Human Evaluation of NMT & Annual Progress Report: A Case Study on Spanish to Korean

Ahrii Kim
Carme Colominas

Ahrii Kim
Grupo de Lingüística Computacional (GLiCom)
Departamento de Traducción y Ciencias del Lenguaje
Universitat Pompeu Fabra
ahriikim@gmail.com
ORCID:
0000-0003-2989-3220

Carme Colominas
Grupo de Lingüística Computacional (GLiCom)
Departamento de Traducción y Ciencias del Lenguaje
Universitat Pompeu Fabra
carme.colominas@upf.cat
ORCID:
0000-0002-0058-294X

Abstract

This paper proposes the first evaluation of NMT in the Spanish-Korean language pair. Four types of human evaluation —Direct Assessment, Ranking Comparison and MT Post-Editing(MTPE) time/effort— and one semi-automatic methods are applied. The NMT engine is represented by Google Translate in newswire domain. After assessed by six professional translators, the engine demonstrates 78% of performance and 37% productivity gain in MTPE. Additionally, 40.249% of the outputs of the engine are modified with an interval of 15 months, showing 11% of progress rate.

Keywords: NMT, MT evaluation, MT Post-Editing, Spanish-Korean translation

Resum

Aquest article proposa la primera avaluació de traducció automàtica neuronal en la combinació lingüística espanyol-coreà. Per fer-ho s'han aplicat quatre mètodes d'avaluació humana: l'avaluació directa, la comparació a través de la classificació dels segments i l'anàlisi del temps i de l'esforç de postedició del text traduït automàticament (en anglès, MTPE), i un mètode d'avaluació semiautomàtica.El motor detraducció automàtica neuronal utilitzat ha estat Google Translate, en concret en el seu domini de notícies. Després de ser avaluat per sis traductors professionals es constata que el motor augmenta el rendiment en un 78% i la productivitat en un 37%. A més, el 40,249% dels resultats del motor es modifiquen amb un interval de 15 mesos, de manera que mostra un índex de millora del 11%.

Paraules clau: Traducción automàtica neuronal, TAN, avaluació de TA, TAPE, postedició de traducció automàtica, traducció espanyol-coreà

Resumen

Este artículo propone la primera evaluación de traducción automática neuronal en la combinación lingüística español-coreano. Se han utilizado cuatro métodos de evaluación humana: la evaluación directa, la comparación mediante ranking y el análisis de tiempo y de esfuerzo de la posedición del texto traducido automáticamente (en inglés, MTPE), y un método de

Rebuda: 26 d'abril de 2020 / Acceptació: 9 de desembre de 2020 / Publicació: 30 de desembre de 2020

Cita recomanada: Ahrii Kim; Carme Colominas (2020). «Human Evaluation of NMT & Annual Progress Report: A Case Study on Spanish to Korean ». Revista Tradumàtica. Tecnologies de la Traducció, 18, 33-48. <https://doi.org/10.5565/rev/tradumatica.241>

evaluación semiautomática. El motor de traducción automática neuronal utilizado ha sido Google Translate, en concreto el dominio de noticias. Después de ser evaluado por seis traductores profesionales se constata que el motor aumenta el rendimiento en un 78% y la productividad en un 37%. Además, el 40,249% de los resultados del motor se modifican con un intervalo de 15 meses, mostrando así un índice de mejora del 11%.

Palabras clave: Traducción automática neuronal, TAN, evaluación de TA, TAPE, posedición de traducción automática, traducción español-coreano

1. Introduction

1.1 Introduction

The birth story of Machine Translation (MT) threw back to March 4, 1947 when Warren Weaver defined the concept of translation with encoding and decoding (Weaver, 1949: p.16). Starting from Rule-based MT (RBMT), this field witnessed two major turning points. The first moment was when Brown et al. (1988) presented Statistical MT (SMT). Instead of creating linguistic rules as in the previous approaches, the focal point of SMT was exploiting annotated data and matching equivalences. Subsequently, the second new wave came from a technological aspect. In around 2014, Neural MT (NMT) was showcased (Bahdanau et al., 2014). In NMT, the original concept of utilizing data in SMT remained identical, but the core technology was originated from the field of Artificial Intelligence (AI). With its growing viability, in just two years after its first debut, it became commercially available starting from Google Translate (Wu et al., 2016), and was widespread at an alarming rate.

The baseline technology of NMT is denominated as Artificial Neural Networks (ANNs), one of the Machine Learning algorithms that is advocated by connectionists who approach AI by imitating the interactions of human brain (Domingos, 2015). Simply put, axons and dendrites transmit/receive chemical and electrical signals by adding or subtracting them via so-called Action Potentials (British Neuroscience Association, 2003). A neuron needs to reach a certain limit to be fired in order to send signals that will strengthen the connections. Such a process is interpreted into the binary system of a 0 and 1 of Computer Science as such: a function $f(x)$ decides a bond of nodes by weakening $(y = 0)$ or strengthening $(y = 1)$ the value, in such a way that it updates information (Russell and Norvig, 1995). It is a gist of the threshold theory. Frank Rosenblatt proposes a single ANN by introducing the new concept of 'weight' to the given theory and names it as "perceptron".

Since then, various types of ANNs have been developed and tested in a number of AI tasks including MT. It started from hybrid architecture of SMT in the realm of n-gram language model (Bengio et al., 2003; Schwenk, 2007). Furthermore, Devlin et al. (2014) applied ANNs in a decoding step as a fully-integrated part that could be applicable to any decoders. Upon their success, in 2014 many scholars presented a purely-ANNs-based MT model (Sutskever et al.,2014; Bahdanau et al., 2014; Cho et al., 2014), opening a new era in MT. As such, NMT was distinctive in origin from the traditional MT paradigms.

1.2. Objective

Not only did the robustness of NMT have a great impact significantly on the MT field, but also the influence was evident in the humanities field. The study of MT evaluation or feasibility of MT post-editing (MTPE) increased markedly, in general, and in Korea, in particular. According to Korean Citation Index (KCI, 2020), the number of articles published under the keyword 'MT' in the humanities skyrocketed in 2017, as shown in Figure 1. Similarly, within the result, papers with the keywords 'MT Evaluation' (including Translation Quality) or 'MTPE' (either 'post-editing' or 'postediting') became noticeable since 2017. Such a trend served as an indication of the possible viability of NMT in the Korean language, which was linguistically distant to English and European languages and computationally low-resourced.

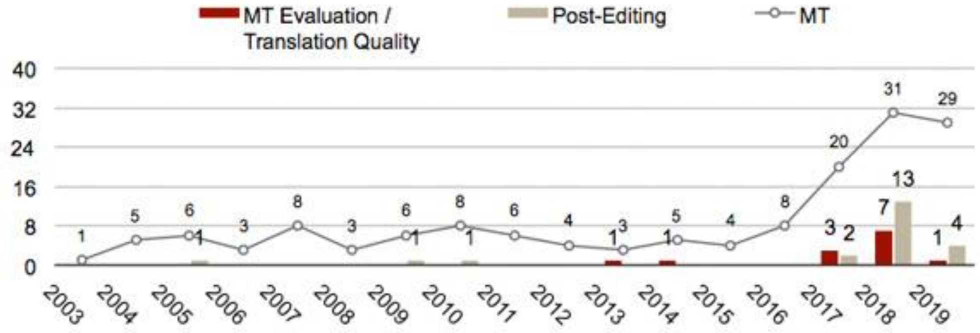


Figure 1. Number of registered articles in KCI under three keywords: MT, MT Evaluation and PE.

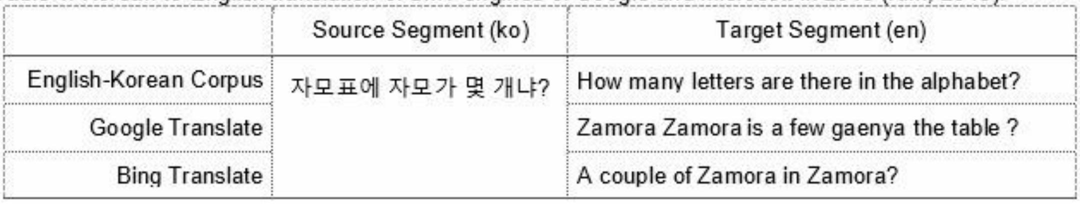
The main objective of this paper is to shed light on the possibility of NMT performance in Korean by reporting results of a fine-grained evaluation of NMT in a Spanish-to-Korean translation. The methodology of the intended evaluation is focused on human metrics of Fluency & Accuracy scoring, Ranking Comparison and MTPE time/effort. The HTER score is also proposed as a semi-automatic metric that can mediate the methodological imbalance. In addition to the report, a progress rate of the intended model is estimated by calculating how much the outputs have changed in the period of a year. From our understanding, it is believed that the current study is the first attempt to explore not only the Korean language but also the Spanish-Korean language combination, applying multifaceted and standardized MT evaluation methods. In this respect, we differ from the studies presented in Chapter 1.3.

This paper constitutes a part of the doctoral thesis of Kim (2019) that includes an NMT performance evaluation, as well as an error analysis. In this present work, we extend the evaluation study with an annual progress report.

1.3. Relevant Studies

More intriguing than the upsurge of interests in NMT in Korea (in Figure 1) was the little attention paid to SMT. Kim (2015: p.34) in her MA thesis speculated that the performance of SMT in Korean was untestable. She illustrated some English-to/from-Korean translation results done by Google Translate and Bing Translate. Although it was a small-scaled preliminary evaluation test, Table 1 clearly proved that the two online SMT engines could largely fail in such a basic sentence.

Table 1. Korean-to-English translation of SMT engines of Google and Microsoft in 2015 (Kim, 2015).



	Source Segment (ko)	Target Segment (en)
English-Korean Corpus	자모표에 자모가 몇 개냐?	How many letters are there in the alphabet?
Google Translate		Zamora Zamora is a few gaenya the table ?
Bing Translate		A couple of Zamora in Zamora?

Beginning from 2017, the advent of NMT arose a burst of enthusiasm for MT evaluation in the Korean academia, especially in relation to Chinese. Chang (2017) manually compared the performance of seven online NMT and SMT engines with 16 sentences selected from various domains. Ki (2018) compared two online NMT engines (Google Translate and Naver's Papago) with 580 sentences. In a 4-point-scale evaluation designed by the author, Papago showed slightly better performance. Similarly, Kang and Lee (2018) assessed three online NMT engines with a 10-point scale of Fidelity and Accuracy. 270 sentences were selected from verbal and literal texts. In terms of Korean-English, Choi and Lee (2017) evaluated an NMT engine with BLEU and a manual scoring system in 100 sentences in patent domain. They reported 22.90 of BLEU score, 47.5% of Fidelity and 45.5% of Readability, concluding that the engine was unproductive in this environment. Kim and Lee (2017) focused on 179 embedded sentences from a movie script to compare SMT and NMT. Their qualitative analysis found that NMT reduced syntactic errors but increased out-of-context errors.

2. Experiment Setup

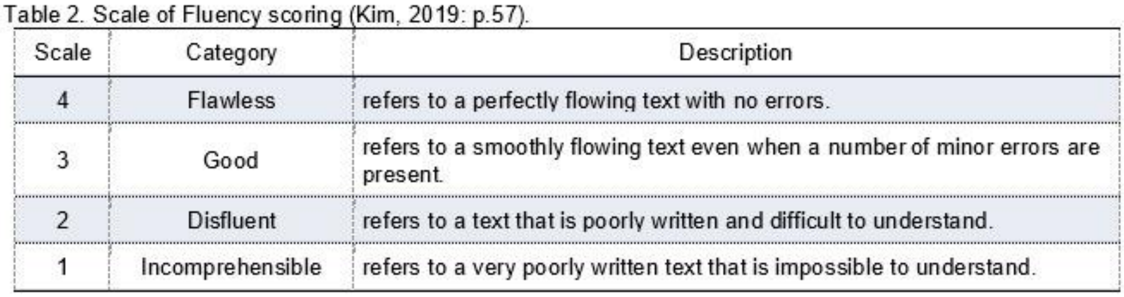
The NMT engine was represented by Google Translate of version 2018. The performance was assessed in newswire domain for the Spanish-to-Korean language direction by five different MT evaluation methods. The empirical experiment was conducted on TAUS Dynamic Quality Framework (DQF) with six professional translators and lasted for two weeks. The details were summarized subsequently.

2.1. Evaluation Method

Fluency Scoring

This test provides a direct judgment on each sentence. An annotator is asked to “capture to what extent the translation is well-formed grammatically, contains correct spellings, adheres to common use of terms, title and names, is intuitively acceptable and can be sensibly interpreted by a native speaker" (Görög, 2014). He/she is instructed to give a rating on a 4-point scale of Table 2 sentence by sentence.

Table 2. Scale of Fluency scoring (Kim, 2019: p.57).

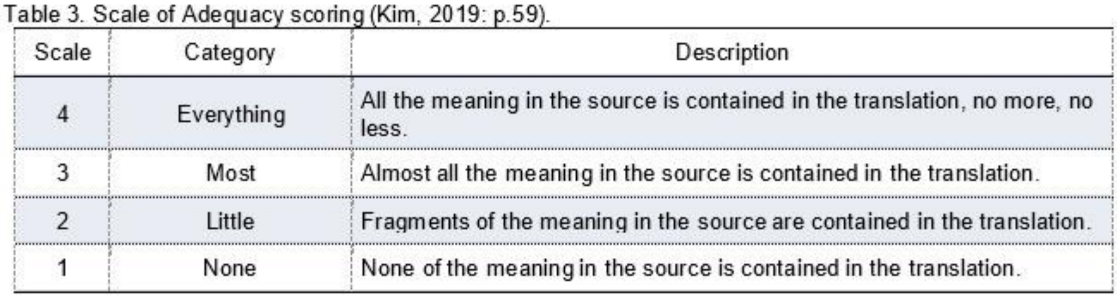


Scale	Category	Description
4	Flawless	refers to a perfectly flowing text with no errors.
3	Good	refers to a smoothly flowing text even when a number of minor errors are present.
2	Disfluent	refers to a text that is poorly written and difficult to understand.
1	Incomprehensible	refers to a very poorly written text that is impossible to understand.

Adequacy Scoring

This test is based on the identical architecture to the Fluency Scoring, but it concerns different aspects of the sentence. An annotator is asked to "capture to what extent the meaning in the source text is expressed in the translation" (Görög, 2014). As such, the current method takes both the source and target texts into consideration. The rating scale is identically 4-point as given in Table 3.

Table 3. Scale of Adequacy scoring (Kim, 2019: p.59).



Scale	Category	Description
4	Everything	All the meaning in the source is contained in the translation, no more, no less.
3	Most	Almost all the meaning in the source is contained in the translation.
2	Little	Fragments of the meaning in the source are contained in the translation.
1	None	None of the meaning in the source is contained in the translation.

Ranking Comparison

This test allows an indirect judgment of the engine by contrasting it to two other candidates —Kakao i (an online NMT engine) and a human translator. Provided anonymously with three translations, an annotator ranks them from the best to the worst, with a possibility of a tie. The rankings are then computed with 3, 2 and 1 points each for the final score. The sum per candidate is normalized by the number of segments.

MTPE Time/Effort

This test makes use of MTPE in MT evaluation by measuring how much time/effort is reduced by performing MTPE of the given engine instead of translating from scratch (TS). An annotator is engaged in full MTPE that aims at the level of a "similar-or-equal-to-human-translation quality" (TAUS, 2010) for half of the dataset and in TS for the rest half. The time and throughputs (words per hour, WPH) are calculated to compute productivity ratio. In MTPE effort, the correlation of MTPE time/throughput and sentence length is observed for temporal MTPE efforts (Koponen, 2012). Edit distance is used to measure technical MTPE efforts (Tatsumi, 2009).

HTER

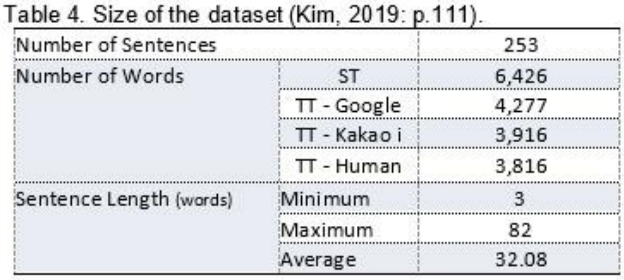
TER (Translation Error Rate) detected the similarity of the system and reference translation by calculating the minimum number of deletions, insertions, substitutions and shifts (reordering) (Snover et al., 2006). Going one step further, Human-mediated TER (HTER) improved TER's correlation to human judgment by filling the linguistic gap between the two texts (Snover et al., 2009). HTER substituted the reference translation to multiple MTPEs that are intentionally created for this purpose, commonly referred to as Targeted Reference (Snover et al., 2009). The minimum scores are selected and normalized by the number of words in the targeted reference.

2.2. Dataset

One of the biggest challenges of the Spanish-Korean pair is a lack of parallel corpora. They are so limited that a big part of available corpora in popular platforms like Wikipedia or OPUS might have been already employed in the training stage of such publicly popular MT systems. To alleviate such concern, we have collected hands-on data and hired a human professional translator to create its reference translation. A total of 253 Spanish sentences are extracted from three major journals —ABC, El País and KBS World Radio—- in the section of Politics. The main topic of all 11 articles is election-related. An example of the articles is given below for readers's information. The size of dataset is detailed in Table 4.

> El último Eurobarómetro, publicado el miércoles 17 de octubre, muestra un aumento del europeísmo incluso en Reino Unido, donde los partidarios de seguir en la UE superaban a los del Brexit. Sin embargo, cuando se acaban de cumplir 25 años de la entrada en vigor del Tratado de Maastricht que diseñó la actual Unión y su moneda única, el euro, la UE se enfrenta a uno de los mayores desafíos de su historia en las próximas Elecciones Europeas: por primera vez se espera que los tradicionales bloques centroizquierda y centroderecha europeístas caigan por debajo del 50% y algunas encuestadoras estiman que en torno a un tercio de los escaños serán ocupados por partidos nacional-populistas, que tratan de torpedear desde dentro los valores europeístas y que paradójicamente se han aprovechado de fondos de la Unión para impulsar sus finanzas. [...]

Table 4. Size of the dataset (Kim, 2019: p.111).



Number of Sentences		253
Number of Words	ST	6,426
	TT - Google	4,277
	TT - Kakao i	3,916
	TT - Human	3,816
Sentence Length (words)	Minimum	3
	Maximum	82
	Average	32.08

2.3. Profile

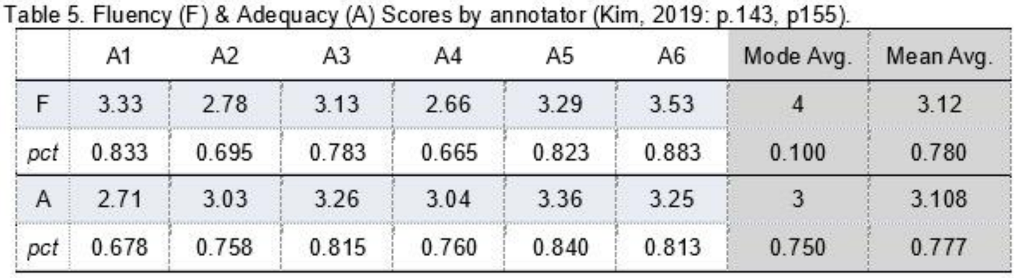
Six professional translators, who have been engaged in the Spanish-Korean language combination for a period of one to five years, are hired for the experiment. They do not have previous experience in MTPE. They are all native.

3. Results

3.1. Fluency & Adequacy

In Table 5, the Fluency & Adequacy scores of each annotator were presented. The Google NMT system obtained on (mean) average 3.12 Fluency and 3.108 Adequacy scores of 4, equal to 78% and 77.7%. With a margin of 0.3% point, the engine was judged to be more fluent than adequate.

Table 5. Fluency (F) & Adequacy (A) Scores by annotator (Kim, 2019: p.143, p155).



	A1	A2	A3	A4	A5	A6	Mode Avg.	Mean Avg.
F	3.33	2.78	3.13	2.66	3.29	3.53	4	3.12
pct	0.833	0.695	0.783	0.665	0.823	0.883	0.100	0.780
A	2.71	3.03	3.26	3.04	3.36	3.25	3	3.108
pct	0.678	0.758	0.815	0.760	0.840	0.813	0.750	0.777

Taking a closer look, Figure 2 and Figure 3 illustrated the distribution of Fluency and Adequacy scores by scale. In Fluency, the biggest pie was taken by Scale 4 with 48.1%. In Adequacy, one the other hand, Scale 3 was predominant with 46.64%. A positive outcome was that 48.1% of the dataset were grammatically flawless and 32.47% contained all elements of the source text.

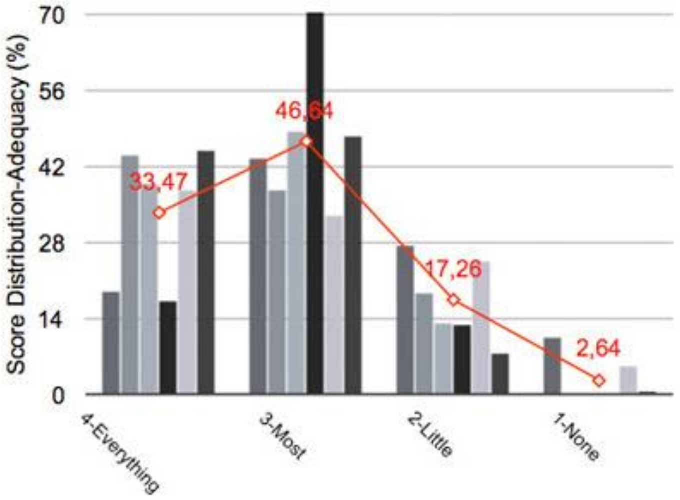


Figure 2. Score distribution of Fluency (Kim, 2019: p.144).

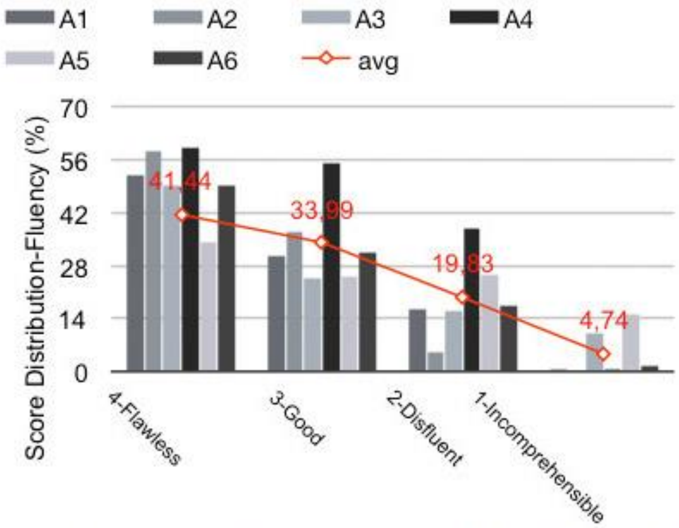
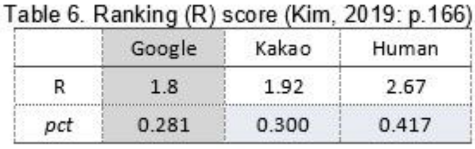


Figure 3. Score distribution of Adequacy (Kim, 2019: p.156).

3.2. Ranking Comparison

Compared to Kakao i and human reference translation, Google Translate earned 1.8 (of 3) ranking score and was considered as the worst candidate with 28.17% of preference, as in Table 6. In the meantime, it came to our attention that when a distinction of human versus machine was drawn, it turned out that the annotators preferred the two system translations (58.22%) to the human translation (41.78%). The reasons were unclear, but some possible scenarios were speculated in Kim (2019).

Table 6. Ranking (R) score (Kim, 2019: p.166).



	Google	Kakao	Human
R	1.8	1.92	2.67
pct	0.281	0.300	0.417

Subsequently, the result was organized by machine and ranking choice in Figure 4 and Figure 5. Focusing on the case of Google Translate, almost the half of the dataset was positioned in the second rank (49.61%). The first-ranked segments took up only 16.54%. When ranking choice was concerned, 14.79% were chosen as Rank 1 and 53.88% were Rank 3 in Google Translate. All in all, this test showed that there were plenty of rooms for improvement in this engine.

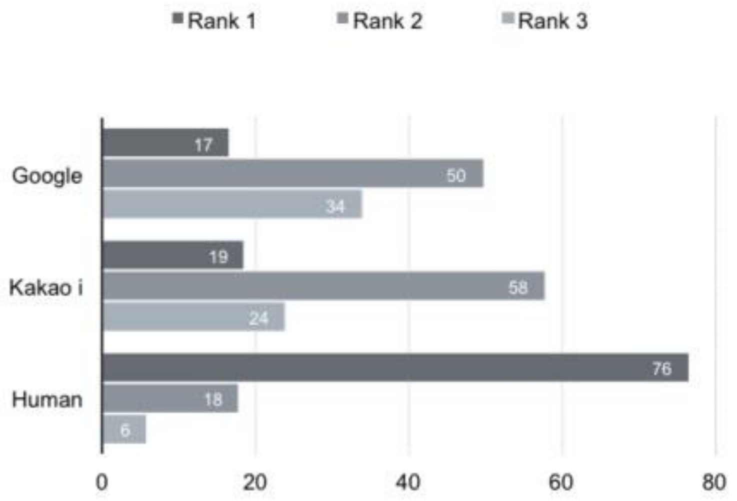


Figure 4. Score distribution of ranking by machine (Kim, 2019: p.170).

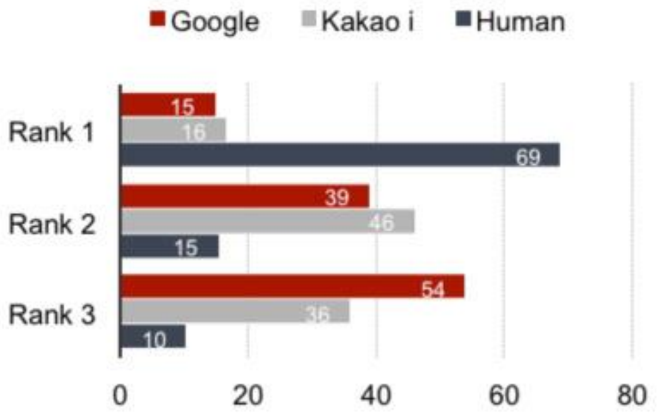
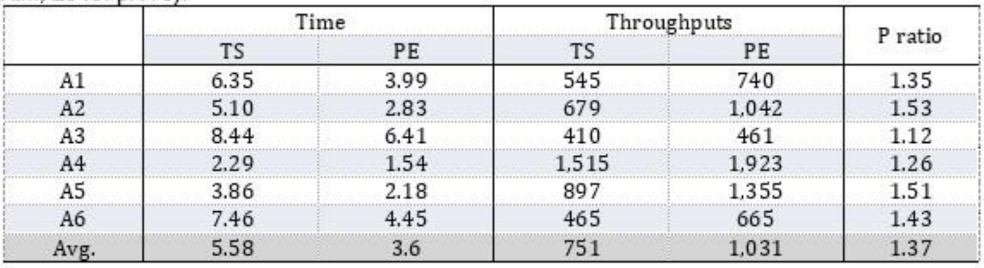


Figure 5. Score distribution of ranking by ranking choice (Kim, 2019: p.171).

3.3. MTPE Time & Effort

In comparison to TS, Table 7 showed that MTPE was 37% faster on average and in a range of 12% - 53%. We, however, could not interpret a real sense of this 37% productivity, as no standard was established currently in this regard. Groves and Schmidtke (2009) reported 6.1% - 28.7% gains while the case of Plitt and Masselot (2010) and Skadina and Pinnis (2017) reached up to 118%. The closest study to ours was Zhechev (2014) who obtained 81.93% gain in the English-Korean pair. From these previous studies, it was soon to declare that MTPE would be always more recommendable than TS in our language pair in this environment, but MTPE was more time-efficient than TS in our study.

Table 7. TS & PE time (unit: hour) and throughputs (unit: WPH) by annotator (A1 - A6) and their average (Kim, 2019: p.176).



	Time		Throughputs		P ratio
	TS	PE	TS	PE	
A1	6.35	3.99	545	740	1.35
A2	5.10	2.83	679	1,042	1.53
A3	8.44	6.41	410	461	1.12
A4	2.29	1.54	1,515	1,923	1.26
A5	3.86	2.18	897	1,355	1.51
A6	7.46	4.45	465	665	1.43
Avg.	5.58	3.6	751	1,031	1.37

In relation to effort reduction in MTPE, the temporal effort was far lower in short sentences ($l <= 13$) and become higher from sentences of $l=31$, as in Figure 6. The tendency, however, was not proportionate. When it was measured with MTPE throughputs (WPH), no clear-cut correlation was observed, as in Figure 7. Interestingly though, all sentences required certain degree of MTPE efforts, with minimum 380 WPH. Hence, it was estimated that MTPE was efficient in sentences of $l <= 13$ but inefficient in those of $l >= 31$. Such a finding also coincided with the comments of the annotators in Kim (2019: p.134-138).

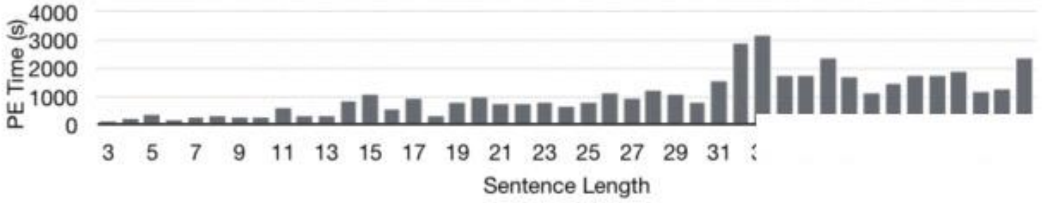


Figure 6. Temporal PE efforts by time (unit: second) per sentence length (Kim, 2019: p.178).

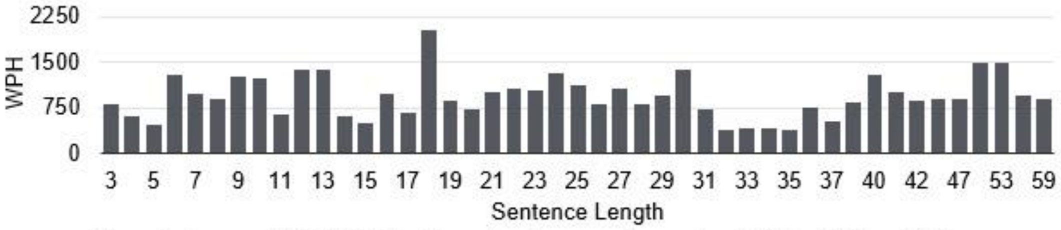
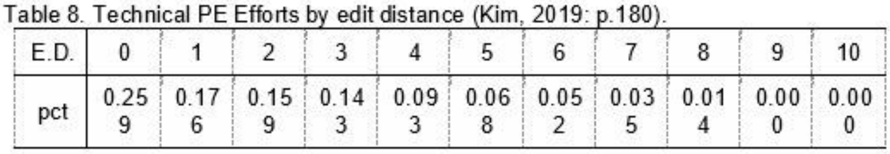


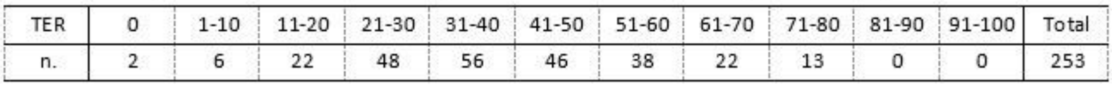
Figure 7. Temporal PE Efforts by throughputs per sentence length (Kim, 2019: p.179).

In terms of the technical effort measured by edit distance, Table 8 displayed that 25.9% of the dataset hardly required any MTPE. It was also noticeable that not a single sentence was entirely deleted to be translated from scratch ($d = 10$).

Table 8. Technical PE Efforts by edit distance (Kim, 2019: p.180).



E.D.	0	1	2	3	4	5	6	7	8	9	10
pct	0.25 9	0.17 6	0.15 9	0.14 3	0.09 3	0.06 8	0.05 2	0.03 5	0.01 4	0.00 0	0.00 0

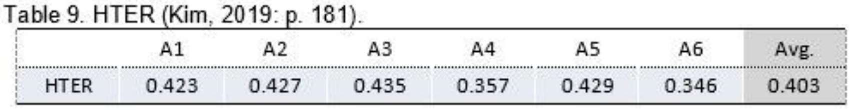


TER	0	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100	Total
n.	2	6	22	48	56	46	38	22	13	0	0	253

3.4. HTER

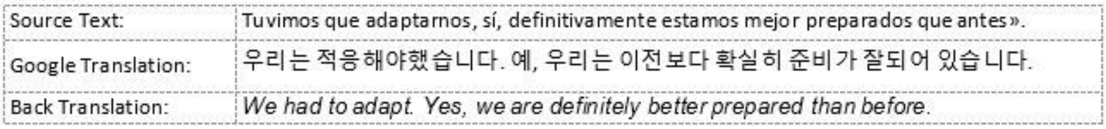
As in Table 9, the HTER score of Google Translate ranged from 35.73 to 43.52, with an average of 40.33. In other words, at least 35.73%% of the dataset were reformed to satisfy the translation quality. We, however, acknowledged a potential bias in this result due to the characteristic of Korean as an agglutinative language, whose word-spacing unit did not match its part-of-speech tagging (Song and Park, 2020). For example, in Table 10, the first word (우리는) was composed of "we (우리-)" and subject case marker (-는). A back translation to English was given for readers' information.

Table 9. HTER (Kim, 2019: p. 181).



	A1	A2	A3	A4	A5	A6	Avg.
HTER	0.423	0.427	0.435	0.357	0.429	0.346	0.403

Table 10.



Source Text:	Tuvimos que adaptarnos, sí, definitivamente estamos mejor preparados que antes».
Google Translation:	우리는 적응해야했습니다. 예, 우리는 이전보다 확실히 준비가 잘되어 있습니다.
Back Translation:	*We had to adapt. Yes, we are definitely better prepared than before.*

4. Annual Progress Report

The aforementioned experiment gave us insight into the status quo of the Google NMT engine in the Spanish-Korean pair, which could be summarized as follows:

- The Direct Assessment on the engine confirmed 78% of performance.
- The comparative study suggested that if a human parity was defined as the first-ranked system translation, the given engine obtained 16.54% of human parity.
- MTPE was 37% faster than TS. It was especially effective in short sentences.

Considering the past performance of SMT in Table 1, the fact that NMT achieved positive results alone was a remarkable phenomenon. At this point in time, we came to inquire into how fast and to where NMT would further grow. To this aim, the performance of NMT was chronologically compared in two different periods of time: November 2018 and February 2020. From our understanding, such temporal approach was a new approach in this area. From this mini-task, two questions were answered:

- How much did the result change?
- Were those changes positive or negative?

The two versions of system translation (named after Old and New) based on the equivalent dataset to that of the evaluation experiment in Chapter 3 —6,426 words in the newswire domain— were prepared and analyzed. The comparison was based on edit distance of TER on a sentence basis on TAUS DQF. Additionally, Ranking Comparison was manually carried out, in which the author played as a sole annotator.

4.1. Change Rate

Table 11 demonstrated edit distance of all 253 sentences displayed in percentage in eleven sections. On average, 40.249% of the dataset were modified after the given period in Google Translate. Observing TER of each sentence as in Figure 8, a certain level of modification was performed throughout the whole dataset, with an exception of two sentences (TER = 0.0). The largest modification was witnessed with one case, with 80% of changes (TER = 0.8) as shown in Table 12. The modifications were witnessed not only in lexicon but also in syntax.

TER	0	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100	Total
n.	2	6	22	48	56	46	38	22	13	0	0	253

Table 11. Distribution of TER of Old vs. New.

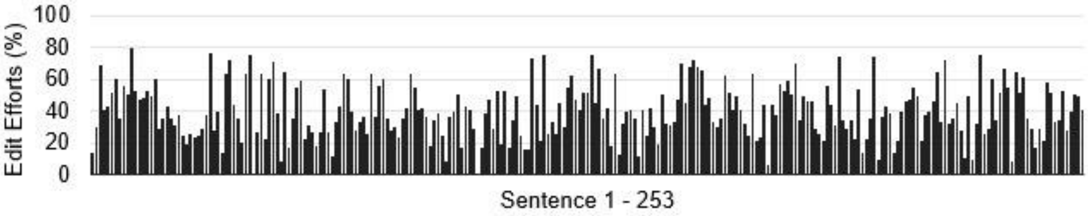
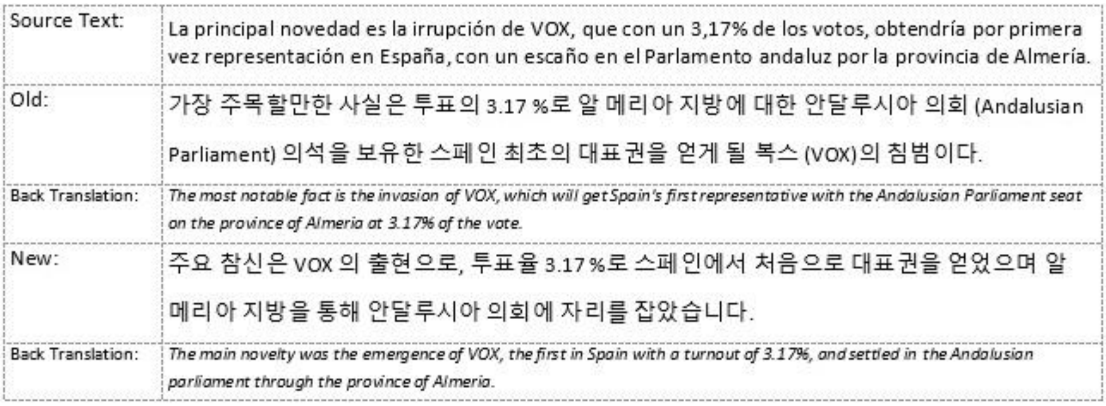


Figure 8. New vs. Old

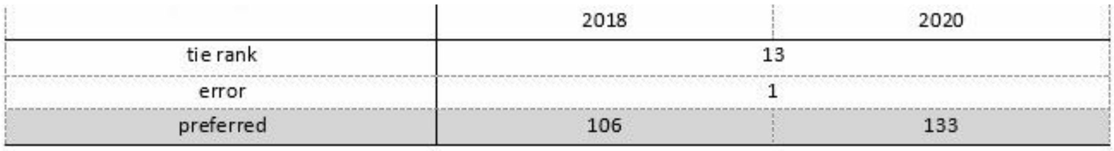


Source Text:	La principal novedad es la irrupción de VOX, que con un 3,17% de los votos, obtendría por primera vez representación en España, con un escaño en el Parlamento andaluz por la provincia de Almería.
Old:	가장 주목할만한 사실은 투표의 3.17 %로 알 메리아 지방에 대한 안달루시아 의회 (Andalusian Parliament) 의석을 보유한 스페인 최초의 대표권을 얻게 될 복스 (VOX)의 침범이다.
Back Translation:	*The most notable fact is the invasion of VOX, which will get Spain's first representative with the Andalusian Parliament seat on the province of Almeria at 3.17% of the vote.*
New:	주요 참신은 VOX 의 출현으로, 투표율 3.17 %로 스페인에서 처음으로 대표권을 얻었으며 알 메리아 지방을 통해 안달루시아 의회에 자리를 잡았습니다.
Back Translation:	*The main novelty was the emergence of VOX, the first in Spain with a turnout of 3.17%, and settled in the Andalusian parliament through the province of Almeria.*

Table 12.

4.2. Positivity

In Ranking Comparison, a better option between the New and Old was directly selected on a sentence basis. It turned out that excluding one erroneous sentence, the New was about 11% more preferred than the Old, with 55.65% versus 44.35% (Table 13). 13 sentences were of equal value. With these results at hand, our study confirmed that there was a strong possibility of progress of the given engine.



	2018	2020
tie rank	13	
error	1	
preferred	106	133

Table 13. Ranking Comparison of Old vs. New.

5. Conclusion

The assorted manual evaluations of NMT-based Google Translate in the newswire domain in the Spanish-Korean language pair was carried out with six professional translators. The engine achieved 78% of fluency and 77.7% of adequacy. The quality was still far behind the human translation with 16.54% of human parity. In regard to MTPE, it was 37% more productive than TS. The MTPE effort reduction rate was distinctive in shorter sentences, but it turned out that all segments required a certain level of MTPE efforts regardless of sentence length. Technically speaking, there was 25.9% of MTPE effort reduction. HTER revealed that at least 35.73% of the dataset were edited.

Taking a comprehensive stance, we could conclude that understanding the meaning of the text with this engine was guaranteed in this setup. Our study proved that NMT was breakthrough technology for this language combination. It also gave hope that MT was tearing down the language barrier. The question was: Is the performance good enough to substitute TS to MTPE? There was no doubt that MTPE was strongly recommended, but at this level we encouraged MTPE only for short sentences of up to 13 words.

As a mini project, we examined the progress rate of Google Translate for a period of 15 months. Compared to year 2018, the engine made 40.35% of modification to the equivalent dataset in year 2020, according to TER. From a quick comparison test, it was estimated that there was 11% of progress in the engine. Taking all into account, we expect a bright future of NMT ahead in the Spanish-Korean language combination.

6. Future Research

Given such circumstance where the linguistic barrier is considerably resolved between the two languages, we assume that automatic evaluation of NMT will be of utmost value. It is our first aim to organize automatic evaluation of NMT in this language pair, with larger dataset and hopefully more annotators. We are also interested in comparing the performance of NMT and SMT in the given environment.

Referencies

Bahdanau, D.; Cho, K.; Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate, in: CoRR, Accepted for oral presentation at the International Conference on Learning Representations (ICLR) 2015. <https://arxiv.org/abs/1409.0473>.

Bengio Y.; Ducharme, R.; Vincent, P. (2003). A neural probabilistic language model, in: Journal of Machine Learning Research, v. 3, pp. 1137-1155. <http://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf>.

British Neuroscience Association (2003). Science of the brain: an introduction for young students. <https://www.bna.org.uk/static/uploads/resources/BNA_English.pdf>. Last updated: 2003. Page consulted on date: 07.05.18.

Brown, P.; Cocke, J.; Della Pietra, S.; Della Pietra, V., Jelinek, F.; Mercer, R.; Roossin, P. (1988). A Statistical Approach to French/English Translation, in: Proceedings of the 2nd International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages (12-14 June 1988). Center for Machine Translation, Carnegie Mellon University, Pittsburgh, Pennsylvania, the United States of America. <https://dl.acm.org/doi/10.5555/3170668.3170681>. https://doi.org/10.1007/978-94-009-3117-6_27

Chang, A. (2017). Analysis of the Current Development of Machine Translation and Interpretation in Korea: Focusing on Korean-Chinese Language Pairs, in: The Journal of Translation Studies v. 18, n. 2, pp. 171-206. <http://doi.org/10.15749/jts.2017.18.2.007 >.

Cho, K.; Merriënboer, B.; Gülcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, in: Proceedings of EMNLP 2014, Doha, Qatar, October: Association for Computational Linguistics. <https://arxiv.org/abs/1406.1078>. https://doi.org/10.3115/v1/D14-1179

Choi, H.; Lee, J. (2017). A study on the evaluation of Korean-English patent machine translation - Focusing on KIPRIS K2E-PAT translation, in: Interpretation & Translation, v. 19, n. 1: pp. 139-178. <http://doi.org/10.20305/it201701139178 >. https://doi.org/10.20305/it201701139178

Devlin, J.; Zbib, R.; Huang, Z.; Lamar, T.; Schwartz, R.; Makhoul, J. (2014). Fast and Robust Neural Network Joint Models for Statistical Machine Translation, v. 1: pp. 1370-1380. <https://doi.org/10.3115/v1/P14-1129>.

Domingos, P. (2015). The Master Algorithm. Basic Books, 1st edition.

Görög, A. (2014). Quality evaluation today: the Dynamic Quality Framework, in: Proceedings of Translating and the Computer 36: ASLING: Proceedings. Geneva: Tradulex, pp. 155-164. <http://www.tradulex.com/varia/TC36-london2014.pdf>.

Groves, D.; Schmidtke, D. (2009). Identification and analysis of post-editing patterns for MT. <http://www.mt-archive.info/MTS-2009-Groves.pdf>.

Kang, B.; Lee, J. (2018). The Operating Principles of Neural Machine Translation and the Accuracy of Translation - Focusing on the Chinese-Korean Translation, in: The Journal of Chinese Language and Literature, v. 73, pp. 253-295. <https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART002389060>. https://doi.org/10.46612/kjcll.2018.09.73.253

Ki, Y. (2018). An Analysis of Errors by sentence pattern in translating Korean sentences into Chinese by Machine Translation - focus on Naver Papago machine translation and Google machine translation, in: Chinese Studies, v. 74, pp. 3-32. <http://doi.org/10.18077/chss.2018.74..001>. https://doi.org/10.18077/chss.2018.74..001

Kim, A. (2015). Reordering of SOV-SVO Pairs in Statistical Machine Translation: In Relation to the Korean Language. Masters thesis at Universitat Pompeu Fabra.

Kim, A. (2019). Neural Machine Translation Evaluation & Error Analysis in a Spanish-Korean Translation. Doctoral thesis at Universitat Pompeu Fabra. Retrieved from https://repositori.upf.edu/handle/10230/42853.

Kim, S.; Lee, H. (2017). A Study on Machine Translation Outputs - Korean to English Translation of Embedded Sentences, in: The Journal of Mirae English Language and Literature, v. 22, n. 4, pp. 123-147. <http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07273221>.

Koponen, M. (2012). Comparing human perceptions of post-editing effort with post-editing operations, in: Proceedings of the 7th Workshop on Statistical Machine Translation. Montreal, Canadá: Association for Computational Linguistics, pp. 181-190. <https://www.aclweb.org/anthology/W12-3123>.

Korean Citation Index <https://www.kci.go.kr/kciportal/main.kci?locale=en>. Page consulted on date: 15.03.20.

Plitt, M.; Masselot, M. (2010). A Productivity Test of Statistical Machine Translation, in: The Prague Bulletin of Mathematical Linguistics, v. 93, pp. 7-16. <http://doi.org/10.2478/v10108-010-0010-x>.

Russell, J.; Norvig, P.; Canny, F.; Malik, M.; Edwards, D. (1995). Artificial Intelligence: a Modern Approach. Vol 2, Englewood Cliffs: Prentice Hall.

Schwenk, H. (2007). Continuous space language models, in: Computer Speech & Language, v. 21, pp. 492-518. <http://doi.org/10.1016/j.csl.2006.09.003>. https://doi.org/10.1016/j.csl.2006.09.003

Skadina, I.; Pinnis, M. (2017). NMT or SMT: Case Study of a Narrow-domain English-Latvian Post-editing Project, in: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Vol 1: Long Papers), pp. 373 - 383. <https://www.aclweb.org/anthology/I17-1038>.

Snover, M.; Dorr, B.; Schwartz, R.; Micciulla, L.; Makhoul, J. (2006). A study of translation edit rate with targeted human annotation, in: Proceedings of the 7th Biennial

Conference of the Association for Machine Translation in the Americas (AMTA-2006). Cambridge, Massachusetts: Association for Machine Translation in the Americas.

Snover, M.; Madnan,i N.; Dorr, J.; Schwartz, R. (2009). Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric, in: Proceedings of the Fourth Workshop on Statistical Machine Translation, pp. 259-268. <https://www.aclweb.org/anthology/W09-0441>. https://doi.org/10.3115/1626431.1626480

Song, H.; Park, S. (2020). Korean Part-of-speech Tagging Based on Morpheme Generation, in: ACM Trans. Asian Low-Resour. Lang. Inf. Process. n. 19, v. 3, Article 41 (January 2020), 10 pages. <https://doi.org/10.1145/3373608>.

Sutskever, I.; Vinyals, O.; Le, Q. (2014). Sequence to Sequence Learning with Neural Networks, in: Proceedings of the Neural Information Processing Systems, Vol 2 (NIPS'14), MIT Press, Cambridge, MA, USA, pp. 3104 - 3112. <https://dl.acm.org/doi/10.5555/2969033.2969173>.

Tatsumi, M. (2009). Correlation Between Automatic Evaluation Metric Scores, Post-Editing Speed, and Some Other Factors. <http://www.mt-archive.info/MTS-2009-Tatsumi.pdf>.

TAUS (2010). Machine Translation Post-editing Guidelines. <https://www.taus.net/think-tank/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines>. Last updated: 2010. Page consulted on date: 04.08.17.

Weaver, W. (1949). "Translation", Reprinted in Locke, W.; Booth, A. (eds.) Machine Translation of Languages: Fourteen Essays, Cambridge, Massachusetts: Technology Press of the Massachusetts Institute of Technology: pp. 15-33.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; Klingner, J.; Shah, A.; Johnson, M.; Liu, X.; Kaiser, L.; Gouws, S.; Kato, Y.; Kudo, T.; Kazawa, H.; Stevens, K.; Kurian, G.; Patil, N.; Wang, W.; Young, C.; Smith, J.; Riesa, J.; Rudnick, A.; Vinyals, O.; Corrado, G.; Hughes, M.; Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation, in: CoRR. <https://arxiv.org/pdf/1609.08144.pdf>.

Zhechev, V. (2014). "Analysing the Post-Editing of Machine Translation at Autodesk," in O'Brien, S.; Balling, L.; Carl, M.; Simard, M.; Specia, L. (eds.) (2014), Post-editing of Machine Translation: Processes and Applications: Cambridge Scholars Publishing.