnovation/tpr-db

[8] Professional translators agree more in what an error is and are clearer about the type of error.

[9] Researching how to calculate inter-annotator agreement on alignment is beyond the scope of our article. For further research, see Artstein and Poesio (2008).

[10] https://cran.r-project.org/web/packages/irr/irr.pdf

[11] For further information on Kappa scores, see Carletta (1996) and Di Eugenio *et al*. (2004).

[12] For the kappa score it does not matter whether the label is numeric or non-numeric.

[13] The model in R for this test was: lm({Dur | TrtS | TrtT} ~ any + STokLen, data = dataframe).

[14] Significance codes: p<0.001: '***', p<0.01: '**', p<0.05: '*', p<0.1: '.'

[15] The *HTra* score extrapolates probabilities of the translations, and is thus, to a certain extent, independent from the absolute amount of alternative translations. Thus, *HTra* values based on a set of 16 and 32 translations can be compared.

[16] Due to the long tail in the distribution of *Dur* values, a log-transformation results in more similar normal distribution. However, all ST words with *Dur=0* had to be taken out. These are words with no (aligned) translation, and translations of ST words in the post-edited texts which have not been modified. This reduced the data set by approximately 50%, so that the data set shrunk from 107867 to 50805 observations.