Neural Machine Translation Evaluation & Error Analysis in a Spanish-Korean Translation

Ahrii Kim

TESI DOCTORAL UPF / 2019

Directora de la tesis

Prof. Carme Colominas Ventura

DEPARTAMENTO DE TRADUCCIÓN Y CIENCIAS DEL LENGUAJE

upf. Universitat Pompeu Fabra Barcelona

[Blank page]

Abstract

From RBMT to SMT and NMT, the MT field witnessed, first, a conceptual turn —from rule-based to data-base— and now, a technological turn —from MT algorithm to ML algorithm. Now that NMT became a new state of the art, this thesis quested for evaluating its performance in a Spanish-to-Korean translation, which, for the best of our knowledge, was the first attempt in this regard. The results reported that the NMT-based Google Translate (GNMT) had about 78% of reliability. In an experiment with post-editing, the post-editing was 37% more productive in GNMT than translation from scratch. An important finding was obtained from quantitative and qualitative error analysis. It reported that only 6% of the errors detected in the dataset were a syntactic error in such a distant pair like this. The results of this thesis served as a proof of a promising future of NMT in distant pairs.

Keywords

Neural Machine Translation, MT Evaluation, Error Analysis, Spanish-Korean translation

Resum

Des de la Traducció Automàtica (TA) basada en regles a la TA estadística i la TA neuronal (TAN), el camp de la TA va presenciar, primer, un gir conceptual - des d'aproximacions basades en regles fins aproximacions basades en dades- i ara, un gir tecnològic –de l'algoritme de la TA al d'Aprenentatge Automàtic. Ara que la TAN s'ha convertit en un nou estat de l'art, busquem avaluar el seu grau de qualitat en la traducció de l'espanyol al coreà,. Aquest estudie constitueix, segons el nostre coneixement, el primer que intenta avaluar aquest parell de llengües. Els resultats informen que Google Translate, basada en la TAN té al voltant el 78% de fiabilitat. En un experiment amb postedició, la postedició és un 37% més productiva que la traducció des de zero. Apartir d'una anàlisi d'errors quantitativa i qualitativa hem pogut fer constatar que només el 6% dels errors detectats van ser de naturalesa sintàctica en un parell de llengües tan distant com aquest. Els resultats obtinguts en aquesta tesi van servir com a prova per a un futur prometedor de la TAN en parells distants.

Paraules Clau

Traducció Automàtica Neuronal, Avaluació de Traducció Automàtica, Anàlisi d'Errors, Traducció de l'Espanyol-Coreà

Resumen

Desde la Traducción Automática (TA) basada en reglas a la TA estadística y la TA neuronal (TAN), el campo de la TA presenció, primero, un giro conceptual —desde aproximaciones basadas en reglas hasta aproximaciones basadas en datos— y ahora, un giro tecnológico —del algoritmo de la TA al de Aprendizaje Automático. Ahora que la TAN se ha convertido en un nuevo estado del arte, buscamos evaluar su desempeño en la traducción del español al coreano, que constituye, según nuestro conocimiento, el primer intento al respecto. Los resultados informan que Google Translate basada en la TAN tenía alrededor del 78% de fiabilidad. En un experimento con posedición, la posedición es un 37% más productiva que la traducción desde cero. Obtuvimos un hallazgo importante a partir de un análisis de errores cuantitativo y cualitativo. Informamos que solo el 6% de los errores detectados fueron sintácticos en un par de lenguas tan distante como este. Nuestros resultados sirvieron como prueba para un futuro prometedor de la TAN en pares distantes.

Palabras Clave

Traducción Automática Neuronal, Evaluación de Traducción Automática, Análisis de Errores, Traducción del Español-Coreano

초록

기계번역 분야는 규칙기반 번역에서 통계기반 번역으로, 또 인공신경망 번역으로 변화를 거듭하고 있다. 규칙에 의존하던 번역 방식이 데이터에 의존하는 **발상의 전환**이 한 차례 일어났고, 그 후 머신러닝 알고리즘을 적용하는 **기술적 전환**이 일어났다. 본지는 인공신경망 기계번역이 새롭게 부상함에 따라 스페인어-한국어 쌍에 대한 인공신경망 번역기의 성능을 평가하고자 했다. 스페인어-한국어 언어쌍에 대한 기계번역 평가는 최초로 이루어지는 연구이다. 평가 결과 인공신경망 기반 구글 번역기는 78%의 신뢰도를 보였다. 포스트에디팅을 통한 연구에서는 포스트에디팅이 번역 대비 37%의 생산성을 보였다. 이어 본지는 정량적, 정성적 오류 분석을 통해 흥미로운 점을 발견하였다. 발견된 모든 오류 중 6%만이 통사적 오류로 드러났으며, 이는 스페인어-한국어와 같이 매우 이질적인 언어 쌍에서 기대하기 어려운 결과였다. 본지는 이질적인 언어쌍에 대한 인공신경망 번역기의 가능성을 입증하며, 인공신경만 기계번역의 무한한 가능성을 드러냈다.

키워드

인공신경망 기계번역, 기계번역 평가, 오류 분석, 스페인어-한국어 번역

Table of contents

List of figures

List of tables

List of abbreviation

AI	Artificial Intelligence
ANN	Artificial Neural Network
BT	Back translation
DeepL	Deep Learning
EBMT	Example-Based Machine Translation
GNMT	Neural-based Google Translate
GT	System translation of Google Translate
HT	Human translation
KT	System translation of Kakao i
LSTM	Long-Short-Term Memory
MT	Machine Translation
NMT	Neural Machine Translation
OOV	Out-of-vocabulary
POS	Part of speech
RBMT	Rule-Based Machine Translation
RNN	Recurrent Neural Network
RT	Reference translation
SMT	Statistical Machine Translation
ST	Source text
TR	Targeted reference translation
WPH	Words per hour

[blank page]

1. INTRODUCTION

> And the LORD said, Behold, the people is one, and they have all one language; and this they begin to do: and now nothing will be restrained from them, which they have imagined to do. Go to, let us go down, and there confound their language, that they may not understand one another's speech. So the LORD scattered them abroad from thence upon the face of all the earth: and they left off to build the city. Therefore is the name of it called Babel; because the LORD did there confound the language of all the earth: and from thence did the LORD scatter them abroad upon the face of all the earth. (Genesis 11:6–9)

Since the creation of human being or mythically the construction of the Tower of Babel, we are living in a world where the language barrier challenges our communication. Humans have fought against the challenge by developing a translating machine that is faster and more cost-effective than us, but that still lacks accuracy. The recent advent of Neural Machine Translation, however, is overstepping the limitation, producing not only efficient but also accurate translation.

This chapter guides the readers to the evolvement of NMT from a historical standpoint, beginning from the birth of the concept of a translating machine to current status of being a possible threat that might substitute human translators. In such a context where unfounded beliefs are spreading over, it is important to assess the performance of NMT and find a way to make the technology useful for human. However, the study of the Spanish-Korean language combination has never been considered in the MT field, and thus, the current thesis hypothesizes a possible outcome of the research by connecting the dots from the former studies in the relevant field.

1.1. Historical Standpoint Towards Neural Machine Translation

> Recognizing fully, even though necessarily vaguely, the semantic difficulties because of multiple meanings, etc., I have wondered if it were unthinkable to design a computer which would translate. Even if it would translate only scientific material (where the semantic difficulties are very notably less), and even if it did produce an inelegant (but intelligible) result, it would seem to be worthwhile. […] One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode". (Warren Weaver (1949, p.16))

Since March 4, 1947, when Warren Weaver suggested a possibility of machine translation in a memorandum to Professor Norbert Wiener of Massachusetts Institute of Technology (Weaver, 1949), it took 40 years to shift from Rule-based Machine Translation (RBMT) and Example-based Machine Translation (EBMT) to Statistical Machine Translation (SMT) that was first reported in 1988 (Brown et al., 1988) and ruled the MT field for the next 25 years[1] until the first debut of Neural Machine Translation (NMT) in 2014, showing potentials in the world (Sutskever et al., 2014)). The biggest turning point in the MT field was perhaps the infamous ALPAC report (Automatic Language Processing Advisory Committee) initiated by the US government in 1964 (Hutchins, 2015) that put a virtual end to all supports and funds for the relevant studies and divert attention to a more "realistic" field of study such as Computer Linguistics and Artificial Intelligence (AI).

> We have already noted that, while we have a machine-aided translation of general scientific text, we do not have useful machine translation. Further, there is no immediate or predictable prospect of useful machine translation. […] The Committee cannot judge what the total annual expenditure for research and development toward improving translation should be. However, it should be spent

[1] The SMT engines that have been reported to be the state-of-the-art in this context include all hybrid versions that combine RBMT or EBMT to SMT, but the core technology is based on SMT.

> hardheadedly toward important, realistic, and relatively short-range goals. (ALPAC (1966, p.33))

In fact, it was a death sentence from the viewpoint of SMT but a life sentence from the point of view of NMT. On the one hand, during the post-ALPAC age, as Hutchins (2015) would call it, the study of MT kept being on-going minorly inside the USA and mostly outside, such as Canada, Japan and other European countries such as Germany and France. With the increasing demands for translation influenced by the tide of globalization in the ‘70s and the appearance of personal computers in the early '80s, however, MT gained worldwide popularity back and Korean started to be studied in the field (Hutchins, 1995). On the other hand, it was also in the ’80s when researchers in the field of AI attempted to apply AI algorithms to MT as one of their sub-fields (Hutchins, 2015). It was in 2015 on the Workshop on Statistical Machine Translation (WMT) competition when NMT officially beat the conventional SMT engines on the *en-de* (Germany) translation (AWS, 2017), and in the subsequent year many large corporations removed the traditional SMT basis from their free MT service and moved on to the new basis: NMT.

1.2. NMT, is it a threat?

In 2016 when many free MT engines began to adopt NMT instead of SMT, from Google Translate (Wu et al., 2016) down, many users would be in awe of the performance of the machine translators. It was a whole new experience that was drastically different when they first encountered Babelfish, a machine that translated automatically for free introduced in 1997 (Hutchins, 2005). While it was a happy-shock for many, it was definitely not for most of the translators. The job automation probability of interpreters and translators calculated by a group of University of Oxford was 0.38 in 2013 (Frey and Osborne, 2013), but reached to a point where they were considered to be one of the eight jobs that were at the risk of extinction by 2030[2]. Of course, some insist that the development of AI is not what threatens us, but what we can take advantage of, as their ability is limited yet to conveying a general meaning of a text (Puchała-

[2] https://www.huffpost.com/entry/8-jobs-that-will-go-extin_b_8010878

ladzińska, 2016).

The discussion is more furious in Korea, with too many journals inciting unfounded beliefs about the perishability of the translating job (Y.Song, 2018). According to a survey[3] conducted to a group of young office workers and student applicants (4,174 people) in 2018, the first job at the risk of extinction was a translator (31%). Moreover, the fear has been rapidly propagated especially after the Google DeepMind challenge match in March 2016 where the computational intelligence AlphaGo beat a top-ranked Korean professional player, Sedol Lee in the game of Go (C.Lee et al., 2016).

In the academic field, the threat that translators start to feel is more evident. Figure 1.1 shows the total number of articles registered in the Korean Citation Index (KCI) from the year 2004 to 2018 in the humanity field with MT, MT evaluation and MT post-editing as a keyword. The graph shows that the number of articles in relation to MT increased considerably since 2016, accounting for 47.4% of the total sum. Consequently, the interests in MT evaluation and post-editing have emerged around 2017.

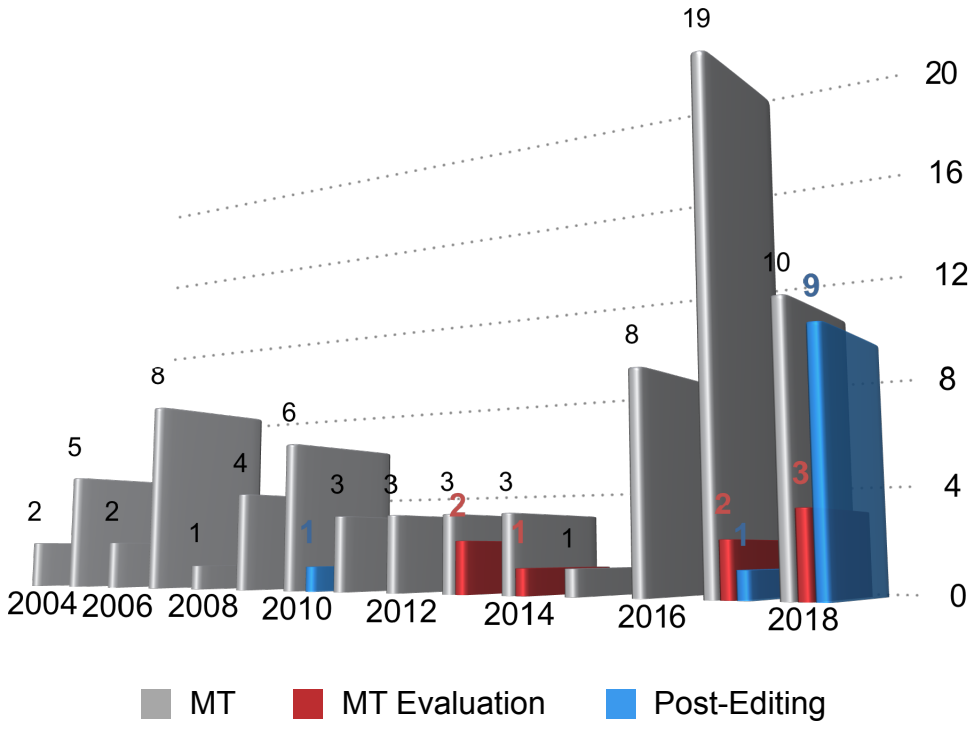


Figure 1.1: Distribution of papers registered in KCI (2004 - 2018) that have MT, MT evaluation or MT post-editing as a keyword (as of January 2019).

[3] https://www.mk.co.kr/news/economy/view/2018/04/210625/

1.3. The Spanish-Korean Language Combination

Korea was registered as the second country with the highest number of patents in AI technologies in the period of 2010-2015 with 17.5, right after Japan (27.9), according to OECD statistics (OECD, 2017). Contrary to the high dominance in AI, however, their interests in MT and post-editing in the practical and academical translation hemispheres have been exceptionally undersized, as seen in Figure 1.1. There are a total of 78 articles in relation to MT, among which 8 papers (10.3%) have ‘MT evaluation’ or 'MT error analysis’ in their titles and 2 papers about post-editing (14.1%)[4]. There is a good chance that more studies of such kind might have been omitted from the KCI server, but considering the status of this institution, there will not be a drastic change in the current tendency despite their possible existence. Many, furthermore, have declared that post-editing is in a nascent stage of development in the Korean academic circle (S.Lee, 2016; H.Choi & J.Lee, 2017; S.Lee, 2018a; S.Mah, 2018, to name a few).

As much new as the interests in MT evaluation and post-editing are, the interests in the *es-ko* pair in this field of study have been ignored. There is no research in terms of the *es-ko* language combination not only in MT but also in MT evaluation and post-editing. In this situation where not a single precedent is available, an intriguing question comes across: Is it possible to predict the result of the *es-ko* pair from the studies of Spanish-to-English and English-to-Korean? Such inference can be useful in that there are many studies about the *es-en* pair and not many but some in the *en-ko* pair. In addition, this section serves as an introduction of the most relevant studies of the *es-ko* pair in an indirect way of dividing it into the *es-en* and *en-ko* cases.[5]

***ES-EN* Pair** Many studies found that the *es-en* pair was one of the most desirable language pairs in European languages, more than French (*fr*), German (*de*) and Czech (*cz*), that facilitated a high performance of SMT

[4] S.Lee (2017) “What can we learn from trainee translators’ post-editing?”, J.Lee (2018) “A study for post-editing education: Difference between post-editing and human translation”.

[5] We are aware of the fact that the automatic evaluation scores are not comparable when the experiment setup is dissimilar (Doherty, 2016). The intention of our analysis is to get a general idea about the *es-ko* pair by means of two separate studies of the *es-en* and *en-ko* groups.

	All Corpora			News		
	Fluency	Adequacy	BLEU	Fluency	Adequacy	BLEU
es-en	3.19	**3.71**	**28.18**	+0.18 0.19	**+0.28 0.12**	**27.92 0.94**
fr-en	**3.25**	3.61	26.09	+0.23 0.09	+0.15 0.14	21.95 + 0.94
de-en	2.87	3.10	21.17	**+0.30 0.12**	+0.21 0.12	18.87 + 0.84

Table 1.1: Average manual and automatic evaluation scores of the three European languages and English in all / news corpora in Koehn and Monz (2006).

engines. For instance, Koehn and Monz (2006) performed manual and automatic MT evaluation of 10 MT engines for French, German, and Spanish to English and back in the Europarl and news (editorial) domain. The manual methods were composed of fluency and adequacy scores, and the automatic method was BLEU. The result indicated that the average scores that the *es-en* pair obtained in all corpora were superior to any other language pairs, with the exception of the fluency score. As given in Table 1.1, the *fr-en* pair obtained a slightly higher score (3.25) over the *es-en* pair (3.19). When the score was compared in the boundary of the news domain, the *es-en* pair still achieved the highest rank in the adequacy score and BLEU, but the highest fluency score was given to the *de-en* pair.

	A	*F*	*R*	*C.R*	METEOR	BLEU	TER	GTM
es-en	0.566	0.543	0.537	0.312	**0.661**	**0.346**	**0.480**	**0.528**
fr-en	**0.576**	**0.596**	0.494	0.312	0.602	0.279	0.405	0.478
de-en	0.552	0.56	0.563	**0.344**	0.577	0.242	0.339	0.459
cz-en	0.550	0.592	**0.627**	-	0.581	0.241	0.355	0.460

Table 1.2: The result of human and automatic evaluation scores of *es-en, fr-en, de-en* and *cz-en* in the news domain (Callison-Burch et al., 2007). Of all engines that were under evaluation, only the result of the best case was given in this table. (*A*=adequacy scre; *F*=fluency score; *R*=ranking score; *C.R*= constituent ranking score).

Callison-Burch et al. (2007) extended the size of the study by implementing a wide range of human and automatic MT evaluations of eight MT engines for French, German, Spanish and Czech to English and back. The human metrics included fluency and adequacy scores, a relative ranking

comparison and a constituent ranking, and eleven different automatic metrics were applied. The results indicated that the *es-en* pair acquired the inferior scores of the four language groups in all human evaluation in both Europarl and news (editorial) domain (see in Table 1.2). It, on the other hand, marked the highest score in all eleven automatic evaluation when the news domain was solely considered. In Europarl, the *fr-en* outperformed the *es-en* group, albeit with a minor difference. Unlike the result of Koehn and Monz (2006), the *es-en* language pair in Callison-Burch et al. (2007) exhibited its powerful aspects only in the automatic evaluation.

Liu et al. (2011) reinforced such finding in their experiment by introducing two new metrics —TESLA-M and TESLA-F— in the newswire domain. The scores shown in Table 1.3 were the highest scores that each language pairs could obtain after tuning and testing with various combinations among the four metrics. The *es-en* pair scored the highest compared against the *fr-en* and *de-en* groups with the exception of TER.

	BLEU	TER	TESLA-M	TESLA-F
es-en	**0.5667**	0.5725	**0.4511**	**0.4409**
fr-en	0.5239	0.6028	0.4170	0.4224
de-en	0.4963	**0.6329**	0.3784	0.4070

Table 1.3: Four automatic evaluation scores of *es-en, fr-en, de-en* and *cz-en* pairs in the news domain (Liu et al., 2011). The scores arranged in this table represent the highest records of each language combination when tuned and tested by the combination of the given automatic metrics.

Considering that the automatic evaluation employed in Callison-Burch et al. (2007) and Liu et al. (2011) were mostly dependent on word similarities, the two studies manifested that i) Spanish and English shared a word-level isomorphism and ii) such similarity had a positive influence on automatic evaluation. In fact, Papineni et al. (2003) had claimed that an n-gram match in the *es-en* pair could go up to 16-gram in their dataset. Unlike the positive outcome, the result of human evaluation did not support such claim and varied case by case. Thus, it could be said that the es-en pair was exceptionally stronger than pairs with other European languages in automatic evaluation of

SMT engines in the news domain.

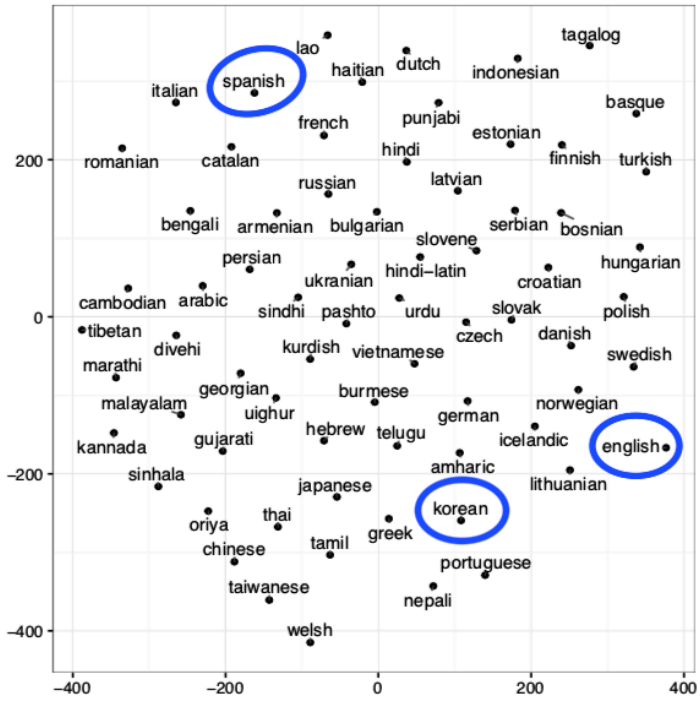


Figure 1.2: t-SNE plot of language embedding vectors (Jaech et al., 2016).

Such lexical similarity between Spanish and English seemed to be of utmost help to enhance the performance of NMT. Conneau et al. (2017) pointed out that the *es-en* pair showed the highest performance of 83.3% accuracy in learning cross-lingual word embeddings without the help of parallel corpus. It was much higher than other pairs such as *fr-en, ru*[6]*-en, de-en, zh*[7]*-en* and *eo*[8]*-en* in both directions. The positive impact of isomorphism of the es-en pair to NMT was reconfirmed by Zhang et al. (2017) who contrasted the evaluation result to European and non-European languages such as *it*[9]*-en*, *jp*[10]*-zh* and *tr*[11]*-en* groups. The two studies showed that the lexical similarity of Spanish and English could help to achieve high scores in NMT evaluations. However, when all linguistic features were considered, the two languages turned out to be

[6] Russian

[7] Chinese

[8] Esperanto

[9] Italian

[10] Japanese

[11] Turkish

considerably dissimilar. Jaech et al. (2016) showed in their T-distributed Stochastic Neighbor Embedding (t-SNE) plot of language embedding vectors where languages were represented in the vector space according to their linguistic similarities (Figure 1.2). Spanish and English were represented in a quite long distance. A rather surprising finding was that Korean was in a shorter distance to English than Spanish to English.

***EN-KO* Pair** Paul et al. (2013) focused their attention on the language pairs by measuring a BLEU score across 12 Indo-European languages and 10 Asian languages. A source language that achieved the highest score was Spanish, and a target language that achieved the highest score was English. Such high scores were dropped drastically when Korean was involved in either side. They reported that the BLEU score took a dive up to 35 points. In Table 1.4, the score of the *es-en* translation achieved 74.4, but the figure decreased to 40.8 in the *es-ko* translation. Similarly, the *en-es* translation obtained 75, which was reduced to 39.8 in the *ko-es* translation.

SL | *TL*	Spanish	Korean	English
Spanish	-	40.8	74.4
Korean	39.8	-	40.4
English	75	42.9	-

Table 1.4: BLEU scores of English, Korean, and Spanish in both directions Paul et al. (2013). (The source language (SL) for the rows and the target language (TL) for the columns).

It was hard to draw a conclusion from a single study. However, considering that no more studies were found in terms of the *en-ko* combination in the area of MT evaluation, it was speculated that the influence of Korean to the automatic metrics was detrimental in SMT. Therefore, with the previous studies, the only thing that was revealed was a possible performance of SMT in the given language pair, which, however, should only be answered by direct evaluation of the intended language pair.

Meanwhile, no precedence in terms of NMT evaluation of such kind in the

en-ko pair was found[12]. In fact, studies about the *es-en* and *en-ko* pairs could not give much insight especially in NMT due to the nature of its architecture: performance of NMT engines was largely dependent on a size of data. In other words, the size of the dataset of the two languages should be directly considered. In that context, a new question arose as to: could English contribute to increasing the size of the dataset of *es-ko* pair? The answer was intensively studied as a form of *Pivot NMT*. The pivot-based approaches alleviated the data scarcity problem of low-resource language pairs by means of training source-to-pivot and pivot-to-target translation models in an independent manner. Various ways of relating the two models increased or decreased the performance of NMT engines. More in-depth theories are provided in Chapter 2.3.2.

In Cheng et al. (2017), the *es-fr* and *de-fr* translations were compared to their pivot translation of having English as a pivot language. In both Europarl and WMT corpus (which included the news domain), BLEU scores were higher in all datasets for the *es-en* and *en-fr* translations (see in Table 1.5), proving the effectiveness of pivot NMT in the *es-fr* language pair.

		Pivot NMT		NMT
		es->en	*en->fr*	*es->fr*
Europarl	Dev-set	**31.53**	30.46	29.52
	Test set	**31.54**	31.42	29.79
WMT	Dev-set	27.62	**27.90**	24.92
	Test set	**29.03**	25.82	24.60

Table 1.5: Comparison of BLEU scores between a pivot translation (*es-en* and *en-fr*) and a direct translation (*es-fr*) in an NMT engine in Europarl and WMT corpus Cheng et al. (2017). (unit: %)

Google Inc., as a matter of fact, was already engaged in these fields to deal with various resource-scarce language pairs inclusive of Korean. They presented a negative effect of English as a pivot language when it came to a Portuguese-to-Spanish translation presumably for their closer linguistic similarities than with English (Johnson et al. 2016). So, they tested a Spanish-

[12] There are some studies that deal with error analysis.

to-Japanese translation with English as an intermediate pivot language in an identical setting, and reported a poor result of a big fall of the BLEU score to 18.00 (see Table 1.6). Such a result was especially meaningful to the current thesis on the grounds that Japanese was one of the closest languages to Korean. How would the model perform for the *es-ko* language pair? Would the poor performance be improved by now in Google Translate? These were one of the issues the current thesis strived to clarify.

	pt-es			*es-jp*
	PBMT pivot	NMT pivot	NMT	NMT pivot
BLEU	28.99	30.91	**31.50**	18.00

Table 1.6: BLEU scores of pivot NMT vs. direct NMT in *pt-es* and *es-jp* language pairs (Johnson et al. 2016).

1.4. Objective

With the rapid growth of NMT and its potential power on the distant language pair, it is important to confirm the performance of NMT in the *es-ko* pair and take advantage of the technology in the course of translation. The necessity also comes from the fact that there are no studies at all in terms of NMT evaluation in the aforementioned language pair up until now. In this context, the principal objective of the dissertation is twofold. In the first place, it aims at reporting on the level of quality of the new state-of-the-art MT system, the NMT, in the *es-ko* translation. The performance, which is represented by two big branches of **accuracy** and **efficiency**, will be measured by a series of MT evaluation methods. In the second place, it attempts to investigate the cause of the errors by way of meticulous error analysis of NMT outputs and post-editing outputs and to address them in a linguistic stance. To accomplish the two meta-objectives, sub-level objectives are pointed out under the two categories:

I. **We assess the performance of NMT for the *es-ko* language combination in the news domain.**
   - It performs a series of evaluation experiments to obtain hands-on data for the given language pair.

- It calculates the level of accuracy of the output generated by *Google Translate* of version 2018 based on a variety of human and automatic evaluation methods.
- It employs HTER as the core automatic metrics based on post-editing results acquired during the evaluation experiment.
- It measures the efficiency of the Google NMT engine by carrying out a translation from scratch and measure the time taken for the task to compare to the machine's case. The time taken for developing their algorithm, however, is exempt from the discussion.
- It performs an in-depth analysis in a quantitative and qualitative manner and proposes a meaningful remark. The qualitative analysis includes feedback from the post-editors on their impression about the overall post-editing task.

II. We attempt to carry out in-depth error analysis of NMT outputs, as well as post-edited outputs.

- It proposes binary error analysis under the title of Analysis I and Analysis II. Analysis I works on errors produced by the NMT engine. Analysis II utilizes the post-edited data and detects errors from the post-editors' perspective.
- To accomplish such analysis, it devises an adapted version of error taxonomy for the *es-ko* pair and news domain from the MQM-based model of TAUS.
- Analysis I is approached both quantitative and qualitatively by detecting some presumable tendencies throughout the errors of each high-level categories and suggesting exemplary sentences.
- Analysis II answers to the question of what types of errors are considered important in post-editing and impede human understanding.
- The two analysis sheds light on the disharmonious conclusion of the experiment —of a robust NMT performance but a low post-editing productivity.

1.5. Limitation

- The core methodology of the current thesis is human evaluation methods. The core reason for the limited use of automatic methods is due to i) the size of the text that is not enough to acquire a credible score and ii) the notoriously low correlation with the human judgment of the automatic methods when it comes to a distant pair.
- The current thesis is not concerned with the superiority of human or automatic methods. We adopt what is believed to be the most adequate to the given circumstances.
- For the same reason, we adopt HTER assuming that it is one of the most reliable methods to the language pair that can compensate for the absence of automatic ones.
- The data obtained from a hands-on experiment is based on a small dataset of about 6,500 words from six people due to a budget limit.
- At the same time, the NMT engine is tested solely on the news domain, leaving unexplored domains as a promising future work.
- Post-editing in our experiment is for research purpose only. The applicability of the suggested means to other fields, i.e. for educational purpose, has to be addressed additionally in independent research.
- The principal focus of the current study is data collection and data analysis, in such a way that discussions over computational theories including the detailed architecture of evaluation methods and MT engines are distant issues and will not be dealt with.
- In the same context, the error analysis serves as a means for detecting errors in the given dataset. It is, thus, not our goal to systemize the error typology or create a new one for Spanish and Korean.

1.6. Thesis Structure

The current thesis is composed of three parts and eight chapters. **Part I: Theories** offers theoretical background information prior to the experiment. The first half of Part I is dedicated to the architecture of NMT, and the second half

deals with MT evaluation theories. **Part II: Experiment** reports on the details of the MT evaluation experiment. The first half describes a pilot study and the second half explains the main experiment composed of various manual and semi-automatic methods. **Part III: Analysis** discusses two-fold analysis of the acquired data —MT evaluation analysis and error analysis— in a quantitative and qualitative manner.

In **Chapter 1**, we narrate the background motives of the thesis, pointing out a dissimilar origin of NMT from SMT and hypothesizing its performance in the *es-ko* pair. We bring up the two meta-goals of this thesis divided further by sub-goals. We also suggest the limitations of the scope of our study.

In **Chapter 2,** where Part I begins, we describe the architecture of NMT starting by a basic unit of AI, the artificial neuron, that is conceptualized from a human brain cell. After explaining key algorithms of artificial neural networks in Machine Learning, we present how and when they are grafted onto the MT field, resulting in the birth of NMT. As a new state-of-the-art, some fundamental NMT models are illustrated along with their merits and demerits. We stress their influence on Korean found in the previous studies. We, then, shift our focus to the architecture of Google NMT, which is the main subject of our experiment. We investigate their initial NMT model and the most up-to-date ones.

In **Chapter 3,** we introduce a theoretic framework in MT evaluation in general because no particular methods for NMT engines exist yet. After working on the relevant theories of a definition of equivalence in MT evaluation, methodologies, and their advantages and disadvantages, we conclude that human and semi-automatic evaluation methods are suitable for the *es-ko* pair. We, then, focus on the five methods that will be employed in the current thesis —fluency scoring, adequacy scoring, segment ranking, post-editing, and HTER. As an additional work, we conduct a survey to delve into the current usage of MT and post-editing in a Korean-related language pair in both global and domestic (Korean) markets.

In **Chapter 4** where Part II begins, a pilot study is proposed. Having defined the objectives of the pilot study, we describe what it is composed of, how it is implemented and the result we obtained. The chapter ends with a reconfirmation of the pre-set goals stating that all the parameters have met

successfully.

In **Chapter 5**, we describe the design and conduct of the main experiment for NMT evaluation in the *es-ko* pair. We detail the parameters of the experiment, such as the dataset, evaluators' profile, and TAUS workbench. And then, we explain how the experiment has been conducted in chronological order. The tree evaluation methods of post-editing, fluency & adequacy scoring, and segment ranking are specified respectively in that regard. Additionally, we report on feedbacks collected from the evaluators regarding their perception on the task in general and the performance of NMT and the post-editing job in the *es-ko* pair in particular.

With the data at hand, in **Chapter 6** where Part III begins, we launch an in-depth analysis and reports on the performance of the NMT engine in a quantitative and qualitative way. Firstly, the fluency and adequacy scoring result shows how fluent and adequate the system translation is. Moreover, we inquire into the correlation of such scores with sentence length. Secondly, the ranking score of GNMT is proposed in contrast to an NMT system of Kakao i and human translation. Finally, the post-edited data are studied in terms of time and effort. By comparing the time taken from post-editing and translation from scratch, we suggest a post-editing productivity. The post-editing efforts are obtained from temporal and technical perspectives with the help of HTER scores.

In **Chapter 7**, error analysis is carried out in a binary way: on the system translation and on the post-edited system translation. Prior to the study, we adapt the error classification to the purpose of our study and suggest 10 error classes. Based on such error typology, we first report on general error analysis on the system translation produced by GNMT. We monitor the most/least common error types and the behavior of the NMT engine in that regard. We additionally initiate the second error analysis on a post-edited version of the system translation. We observe error types detected by the post-editor and see what error types are the most concerned in post-editing.

In **Chapter 8**, we draw a conclusion of the thesis. We summarize new findings in the *es-ko* pair and state our contribution to the MT research community.

[blank page]

PART I. THEORIES

Part I is dedicated to a theoretical study. As the main goal of this thesis is to evaluate the performance of an NMT engine and to analyze the errors, we believe that fine-grained background studies should be preceded before the experiment in a realm of NMT and MT evaluation.

Theoretical background in NMT is provided in Chapter 2, covering from Machine Learning to MT. We present the mainstream ANN models and their application to MT, which is denominated as NMT. While presenting fundamental NMT models, we investigate the most up-to-date models. In the meantime, we narrow down the scope of our study to multilingual NMT models that are specialized in low-resource languages. We put special attention on zero-shot approach designed by Google.

In Chapter 3, the most relevant theories in MT evaluation and post-editing are described. Starting from a definition of equivalence in MT evaluation, we introduce the existing MT evaluation models and discuss their advantages and disadvantages. We detail the three human evaluation methods —fluency & adequacy scoring, segment ranking, post-editing— and the semi-automatic method, HTER. As an additional work, we launch a survey in terms of the current usage of MT and post-editing in a Korean-related language pair.

[blank page]

2. NEURAL MACHINE TRANSLATION

> “Unlike the traditional phrase-based translation system which consists of many small sub-components that are tuned separately, neural machine translation attempts to build and train a single, large neural network that read a sentence and outputs a correct translation.” (Bahdanau et al., 2014)

The official introduction of SMT back in around 1988 was a tremendous shift of *concept* in the MT field, from relying on linguistic rules designed by a human to enabling translation chiefly by pre-existing data (Hutchins, 2015). In SMT, human lost a large part of control over the machine because with data, in theory, it could align the source and target texts and compute the similarity to come up with an output. The role of human, of course, was still crucial but was shoved to the side of the realm, developing parallel corpus, selecting suitable algorithms and tuning parameters. The shift from SMT to NMT, meanwhile, was as extreme as the former transformation in the MT world from a point of view of *technology*. Although both SMT and NMT were founded upon the identical idea of taking advantage of the available data, NMT took control away even more from human over machine by facilitating self-control in some sense. That is, it was a shift from an automatic machine to an organic machine. If the change from RBMT to SMT could be understood as a *paradigm shift*, the change from SMT to NMT was a *technological shift*.

In that sense, this chapter attempts to address technological features of NMT in general and of Google NMT in particular that make NMT distinctive enough to claim a new state-of-the-art, but in a lay-public-friendly manner limiting computational discussions aside. Chapter 2.1 addresses Artificial Neural Networks, a Machine Learning algorithm and at the same time a core NMT algorithm. Chapter 2.2 defines NMT and introduces mainstream NMT models. Then, some studies that claim NMT to be state-of-the-art in a Korean translation are described. Chapter 2.3 narrows down the scope to Google NMT models, one of which is the main subject under evaluation of this thesis.

2.1. Artificial Neural Networks

The key algorithm of NMT is artificial neural networks, or ANNs, that is an integral part of various Machine Learning and DeepL tasks. It is believed that the first step to understanding the architecture of NMT is to know the concept and mainstream models of ANNs. To this aim, this chapter introduces the nature of ANNs from the birth of the concept coming from biological nerve cells (Chapter 2.1.1) to the improvement of major ANN models starting by Perceptron to Recurrent Neural Networks (Chapter 2.1.2), to the application of the neural networks to MT from a partial to a full way that opens a new chapter of MT history (Chapter 2.1.3).

2.1.1. Origin of Concept Borrowed from Neuroscience

While the idea of *a machine that thinks* could be traced back to as far as Homer mentioning “mechanical tripods waiting on the gods at dinner” or the novel *Wizard of Oz* by L. Frank Baum, the conceptual discussion started to be discussed in the late ’30s and the materialization of the idea came in the early ’50s when Alan Turing proposed in his writing a possibility of realizing a thinkable machine by designing a famous experiment called *Turing’s Test* (Buchanan, 2005). The idea behind the test was that the intelligent behavior of a machine could be witnessed by interrogation of the machine, and if a human did not conceive it as a machine, it had an intelligence (Russell and Norvig, 1995). Such a concept is what is now called Artificial Intelligence, Intelligent Machine or Machine Intelligence.

The level of AI was also discussed as to what an intelligent act meant. That was, to be an agent with intelligence, was it enough that it thinks or should it think and behave intelligently? Moreover, in what way an agent should do, humanly or rationally? All these questions concerning the definition of AI have been fierce[13], but it is clear that the agent should demonstrate intelligence that is “in contrast to the natural intelligence displayed by human or animals”[14] and that the intelligence should be comparable to the intelligence of human being

[13] Those who are interested in the dispute over the definition of AI are directed to Russell and Norvig (1995).

[14] https://en.wikipedia.org/wiki/Artificial_intelligence

(Russell and Norvig, 1995)".

To this aim, AI is required to perform various tasks such as knowledge reasoning, planning, natural language processing, etc., including machine learning. There are five approaches in machine learning: symbolists, connectionists, evolutionaries, Bayesians, and analogizers, each of which has master algorithms of inverse deduction, backpropagation, genetic programming, Bayesian inference and support vector machine, respectively (Domingos, 2015). The Artificial Neural Networks or Artificial Neural Nets (ANNs) that is going to be detailed in this section is a core algorithm of the school of connectionism who has a master algorithm of backpropagation, and the so-called *artificial neuron* is a fundamental component that constitutes ANNs. The connectionists came up with the idea that for an intelligent agent to be able to behave like a human ('behave' in a sense of both thinking and acting and 'like human' meaning from a human-centered approach), the best strategy to achieve the goal is to mimic the principles of biological neurons of the human brain. As such, understanding the artificial neuron starts from knowing the conceptual model of human brain from the perspective of neuroscience.

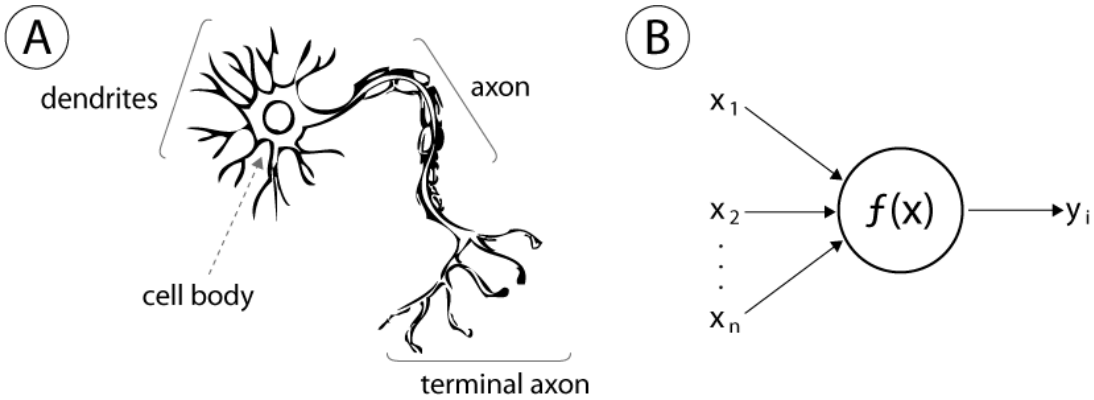


Figure 2.1: A human biological neuron (A) with the name of each part and how the concept is introduced in an artificial neuron (B) (Maltarollo, 2013).

There are approximately 12 billion neurons in the human brain in the form of myriads of *connections* to and from a cell body as shown in Figure 2.1A. The interaction among neurons or nerve cells is based on the transmission of chemical and electrical signals that flow across the neurons. Dendrites receive the chemical signals from axons of other neurons and transmit them to their

axons where the signals are updated by adding or subtracting information and forwarded to synapses that border dendrites of other neurons (British Neuroscience Association, 2003). This biological system is interpreted in computer science in a more compact way to form the artificial neuron that comprises of multiple *nodes* or *units* like dendrites and axons that were connected by *links* (Russell and Norvig, 1995). In Figure 2.1B, the signals with information represented by $x_1, x_2 \dots x_n$ are transmitted from one cell (or node) to another through a certain function $f(x)$, and the information is updated and saved to y_i.

Not only were the connectionists inspired by the feature of the nerve cells, but also they put considerable attention on the process of communication between the neurons enabled by $f(x)$ called *action potentials*. It refers to a change of state of signals in a neuron that is known to usually happen in axons (British Neuroscience Association, 2003). In order for a neuron to fire, it needs to activate the action potentials by means of changing the concentration of ions in the axonal membrane. Such action strengthens the connection of the cells. This procedure has a similarity with the binary system of a computer that processes information with 0 and 1. If the signal does not reach the threshold, a system outputs a 0, and when it goes over the limit, the system outputs a 1. This so-called threshold theory can be mapped into a function in Figure 2.2 where all inputs that are above the line (or function) will have a 1 as an output and those that are below the line get a 0 as an output. In such a way, new information is updated in a way that more information is added or subtracted in each node depending on a level of the bond between the nodes (Jackson, 1985). In other words, the action potentials also decide the relevance between the nodes by monitoring their bond, which forms a *variable weight* in ANNs. These two processes of communication and memory update of neurons constitute the core concept of ANNs. The initial model of the artificial neuron was officially presented in 1943 by Warren McCulloch and Walter Pitts. Years later in 1957, an American physiologist Frank Rosenblatt completed the model by adding a concept of *weights* to the artificial neuron. It was the moment of birth of *perceptron*, a single neural network that could actually learn data.

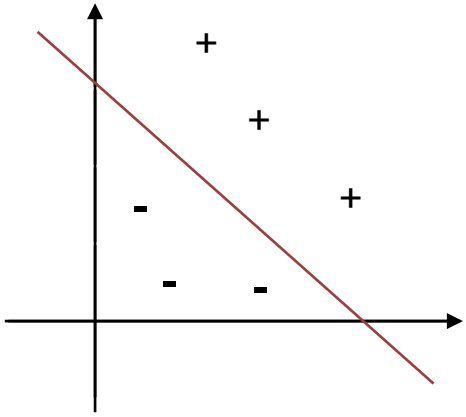


Figure 2.2: A graphical presentation of threshold theory that explains action potentials (Domingos, 2015).

2.1.2. Evolution of ANN Models (in relation to MT)

In this section, we introduce the most basic ANN model with a single layer, Perceptron, and its developed version with multiple layers, Feedforward Neural Networks. And basic models that can learn sequential data, Recurrent Neural Networks and Attention Mechanism, are described, subsequently.

2.1.2.1. Perceptron

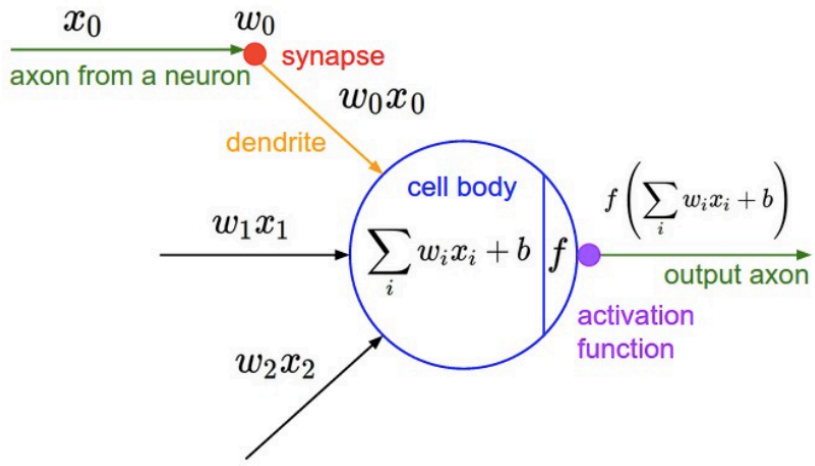


Figure 2.3: Architecture of perceptron of Rosenblatt in comparison to the biological action functions (Kurenkov, 2015).

Perceptron, the algorithm motivated by the cognitive ability of the human brain to 'perceive' things, was a single-layered artificial neuron. The initial version of perceptron (that of the 1950s) was effective in classifying data into a binary pattern with a linear separation as presented in Figure 2.2. Shaped from the

basic idea expressed in Figure 2.1B, the architecture of perceptron in comparison to that of the biological neural cell is detailed in Figure 2.3. The activation function of perceptron also shown in the given figure is more briefly explained with the following formula:

$$f(x) = 1 \text{ if } wx + b > 0,$$
$$0 \text{ otherwise} \tag{2.1}$$

where w, x, and b represent weight, input, and bias, respectively. If the weighted sum of inputs and biases are bigger than a 0 the function returns a 1, and if not, it returns a 0. The key concept in the algorithm is that for each input variable x, some form of weights w is calculated randomly to represent the relevance of each input to an output, and bias b decide the minimum boundary of the activation functions. In other words, the computation is understood as a linear function of $y = ax + b$, where the value a decides the degree of the functional slope and the value b is in charge of the lowest number of the value y when the value x is 0. As such, the weights change the slope of the borderline and the bias set a minimum number of activation functions in terms of a weighted sum of the inputs.

2.1.2.2. Feedforward Neural Networks

Based on a linearly separable function, however, the model had a critical limitation of being unable to represent more than two patterns while most of what happened in a real world required a complex classification. That was what two knowledge engineers officially pointed out about this algorithm (Minsky and Papert, 1969), and it took about 15 years to come up with a refined model known as *Multilayer Perceptron* (Hinton and Sejnowski, 1985) proposed by three professors of Geoffrey Hinton, David Ackley and Terry Sejnowski in 1985. This new model that alleviated the issue of perceptron was Feedforward Neural Networks that was known to be the most basic type of ANNs and became a new definition of perceptron. The difference in contrast to the previous model laid in the fact that more layers were added to tackle complex classifications. And with

that, the pivotal point of this model was focused on how to process the vast amount of procedures with efficacy and accuracy. The key technique in charge of the aforementioned issue was *Gradient Descent* with the help of *Back-propagation*.

Here is a good example that shows how gradient descent and backpropagation work, presented by 3blue1brown[15]. In handwriting recognition, a model has to distinguish a couple of numbers handwritten by a human. The process of handwriting recognition starts from training the model with a large volume of annotated data aligning actual human handwriting images and their expected outputs. This stage is where a machine “learns” data with certain Machine Learning algorithms. After the training, the model gets engaged in an actual task with a test set that requires answers for random human handwritings. When trained, the machine with feedforward neural networks is fed with an input of an actual handwriting sample of a number, i.e. ( , 4), decomposed into various pixels, each of which is in the form of some numeric vectors from 0 to 1 depending on their shades of black and white. The bond between the nodes is represented into weights and biases, as explained in advance. Once it produces a candidate output, it is then compared to an expected output. The model enhances the accuracy of the output by minimizing the differences between them, called *cost,* by gradient descent and back-propagation. That is, the cost is reduced in consideration of the previous path that each input has passed to monitor a level of the bond among the networks, and based on it the model re-calculates and tunes the cost in proportion. In this way of proportional representation, the process of learning is transformed from a simple binary procedure to a probabilistic matter.

2.1.2.3. Recurrent Neural Networks

Unlike the feedforward neural nets whose inputs behave independently as seen in the example of the handwriting recognition, there are tasks where previous inputs have an influence on the next inputs such as voice recognition and word prediction, not to mention machine translation. That is where Recurrent Neural Networks (RNNs) are applied to learn sequential data. The core feature of

[15] https://www.3blue1brown.com

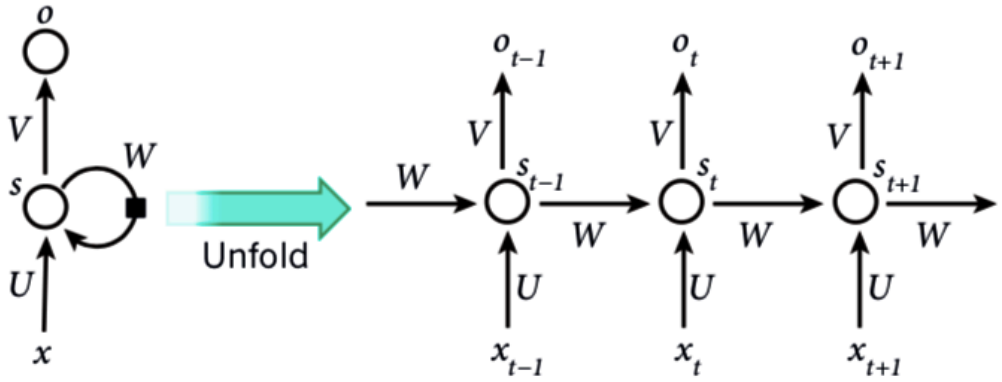


Figure 2.4. Cyclic process of Recurrent Neural Nets (LeCun et al., 2015).

RNNs is that the decision on the previous time step has an influence on the current decision in the form of a loop in the process (LeCun et al., 2015). To be specific, the model processes the input vector one at a time and preserves the history of each input at a given time step in the form of the *state vector* in a hidden layer. As in Figure 2.4, an input x in a time step t-1 is first mapped to a state vector s_{t-1} and produces an output o_{t-1}. In the next step t, a new input x_t is introduced along with the state vector s_{t-1} and s_t in the loop in such a way that the information of the previous data is preserved while the model keeps updated by the following equation:

$$h_{<t>} = f(h_{<t-1>}, x_t). \tag{2.2}$$

where the $h_{<t>}$ represents a hidden state of input x_t in a time step t. The activation function $f(x)$ can be "as simple as a sigmoid function (an S-shaped curve) and as complicated as LSTM (Information about LSTM will be given subsequently)" (Cho et al., 2014a). In other words, the vanilla RNNs enables to keep track of the previous history by updating the hidden state, as well displayed in the example (iamtrask, 2015)[16] given below:

[16] https://iamtrask.github.io/2015/11/15/anyone-can-code-lstm

```
(input + empty_hidden) -> hidden -> output
(input + prev_hidden) -> hidden -> output
(input + prev_hidden) -> hidden -> output
(input + prev_hidden) -> hidden -> output
```

where four inputs are expressed in four different colors in each time step, along with their hidden states and the outputs. In each time step, the previous hidden state (`prev_hidden`) is inserted together with the next input to preserve the previous history. The color illustrates how the previous information is kept in each hidden state vector. Note that in the hidden layer, the model preserves and removes the memory of sequential data in its own fashion, i.e. from `hidden` to `hidden`. In this way, RNNs can learn, for instance, a probability of the character `o` given the previous inputs `h-e-l-l` or a probability of the word `clear` given the inputs `The-sky-is`. This is a *recurrent* way that keeps track of the previous history.

2.1.2.4. Long Short Term Memory

The focal point of the evolvement of the models, after all, has shifted from reducing costs (or errors) (in feedforward neural networks) to preserving as many memories as possible at the same time reducing the computational costs (in RNNs). One of the bottlenecks of RNNs, however, is that the memory is considerably limited so that many of the previous states are lost, which results in an inaccurate translation. A part of the reason of such phenomenon is related to a vanishing and exploding problem of the architecture of RNNs that the values close to 0 tend to vanish while higher values tend to explode o infinity[17]. To alleviate the issue and improve the memory, a DeepL technique called Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) is applied. The primary idea of LSTM is that there is a unit called *cell,* which is in charge of memory at each time step, and the cell is regulated by three different-purposed gates —input gate, output gate and forget gate—, each of which controls to what extent the information at each time step should be preserved or removed. [18] Sutskever et al. (2014) applied LSTM on both sides of the encoder and

[17] https://en.wikipedia.org/wiki/Vanishing_gradient_problem

[18] https://en.wikipedia.org/wiki/Long_short-term_memory

decoder to map “variable-length inputs into a fixed-dimensional vector representation”. Such a technique was an integral part of processing sequential data because DeepL usually could not process inputs of multiple dimensionalities (Sutskever et al., 2014).

2.1.3. Application of ANNs to MT

The ANN models that are described in the previous section are just a small part of a vast number of ANN models and algorithms that exist in performing various Machine Learning tasks such as image recognition, word prediction, voice recognition and so forth. As the nature of such tasks is divergent, the importance is in finding a suitable algorithm and combining them in a way that they can perform the task with the best efficacy and proficiency. As such, the application of neural networks to MT is extended gradually from a small part of SMT to a whole architecture of MT. In that sense, a question arises as to what kind of application is defined to be neural-based MT. Starting from the earlier introduction of ANNs to the MT world, this chapter defines NMT.

The earlier stage of neural approaches to MT was in the form of substituting n-gram language models of SMT baselines. Bengio et al. (2003) took advantage of word feature vectors that were distributed depending on the lexical similarities and probabilities of word sequences calculated by the neural model. Similarly, the neural network language model of Schwenk (2007), also known as a *continuous space* language model, was also based on feedforward neural nets with two hidden layers. The basic idea was to use the probabilities suggested as an output by the neural model as a probability estimator of the language model. It could be regarded as another form of n-grams. Their promising results obtained from the involvement of neural-based language modeling in the process of SMT gained interests and facilitated further applications of ANNs to other stages. Devlin et al. (2014) applied ANNs to a decoding step so that neural nets became a fully-integrated part of SMT that could be implemented in any decoders (see Figure 2.5).

The break-point of NMT and SMT, however, was the development of a continuous neural-based model that did not involve SMT in the process and was solely run by ANN algorithms. This was the birth of the end-to-end type of

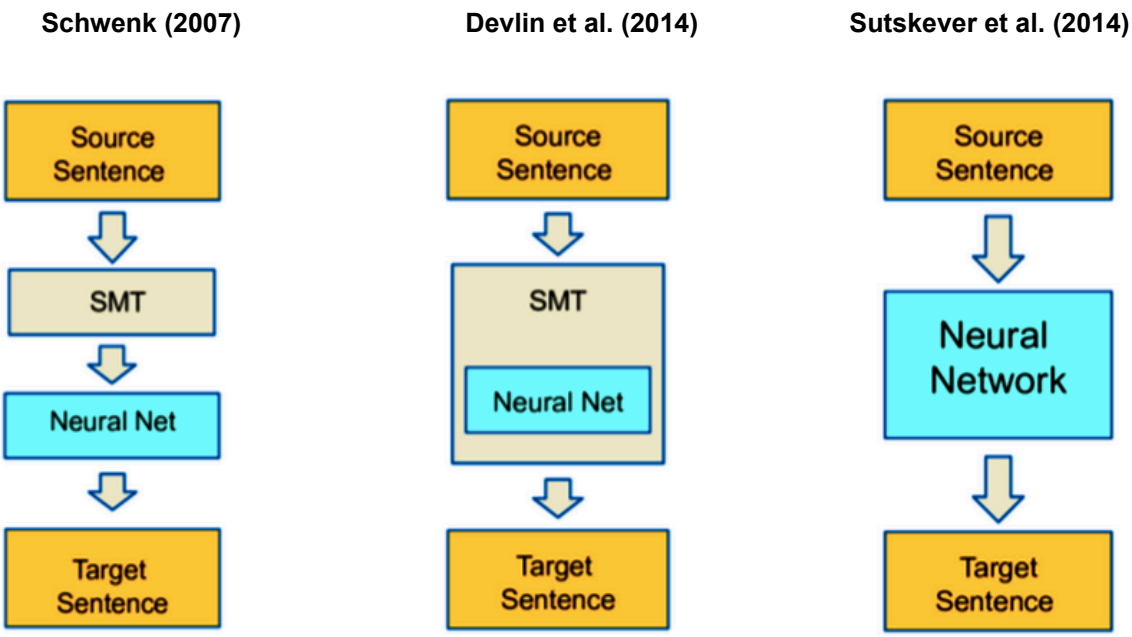


Figure 2.5: Level of integration of neural networks in MT (Stevens, 2016).

neural-based MT model initially proposed by many, but not limited to: Sutskever et al. (2014) who converted the source and target input of various length into a fixed-length vector representation, and learned them sequence by sequence based on two separate LSTMs, briefly explained in Chapter 2.1.2. This powerful DeepL technique made a purely-neural-based MT model possible, and it was called *sequence-to-sequence* model. Such a model constitutes the baseline of most of the NMT models, which will be further explained in the subsequent chapter. As such, although it has been long since ANNs was developed, the circumstantial and technological supports enabled a birth of the pure NMT model "that aimed at building a single neural network that could jointly tuned to maximize the translation performance, unlike the traditional statistical machine translation (Bahdanau et al., 2014)".

2.2. Neural Machine Translation

For the last five years after NMT achieved comparable performance to some state-of-the-art SMT models (Sutskever et al. 2014), newer and stronger models have been developed at an exponential rate. Behind the success of NMT was a materialization of DeepL, as described in the preceding chapters, enabled by the emergence of big data and advances in computer power. For NMT whose performance could be enhanced primarily by a volume of data (in

terms of quality) and computer power (in terms of efficiency), DeepL played a key role in processing more hidden layers at a more rapid pace (Schmidhuber, 2015).

With the basic knowledge in neural networks at hand, this chapter investigates NMT itself and its status in the MT field. Chapter 2.2.1 introduces mainstream NMT models starting from the sequence-to-sequence model. Chapter 2.2.2 argues the strengths and weaknesses of NMT that have been claimed and highlights three common traits of NMT that have been witnessed in many comparative studies between SMT and NMT. Chapter 2.2.3 presents some of the studies that claimed an outperformance of NMT over SMT and stresses why NMT should be evaluated in this thesis.

2.2.1. Mainstream System Architecture

The most basic framework of NMT that constitutes most of the NMT models is the encoder-decoder model that separates the translation process in two steps. In such binary-step approach, identical or different types of algorithms can be applied. The initial encoder-decoder model was proposed by Sutskever et al. (2014) and Cho et al. (2014a), who were inspired by the challenge that DeepL technologies posed in terms of mapping sequences of different length to a fixed vector representation. Such framework facilitated the effective use of RNNs in MT.

2.2.1.1. Sutskever's encoder-decoder model

For the sake of convenience, the NMT model of Sutskever et al. (2014) is named as Sutskever's encoder-decoder model or Sutskever's model. As shown in Figure 2.6, the input sequences *A, B,* and *C* are learned one at a time step in an encoder to acquire a large fixed-dimensional vector representation *W* that is decompressed in the decoding stage to produce output sequences *X, Y* and *Z*. The process can be understood as a combination of many-to-one and one-to-many ANN models based on two different LSTM units that better capture long-term dependencies than RNNs. The key benefit of this model is that the computational costs are reduced by employing two independent ANNs. Moreover, compressing all relevant information to some numeric vector enables

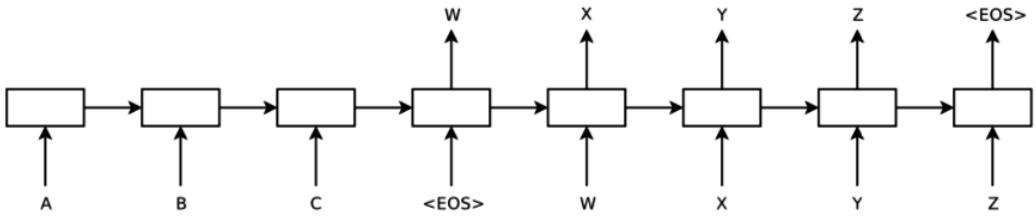


Figure 2.6: The straightforward Sutskever NMT model. The encoder reads the inputs until the end-of-sentence (<EOS>) token and creates a large fixed-length vector representation. The decoder extracts the outputs from the vector *W* til it reads <EOS> (Sutskever et al., 2014).

the model to better learn semantic (and contextual) information. In fact, Sutskever et al. (2014) claimed that their qualitative analysis revealed that the model dealt well with word order and grammatical components such as active/ passive voice. In addition to the baseline encoder-decoder framework, a special feature of the Sutskever's model is the use of LSTM (instead of RNNs) that is known to be more robust in processing long-term dependencies and the introduction of a source language in a reversed way (*a, b, c -> c, b, a*) that supplies the model with more short-term dependencies. It was known to be the first model that outperformed the state-of-the-art SMT model by 1.5 BLEU point.

2.2.1.2. Cho's encoder-decoder model

Prior to Sutskever et al. (2014), the concept of the encoder-decoder model was initially proposed by Cho et al. (2014a)[19] where an encoder and decoder were run by two RNNs and mapped input sequences into a fixed-length vector representation as in Figure 2.7. The range of application of the study at that moment, however, was limited to a subpart of an SMT baseline as a rescoring method of phrase pairs and was not considered as a full-fledged NMT model. They further refined the model and proposed an NMT model based on the encoder-decoder platform in Cho et al. (2014b). The basic workflow of the model was similar to that of Sutskever's model, but the difference was that the encoder was based on gated recursive convolutional neural networks, and the decoder was based on vanilla RNNs with gated hidden units.

[19] The study of Sutskever et al. (2014) was published in December while that of Cho et al. (2014a) was published in October.

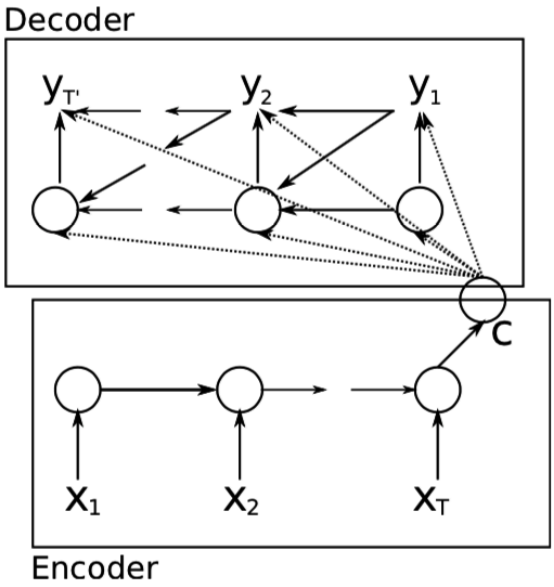


Figure 2.7: Architecture of the encoder-decoder model by Cho et al. (2014a).

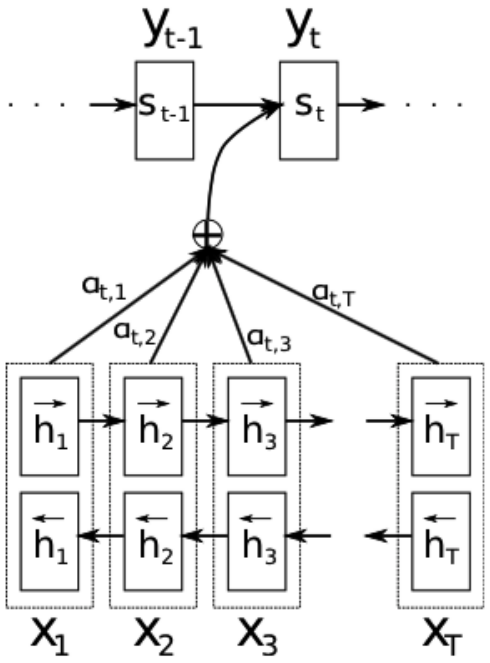


Figure 2.8: The principle of the decoder of the encoder-decoder model with the attention mechanism proposed by Bahdanau et al. (2014).

2.2.1.3. Attention Mechanism

The biggest challenge of the two encoder-decoder models was that the performance deteriorated drastically in the longer sentences. Cho et al. (2014b) claimed that the BLEU score decreased sharply after the sentence length reached 10 ~ 20 words. To mediate such issue of dealing with a long-term dependency, Bahdanau et al. (2014) applied the so-called attention mechanism (Graves, 2013) to the baseline model and achieved a breakthrough in MT with a comparable or slightly better performance against SMT engines. The core idea of attention mechanism was to relieve the burden of input vectors from having to contain a vast amount of context information by means of the attention mechanism in a decoder that could put special attention on the relevant information adaptively when reading the vectors. Unlike the vanilla RNNs, a hidden state in their model was conditioned on the previous hidden state (S_i) along with the previous output (Y_{i-1}) and a context vector (C_i) as in:

$$S_i = f(S_{i-1}, Y_{i-1}, C_i). \tag{2.3}$$

The context vector was where the attention mechanism displayed its ability. As the graphical illustration of the model in Figure 2.8 showed, it was composed of a) *annotations* (h_T) of the preceding and following words bi-directionally learned, multiplied by b) *weights* (a_{ij}) that represented the probability of the j-th word aligned to the word at the i-th position. In that sense, the weights served as an alignment model, as well as a translation model, that were jointly learned with feedforward neural nets. Through the two features — annotations and alignment information—, the context vectors were capable of highlighting a certain sequence vector and extract the relevant information in the decoding step. The key benefit was that the model captured broader context information, especially when translating very long sentences (Bahdanau et al., 2014).

As such, more NMT models have been investigated with different algorithms and techniques but all based on the aforementioned baselines. Better NMT models can process a bigger volume of the dataset in a more efficient way and can capture more context information, in such a way that the quality of translation is improved.

2.2.2. Advantages & Disadvantages of NMT

> "… The NMT process is less transparent than previous paradigms. Indeed, it represents a further step in the evolution from rule-based approaches that explicitly manipulate knowledge, to the statistical/ data-driven framework, still comprehensible in its inner workings, to a sub-symbolic framework in which the translation process is totally opaque to the analysis." (Bentivogli et al., 2016)

Despite the robust performance of NMT, many disadvantages of the model have been claimed. Bentivogli et al. (2016) who performed an exhaustive evaluation experiment of SMT versus NMT pointed out that the primary disadvantage of NMT was its opaque architecture that did not give information about the process

of the system so that detecting errors could not contribute much to the improvement of the system. B.Kang and J.Lee (2018) stated that the NMT required high computational cost in the training step and that it was data-hungry, meaning that it took days to train the model and the maintenance cost of the system was usually 20 times higher than that of SMT. And a competitive performance could not be guaranteed with an insufficient number of data, i.e. less than a few millions of words (Koehn and Knowles, 2017). Besides such individual aspects, some studies claimed that NMT was not ready to be deployed from a general perspective because the performance of NMT was worse than that of SMT yet. Castilho et al. (2017a) stated that although the automatic evaluation scores saw a promising aspect of NMT, a series of human evaluations confirmed that SMT surpassed NMT. Farajian et al. (2017) also claimed the outperformance of SMT over NMT through a series of comparative evaluation test of NMT and SMT in multiple domains. Skadiņa and Pinnis (2017) supported the claim by evaluating them in a narrow domain. Toral and Way (2018), despite showing the outperformance of NMT, argued that human parity was still a long way to go.

Those negative opinions, albeit partially true, leave much room for discussion. It is a known fact that the performance of MT depends on the domain type and size of the dataset. The fact that a model performs more poorly than the other model when trained and tested in domains of a different nature does not necessarily justify the robustness of one model over another. It just manifests which model is weaker in the given setup. In that sense, the experiment of Farajian et al. (2017) shows that NMT is more vulnerable to a domain variation. Furthermore, Skadiņa and Pinnis (2017) acknowledged that the negative result for NMT could be due in part to the lack of available dataset of their study.

In fact, NMT exhibits many positive sides. The way input sequences are represented into some vectors allows the model to generalize the data and capture context information with excellent proficiency (Toral and Sánchez-Cartagena, 2017). It also leads to a great accuracy of the output, more flexibility in terms of dealing with syntactic features and remarkable freedom to the developers not having to tune every parameters (B.Kang and J.Lee, 2018), not

to mention that the utility of available data has been enhanced in the neural (Bentivogli et al. 2016). Moreover, the system uses a small fraction of memory (about 500MB) in contrast to the amount SMT requires (tens of gigabytes) (Cho et al., 2014b).

In the middle of such mixture of positivism and negativism, there are some common grounds that are witnessed in the majority of the studies that compared the performance of SMT and NMT, which are given below point by point. Each point will be discussed with the center of attention on the language combination.

- Great accuracy with the help of strong reordering skill
- Poor performance in translating very long sentences
- Major errors stemming from omission and mistranslation

High Accuracy In the first place, what every research in this regard coincided in was that NMT produced a more accurate translation, in a sense that the output translation was linguistically flawless. The claim was made in multiple language pairs such as English to or to/from German (Bentivogli et al., 2016; Popović, 2017; Castilho et al., 2017b; Bentivogli et al., 2018), Czech, Romanian, Russian, Finnish, Turkish (Toral and Sánchez-Cartagena, 2017), Serbian (Popović, 2018a), Greek, Portuguese (Castilho et al., 2017b), French (Bentivogli et al., 2018), Croatian (Klubička et al., 2017), Catalan (Toral et al., 2018), Korean (S.Kim and H.Lee, 2017) and so on. The great achievement of NMT in accuracy stemmed, in part, from its capability to capture syntactic features such as word reordering. Bentivogli et al. (2016) emphasized that the key benefit of the NMT engine that was under evaluation in their study was the strong reordering ability, reducing in general 50% of the errors produced by the SMT engine. Special attention was paid on the movement of verbs (by being reduced up to 70%) which was the main error that other three state-of-the-art PBMT models committed. The study concluded that the engine had an impressive ability to build well-formed sentences in distant pairs such as *en-de*. Such finding was reconfirmed in Bentivogli et al. (2018) and Popović (2017).

Poor performance in long sentences Many studies coincided in that incapability in longer sentences was as much challenging to NMT as to SMT, but many of the studies found that at some point, the quality of NMT deteriorated more sharply than SMT. The break-even point differed per studies. For instance, in the case of Bentivogli et al. (2016), the quality of the output produced by NMT started to be degraded from sentences longer than 35 words, and the quality was lower than that of SMT. A similar observation was claimed in Toral and Sánchez-Cartagena (2017) where NMT performed poorer than SMT beginning from sentences with 36 ~ 40 words. Koehn and Knowles (2017) claimed slightly in a more positive way, saying that the performance of NMT was better than SMT in sentences with 60 words, but when they reached to 80 words, the quality fell to an extremely low level (a judgment based on BLEU scores).

Major error in omission & mistranslation Lastly, most of the errors were detected in the category of omission and mistranslation. Many of the studies found a reason for not dealing well with rare or unseen words which were also called out-of-vocabulary (OOV) words. While many of them pointed out that it was problematic for SMT and NMT equally (Castilho et al., 2017b; Koehn and Knowles, 2017; Klubička et al., 2017), Popović (2018a) and Bentivogli et al. (2018) performed error analysis by means of POS tagging and found that the most common POS that was omitted or mistranslated was proper nouns.

All these interesting points made by previous studies pose questions over the case of the *es-ko* language combination. How will the word reordering be managed? What will be the relation between translation quality and sentence length? How will NMT perform as sentences get longer? What types of errors will be most frequently detected? Will there be many omissions and mistranslations? These are the questions that this thesis strives to answer.

2.2.3. Achievements of NMT over SMT in a Korean translation

As seen in the previous chapter, more and more studies are claiming that NMT outperforms SMT either slightly or markedly in various language combinations

and domains based on both automatic and human evaluation methods. Most of such comparative studies are overviewed by Popescu-Belis (2019), but Korean-related studies are not introduced in this regard. As such, this chapter introduces studies that compared the performance SMT and NMT in terms of a Korean translation to give insights to the Korean-involved language pairs. Albeit few, these evaluation studies also shed light on the robustness of NMT in the Korean translation.

A.Chang (2017) carried out descriptive and comparative analysis of seven freely available NMT engines and one SMT engine that could translate Korean and Chinese in both directions —two versions (SMT and NMT) of Papago, Genie Talk (interpretation application) of Hancom Interfree[20], PNS (Pure Neural Server) and exTalky (interpretation application) of Systran International, Google Translate and Baidu[21]. Five sentences in the Chinese-to-Korean translation and another five in the Korean-to-Chinese translation, both of which were selected from a variety of domain, were analyzed one by one based on the author's intuitive judgment. The study asserted that NMT (of Papago) achieved outstanding performance over the proposed SMT engine in the *zh-ko* translation in both directions. From qualitative analysis, they claimed that NMT tended to output more accurate translations than SMT. However, an accurate translation was only observed in domains that required a literal translation. In a humorous text, for example, the system exhibited the worst performance. They reasoned out that the sentences in such domain required the system to have cultural background information and a pragmatic linguistic approach. Moreover, similar to most of the comparative studies, she claimed that the translation quality drastically deteriorated in very long sentences, but did not specify the number of words. Despite the valuable insights from her research, they fall short of credibility due to the fact that they are based on only 10 sentences.

In a similar manner, S.Kim and H.Lee (2017) compared the performance

[20] http://www.interfree.com/if/main/main.do

[21] https://fanyi.baidu.com

of unknown NMT and SMT systems (assumedly Papago[22]) in embedded sentences obtained from an English-to-Korean translation in a movie script, from a descriptive and phenomenological standpoint. The data summed up to 179 embedded clauses composed of simple and complex sentences. The study stated that NMT had less syntactic errors than SMT while they tended to produce an out-of-context translation that had nothing to do with the original text, suggesting evidence that the engine sacrificed the fluency to obtain higher accuracy.

2.3. Google NMT

Google LLC, as a front-runner of NMT, has proposed various groundbreaking models, starting from Sutskever et al. (2014) who initially proposed the end-to-end NMT architecture and the sequence-to-sequence model, opening the door to the NMT world. The successful evolution of many NMT models led to the development of Google's own NMT model in 2016 (Wu et al., 2016) and its official deployment on their open MT service platform, Google Translate, in November of the same year, upgrading the quality of a freely available machine translator remarkably. The detailed architecture of Google Translate at this moment is not officially described, but the influence of Google research teams in NMT is exceptional.

While developing up-to-date NMT models, Google is also actively engaged in the field of Multilingual NMT, which aims to translate multiple languages in one single model, by carving out a new field of namely zero-shot translation (Johnson et al., 2016). Their multilingual model contributed to boosting the power of the engine, handling from nine languages in 2016 (English and eight non-English) to 103 (including English) in 2019[23] in Google Translate, enabling quality translations of 10,506 language directions.

This chapter tries to take on research on the architecture of Google' NMT models as Google Translate is under evaluation in this thesis. Chapter 2.3.1

---

[22] The paper did not clearly mention with which system they performed an experiment, but considering that they found a reason of NMT producing out-of-context translation from materials of Naver Lab's, it is assumed that the engine in question is Papago of Naver Inc.

[23] The date refers to the moment of judgment saying that currently 103 languages are translated in Google Translate. It does not mean that Google started to cover those languages in 2019.

addresses the development of Google's NMT models, starting from the initial version. As the current architecture of Google Translate is unknown, this chapter introduces the first NMT model that is claimed to be applied in Google Translate and other newer models that are studied by Google research teams. Chapter 2.3.2 shifts the center of attention to the multilingual NMT field that is proved to be especially effective in handling low-resource language pairs like the *es-ko*. Starting from a general overview of the atmosphere of this line of research, such as Pivot NMT, this chapter reviews the contribution of Google in this regard.

2.3.1. Evolution of Google's NMT Models

In this section, two Google's NMT models are introduced that have been developed and claimed a new state of the art: namely Google NMT (GNMT) and Transformers. Subsequently, some advanced techniques based on the vanilla Transformers are described as the most up-to-date system. It is important to stress again the fact that the purpose of the description in this section is to show the improvement of Google Translate and the contribution of Google to the NMT field. As such, the computations and algorithms are briefly summarized and the exhaustive information of the model architecture is left out.

2.3.1.1. GNMT by Wu et al. (2016)

The first Google's NMT model revealed its appearance in public in Wu et al. (2016). The core architecture of the model was based on the conventional sequence-to-sequence encoder-decoder model of Sutskever et al. (2014) with attention mechanism (Bahdanau et al., 2014). As shown in Figure 2.9, the model consisted of two LSTM RNNs in the encoder and decoder side and attention module in the middle. Each side had eight layers in an independent GPUs to boost the computational speed. The key features that differentiated it from other models laid in the following three aspects:

- The low computational cost for training and inference time
- Better dealing with OOVs
- Broader input coverage

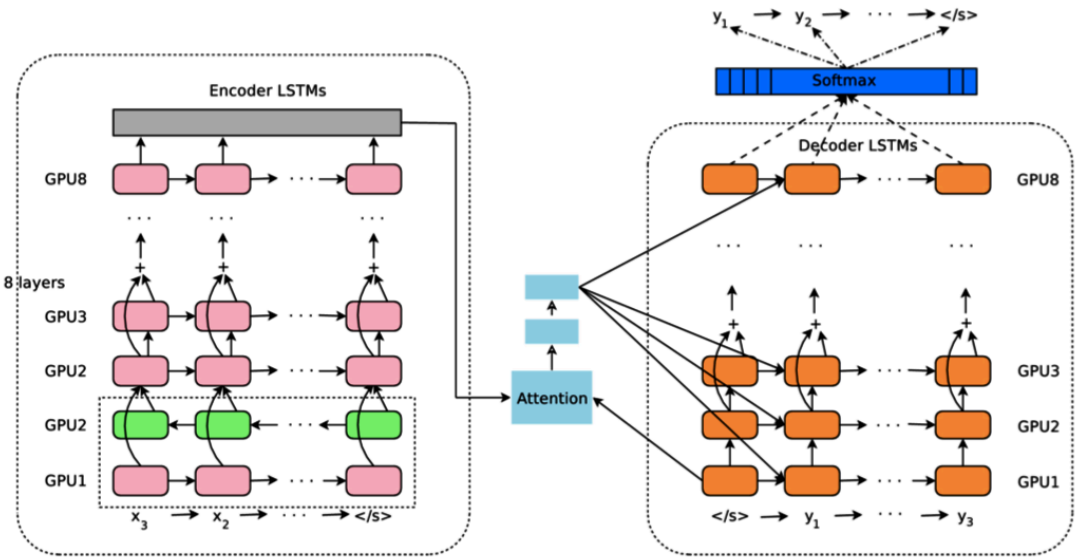


Figure 2.9: Architecture of GNMT (Wu et al., 2016).

To speed up the process of training and inference task, the parallelism was maximized by i) maintaining the eight deep layers with the help of residual connections in the stacked LSTMs from the third layer to the eighth that update the previous input information and so, facilitate gradient flow; ii) running on bi-directional RNNs; iii) connecting attention module from the top to bottom decoder in a straightforward manner; iv) introducing special hardware — Google's Tensor Processing Unit (TPU)— to enable quantized arithmetic for faster inference. Secondly, to better capture the rare words or out-of-vocabulary (OOV) words, the model made use of sub-word units that broke the ambiguous word into pieces to later recover the original word with a special mark in front of every word. Lastly, to broaden the input coverage, the decoder was based on beam search that maximized the probability of the target sequence, with the addition of length normalization and coverage penalty algorithm such that the judgment of the decoder was not biased against different sequence length and fully cover the whole sentence. The human evaluation of the given model reported on a 60% error reduction compared to Google's SMT model.

2.3.1.2. Transformers by Vaswani et al. (2017)

In the subsequent year, a new model called *Transformers* was developed by a group of Google researchers (Vaswani et al., 2017), claiming a new state of the art in MT. The core difference of this model in contrast to the previous sequence

models, or so-called transduction models, that were typically based on RNNs and convolutional neural networks was that Transformer was entirely based on attention mechanism that constituted one part of the previous models. The key benefits of this model was that:

- It facilitated a parallelized computation, which was more suitable for modern GPUs and boosted the process by reducing training time to 1/4 (3.5 days).
- The input and output were learned regardless of their variable length.
- The current model outperformed former state-of-the-art models in the *en-de* and *en-fr* pairs by BLEU scores.

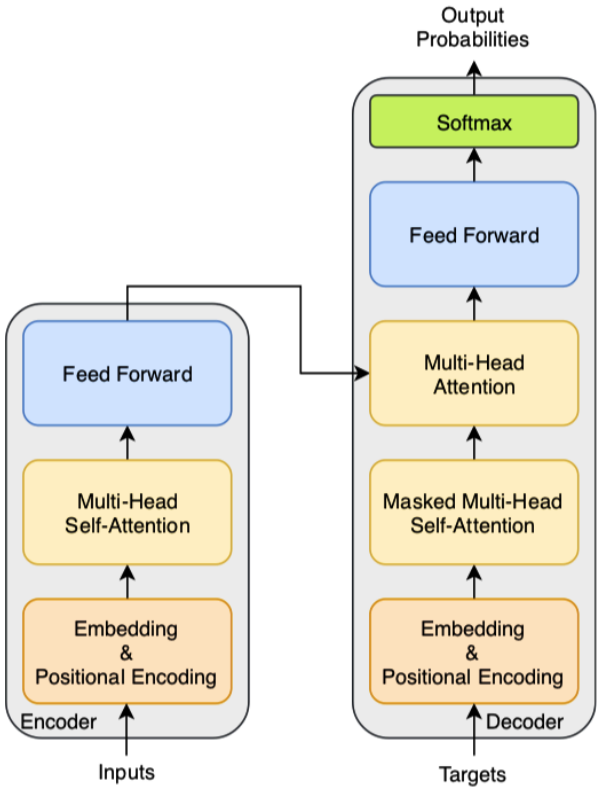


Figure 2.10: Simplified version of the architecture of Transformer by Tubay and Costa-jussà (2018).

The simplified architecture of the model designed by Tubay and Costa-jussà (2018) is given in Figure 2.10. Compared to Wu et al. (2016), the noticeable difference is that the attention mechanism (self-attention) is multiple and is included in both the encoder and decoder. The role of *multi-head self-attention* is to "let vectors learn all sequences in all different positions" in a more concrete but refined way. The attention mechanism in the encoder lets each

sequence contain all information of the rest sequences in the encoder, which is added later to the attention mechanism of the decoder in the third layer. It is mixed with the vectors of the other attention mechanism in the second layer of the decoder that contains information in the decoder so that in the final stage, the model gets vectors with refined and concrete information of both the encoder and decoder. In such a way, this structure lets the model learn information about each sequence in any positions in the data. After that, the different dimension on the vectors is computed by Feedforward neural nets.

This model alleviated one of the biggest issues that the previous models, i.e. Convolutional Sequence to Sequence Model, had is a long-range dependency. That is, “the number of steps required to combine information from distant parts of the input grows with increasing distance (Uszkoreit 2017)”. For instance in the example given in Figure 2.11, normally the farther the relevant information is (to “it”), the less accurate the output gets in the previous models. On the other hand, the current model considers all relations of all segments irrespective of their position and computes the level of the bond as expressed in Figure 2.11 with a depth of blue. As such, in the first sentence, the model learns that “it” is more related to “animal” in the context while in the second sentence, it is more relevant to “street”.

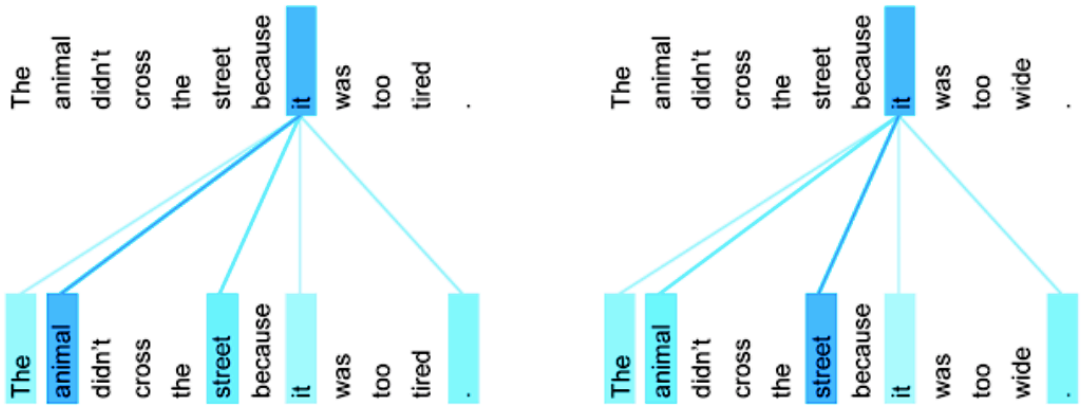


Figure 2.11: Distribution of level of bond in an encoder trained on the en-fr translation (Uszkoreit, 2017).

2.3.1.2. Further Improvements from Vanilla Transformers

The initial model called vanilla Transformers is gaining popularity in many natural language processing tasks, especially in language modeling. One of the innovative models is **BERT** (Bidirectional Encoder Representations from Transformers) developed by Devlin et al. (2018) that learns pre-trained representations from a vast amount of general-purpose data in a deeply bidirectional manner and fine-tune the model afterward with the small task-specific dataset. The core benefit of this model is that it alleviates the shortage of annotated training data of NMT by enabling the usage of general-purpose, unannotated raw data (Devlin and Chang, 2018). The feasibility of this model has been confirmed in other NLP tasks as well such as sentence classification and question answering, achieving state-of-the-art on 11 NLP tasks (Devlin et al., 2018).

Dai et al. (2019) strived to mediate one of the biggest challenges of sequential NMT model in general and of Transformer in particular in dealing with the long-term dependency. Although Transformer alleviated this issue, the fact that it was based on fixed-length vectors limited the model from controlling sequences that are farther than the fixed-length properly (Yang and Le, 2019). The new language model called **Transformer-XL** (meaning extra long) improved the Transformer model by involving the recurrence mechanism to the architecture. During training, “the hidden state sequence computed for the previous segment was fixed and cached to be reused as an extended context when the model processed the next new segment”, in such a way that the network learns the context in the history (Dai et al., 2019). This model claimed to learn 80% longer dependencies than the state-of-the-art RNNs and 450% longer than the vanilla Transformers (Yang and Le, 2019).

More robust models are being claimed day by day with the blink of an eye, and at this right moment of writing the current thesis, newer models are refined and developed. The detailed architecture of the BERT and Transformer-XL is referred to each study that also introduces the application of the vanilla Transformers to other NLP tasks.

2.3.2. Multilingual NMT

In spite of the sturdy performance of varicolored NMT models, their ability to handle some language combinations that fall short of parallel data is significantly limited. One of the encouraging signs in translating low-resource language pairs comes from a multilingual model that translate one-to-many, many-to-one or many-to-many languages in a single system. The idea of translating multiple languages at once is not a new approach in MT field of study, but it was more or less a novel direction in NMT back in 2016 when Firat et al. (2016a) proposed a multi-way multilingual NMT model that could handle many-to-many translation without the cost of efficiency and accuracy. Founded on an assumption that the scenario of introducing multiple sequences would be much more feasible to the nature of RNNs, the multilingual model was successfully devised with a baseline of the basic encoder-decoder NMT model with one shared attention mechanism (Bahdanau et al., 2014). The experiment in English to/from French, Czech, German, Russian, and Finnish reported that the BLEU scores of all language pairs were enhanced in the into-English translation, albeit slightly, but was inferior to the single model in the from-English translation. It was fascinating attainment considering the amount of computational cost it cut down. Were it not for this technique, it would have required many one-to-one models, with parameters growing quadratically (Firat et al., 2016a). Not to mention the great achievement in a large-scale translation, the study shed light on the betterment of handling low-resource language pairs. The study found out that the multilingual model learned *interlingua* from various languages, in such a way that the lack of parallel corpus of resource-scarce language pairs could be compromised by the generalized linguistic data. The performance of the given model surpassed the single model with and without the addition of monolingual data. The feasibility of the multilingual NMT architecture to low-resource language pairs has been proved afterward in many other studies as well.

2.3.2.1. Pivot-Based Approach

The major improvement of multilingual NMT was achieved with a pivot-based

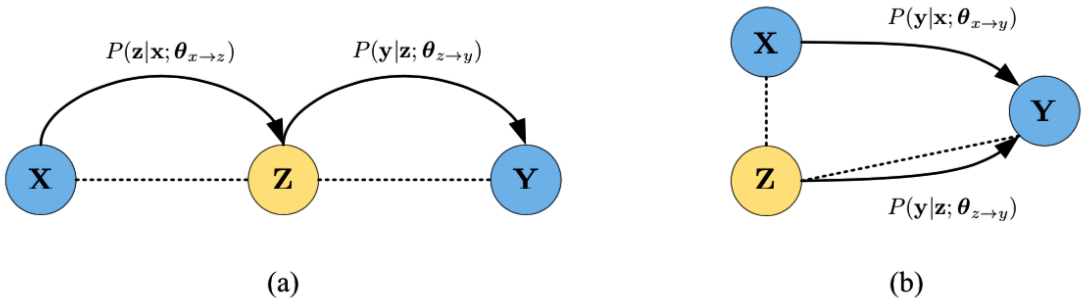


Figure 2.12: The different architecture of (a) the conventional pivot-based and (b) zero-resource approach (Chen et al., 2017). The straight lines represent a translation direction and the dashed lines denote the usage of the parallel corpus.

approach. Among many strategies of pivoting, in general,[24] the most popular approach is cascading two translation systems where the source text is translated to a pivot language (usually English because of the abundant available data) and the pivot-translated texts are translated into a target language. This two-step explicit transference yielded promising outcomes especially for resource-poor languages (Cheng et al., 2016), and recently Cheng et al., (2017) claimed that this approach was more effective in NMT than SMT.

2.3.2.2. Zero-Resource Approach

The primary bottleneck of the pivot NMT, however, is the high computational cost it involves in order to perform two rounds of translation (source-to-pivot, pivot-to-target) and possible data loss during the cascade, as well as an error propagation problem, meaning the errors from the source-to-pivot translation severely damage the pivot-to-target translation (Cheng et al., 2017). As a way of enhancing the accuracy of the multilingual model with lower cost, the *zero-resource* approach was newly proposed (Firat et al., 2016b; Chen et al., 2017). This new approach simplified the process of the pivoting into one-step decoding by directly translate X to Y language with the guidance of a pseudo-parallel corpus of X-to-Z (the pivot language) and Z-to-Y serving as a fine-tuning algorithm (see Figure 2.12). Despite the great performance achieved over the

[24] Those who are interested in pivoting strategy are directed to Paul et al. (2013).

conventional pivot-based, the computational cost was still too high to build multiple language models and the parameters were multiplied quadramatically in the number of languages.

2.3.2.3. Zero-Shot Approach by Google

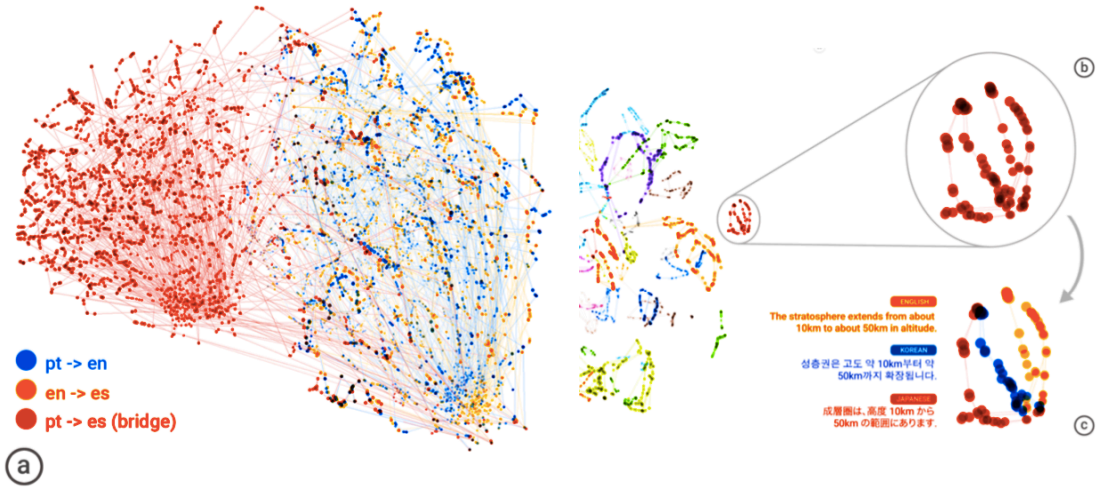


Figure 2.13: A t-SNE projection of Johnson et al. (2016). (A) shows a zero-shot translation of pt-es with the help of English. (B) shows vector representations of English, Korean, and Japanese.

A promising line of research was put forward in this regard, namely *zero-shot* approach by researchers at Google (Johnson et al., 2016). The principal difference between the pivot-based and zero-shot model was the way a language pair was bridged; in an explicit way in the pivot-based and in an implicit way in the zero-shot approach. The zero-shot approach made use of the trained data from the existing multilingual model (i.e. *a->b* and *b->c*) and translated between a language combination that had no previous explicit contact (or alignment) in the training set (*a->c*). It was ultimately a straightforward model that translated two resource-poor languages without explicit parallel corpus that did not require additional language modeling like pivoting or the fine-tuning step like zero-resource.

The key benefit of the zero-shot model is simplicity, feasibility, and cost-effectiveness. It is simple because it does not require any changes in the model but the addition of token indicating its target language. It is feasible because the experiment outcome reported that there was a minimal decrease in BLEU score

compared to pivot-based and a single NMT model, obtaining an affordable quality (above 20). It is cost-effective because it reduces the size of parameters by five times and translation time, by 12 times. Moreover, it reconfirmed the fact that multilingual model had strong power of generalization by learning interlingua as shown in many multilingual models. Figure 2.13b and 2.13c show a t-SNE projection of 74 semantically similar sentences of six language combinations of English, Korean, and Japanese from many-to-many translation model. The clustering of the representations in different languages gives evidence of the existence of interlingua between them.

The challenge of the model, however, was that the generalization was incomplete. As shown in Figure 2.13a, out of all the vector representations, the semantically identical projections occupy only one-third of the space, meaning that the rest of the representations have inadequate target language representations (either of English or Portuguese, in this example). To remedy the issue and enhance the power of generalization, the Google AI team refined the model by regarding it as a domain adaptation problem (Arivazhagan et al. 2019), the idea being that instead of sharing parameters among all languages, non-English languages were represented as a different form of English by minimizing the discrepancy between them and English sequences increased the generalization. The experiment of the *de-fr* translation with English based on Transformers (Vaswani et al., 2017) reported on a striking improvement of the previous zero-shot model and outperformed the pivot-based model by a good margin.

2.4. Chapter Summary

Chapter 2 embarked on a theoretic study of NMT, demonstrating the importance of NMT in the current MT evaluation field. With the main focus on the different origin of NMT from the conventional MT systems, we denominated the turn of SMT to NMT as a *technical shift*.

We started by presenting the concept of the Machine Learning algorithm, the Artificial Neural Networks (ANNs). The most basic unit of ANNs was an artificial neuron whose concept was borrowed from a human brain cell. From the action potentials of human neurons, the connectionists devised Perceptron,

the initial artificial neural network model with a single layer. It was refined into Feedforward Neural Networks that could handle multiple layers with Backpropagation and Gradient Descent. We, then, introduced two most basic models that could learn sequential data: Recurrent Neural Networks and its developed model, Long Short Term Memory. We, then, defined NMT by explaining the level of application of ANNs to MT engines. We introduced a sequence-to-sequence model of Sutskever et al. (2014) as one of the first purely-neural-based models.

In Chapter 2.2, we described mainstream NMT models starting from the encoder-decoder model by Sutskever et al. (2014) and Cho et al. (2014) and their improved version with the help of Attention Mechanism. We discussed the advantages and disadvantages of such NMT models and put special attention on the three core features: i) high accuracy and reordering skill, ii) exceptionally poor performance in long sentences, and iii) major errors in omission and mistranslation. We closed the sub-chapter by investigating previous studies that claimed the outperformance of NMT over SMT in a Korean translation.

In Chapter 2.3, we narrowed down our focus on the architecture of NMT engines of Google Inc. Firstly, we described the first official NMT model of Google that was applied in Google Translate. And then, we introduced currently the state-of-the-art engine, Transformers, and its improved versions such as BERT and Transformer-XL. Secondly, we approached another line of research that aimed at low-resource language pairs such as the *es-ko* pair: Multilingual NMT. After introducing some mainstream models such as Pivot-Based Approach and Zero-Resource Approach, we presented Zero-Shot Approach devised by a Google research team.

3. NMT EVALUATION & POST-EDITING

Probably one of the most influential MT evaluations in history would be the ALPAC report in 1966, as it caused a decade of a dark age in the MT field. Despite the MT evaluation being a sub-field of MT, its influence on the development of MT has always been significant throughout history (Babych, 2014). Before the emergence of SMT, the MT evaluation was largely dependent on human evaluation methods based on criteria of comprehensibility, fluency, and fidelity (Hutchins, 2015). It was since 2000 that the MT evaluation started to be performed in an automatic way, beginning with BLEU of Papineni et al. (2002). The automatic methods made it possible to evaluate MT quickly and immediately, in such a way that the systems were optimized instantly (Babych, 2014). As such, MT evaluation is an indispensable part of MT. This chapter is dedicated to investigating the MT evaluation from its objective and methodologies to its application to the Korean language. While delving into the relevant theories of the MT evaluation, this chapter also features the role of post-editing as an evaluation method.

3.1. Objective of MT Evaluation

The role of MT evaluation is “to determine the effectiveness of existing MT systems and to optimize the performance of MT systems” (Dorr, 2009). The principal criteria for assessing the effectiveness of the engine are speed and quality. First of all, the speed would be the first and foremost reason for the existence of a machine in contrast to human, in that it cuts off the total translation cost and human efforts. The quick obtention of a rough idea of a foreign text in a second has saved considerable time and money. *Assimilation* is termed in that regard to describe a way of using MT to produce a draft translation in contrast to publishable quality, which is called *dissemination* (Hutchins, 2015). Although the speed is crucial, it can be, as a matter of fact, simply improved by a better version of hardware and software (Arnold et al., 1994). If the computing power is upgraded to handle more tasks at a given time than before, for example, the result will be obtained a lot faster. It, then, leaves us with the second criteria of the engine performance —the quality.

The quality of MT is a major criterion of the MT evaluation whose goal is "to meet a certain degree of translation quality and deliver the translation as a product with that quality maintained" (Doherty, 2016). Doherty (2016) adds an additional explanation as to the definition of an ideal MT evaluation to be objective at the same time the turnaround time is rapid and subsequently cost-effective, not to mention that the results are comparable to any language pairs and any MT systems. Nevertheless, the first and foremost objective in MT evaluation is to be able to provide a conclusive remark on the level of "goodness" of an output. Unlike the precise goal it has, ways to reach the goal are far from being precise due to the absence of an absolute standard of the quality. The main reason is that translation itself, by definition, is a mixture of "cognitive, social, cross-linguistic, and cross-cultural" acts of a human being (House 2015). That is, there is no ground truth in terms of the standard of translation quality. It is, therefore, crucial to clarify the definition of a correct translation or equivalence in MT evaluation.

The definition of equivalence in translation studies differs in time, but it is more or less standardized in the MT evaluation field. There are two agents that judge the quality of MT —human and machine—, and each agent decides such definition accordingly. From the perspective of human evaluation, an equivalent translation depends on the instinctive judgment of an annotator (Sanders et al., 2009). From the perspective of a machine evaluation, an equivalent translation is the most approximate version to a translation prepared by a human translator. In that sense, the concept of so-called *Gold Standard* is that the closer a hypothetical translation is to a human translation, the better. All in all, a typical way of performing the MT evaluation is either to judge directly the quality by the hand of human or to indirectly calculate the degree of similarity of the translation to human-produced outputs automatically. Both of them, therefore, imply that the concept of equivalence in MT evaluation is a human parity.

3.2. MT Evaluation Typology

Various MT evaluation methods delved into measuring the level of a human parity of translation output produced by a system. While a concerted effort has been made in that regard to evaluate the conventional MT engines such as

SMT and RBMT, a tailor-made evaluation method for NMT has not yet been discussed. The evaluation of NMT engines is still largely dependent on the conventional automatic methods that have been used in SMT. In fact, Jean et al. (2015) are known to be the first to initiate a series of manual evaluations on NMT engines (Way, 2018). Since the emergence of NMT on the surface, it is important to have a method that is designed to take its different architecture into account. However, as no such method is devised to date, this section inquires into the theories of the conventional MT evaluation starting by presenting a typology of existing MT evaluation measurements.

3.2.1. Traditional Paradigm

Traditionally, a binary distinction of the MT evaluation methods has existed in a broad sense: glass box and black box. **Glass Box** evaluation assesses the performance in consideration of a property of a system. It is usually applied to an RBMT system to verify its linguistic features and theories (Dorr, 2009). On the contrary, **Black Box** evaluation is concerned with the general performance of a system in any given circumstances, the assessment depending solely on the outputs (Dorr, 2009). Such an objective inspection allows comparative evaluation of various MT systems irrespective of their architecture. That is why it has widely used in the MT field of research. Black Box evaluation can measure general quality of a system in an **intrinsic** manner by assessing a fluency and adequacy score of the outputs, which is called *metric-based* evaluation. It can assess task-specific quality in an **extrinsic** manner by monitoring the usefulness of a system in a certain task, which is known to be *task-based* evaluation (Doyon et al., 1999). In the context of human evaluation in specific, the intrinsic and extrinsic evaluation are also denominated as *declarative* evaluation, which inquires into how good a system is, and *operation* evaluation, which measures how effective a system is (Humphreys, 1991).

3.2.2. Human/Machine Paradigm

The most fundamental and probably the most common classification of the MT evaluation is a division between **human** and **machine**, two agents that can perform the task. The human or manual evaluation is based on the subjective

judgment of one or more annotators. The machine or automatic evaluation, on the other hand, focuses on detecting certain linguistic features or similarities of translation on a lexical or syntactic level by applying algorithms (Sanders et al., 2009). The key benefit of the human assessment over the automatic one is a comprehensiveness that approaches from a holistic perspective. It is, however, challenged greatly by the subjectiveness and inconsistency that are inherent consequences as human behavior. Moreover, it is confronted with limitations such as labor-intensiveness, inefficiency, and high cost. The automatic evaluation has gained attention in this regard because it alleviates such weaknesses, although it is originally designed to “monitor the development of the same MT engine” (Babych, 2014). Unlike the manual evaluation method, it is fast, robust, time- and effort- saving and consistent. Despite such merits, the automatic evaluation methods also hold the issue of subjectivity with them, albeit less than the manual method possesses, in that equivalence is judged entirely based on a translation produced by humans.

Not only can the score vary by the quality of the reference translation, but also the judgment is decided on a few versions of translation. That is why HTER metrics have gained much attention because it remedies such limitation by involving human in the loop of the automatic evaluation. The core idea is identical in that it measures the similarity between a system translation its reference translation. The difference is that it requires human to produce multiple *human-targeted reference translations* that are the closest version of reference translation to the given system translation via post-editing (Snover et al., 2006). These types of MT evaluation metrics are called semi-automatic evaluation.

3.2.3. Technique-based Paradigm

Han et al. (2016), in their review of MT evaluation methods, draws a further distinction from the human and machine paradigm to cover up-to-date information based on technical features of the methods. As in Figure 3.1, the human and automatic evaluation methods are grouped into *traditional* and *advanced* categories, respectively. In human evaluation, the traditional methods include the initial models such as intelligibility and fidelity scorings and their

developed versions —fluency and adequacy scorings, all of which approach the task in a direct way. The advanced category reflects a more indirect way of evaluation: segment ranking and extended criteria (from the traditional group) such as suitability, reliability, maintainability, etc. of King et al. (2003), as well as task-based evaluation and post-editing-related methods that measure the performance of an MT system in relation to its effectiveness in post-editing. Meanwhile, the automatic evaluation methods feature more traditional methods than the advanced ones. The traditional methods are sub-divided into their coverage in a lexical, syntactic and semantic realm. The most commonly used automatic metrics in the field such as BLEU (Papineni et al., 2002), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), TER (Snover et al., 2006), etc. belong to the lexical category. They are computed on a basis of lexical similarity between a system translation and reference translation or an edit distance in the case of TER. The technically-advanced category includes some DeepL-applied models. The basic architecture of some of the automatic measures is provided in Chapter 3.5.3.1. One noticeable fact from such division is that the human methods are holistic and totalistic whereas the automatic methods are partial and purposeful.

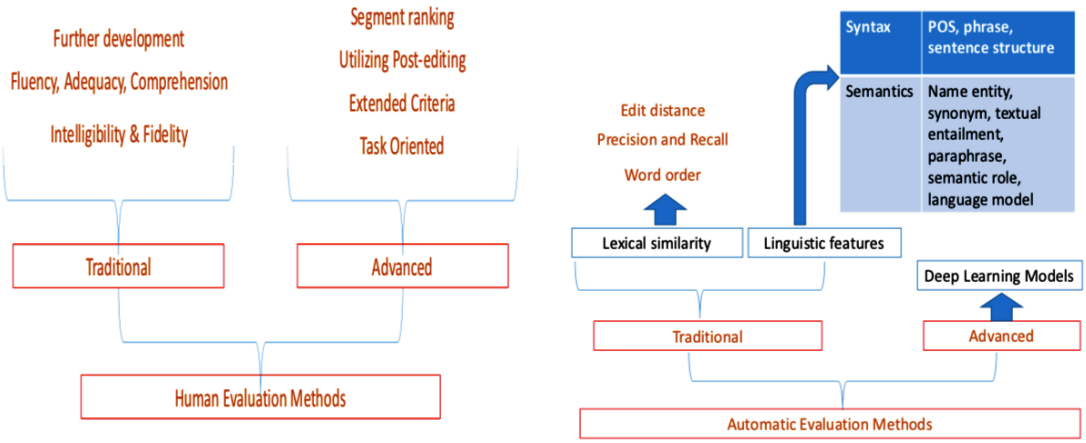


Figure 3.1: MT evaluation typology of Han et al. (2016).

3.3. Inherent Issues of MT Evaluation Methods

The underlying hypothesis of the current notion of equivalence in MT evaluation insinuates that a perfect form of MT is one that can produce a human-styled translation. To this end, various evaluation methodologies have been devised to grade how *human* a system is. In this context, the best evaluation method that can measure a human parity is the human itself who would either directly assess the quality of a system or indirectly observe its influence on certain tasks. While the manual methods are comprehensive and best suited for the goal, their greatest limitation, as discussed in advance, is its subjectiveness and inconsistency in the course of action. That is, the result can vary depending on many qualities of the evaluators because each one would make a different judgment based on their subjective criteria. Moreover, the judgment is easily biased by circumstantial factors, which results in an inconsistent result inter-personally and intra-personally. In that regard, the automatic evaluation is a good alternative, in that as far as the requirements are met, it guarantees the consistency and objectivity of the outcome.

Although its calculation is automatized, the automatic metrics also have the subjectivity issue in nature due to the subjectiveness of gold standard. While gold standard is identified with a human translation, there is no agreed-upon consensus as to what level of quality of translation is accepted as gold standard. Such an open selection of a reference translation makes the method vulnerable to biases. Another problem of the automatic metrics is that while the meaning of a source language can be expressed in a diverse way in a target language, the method depends on one version of them, not being able to cover the versatility of desirable examples. A good way to alleviate the problem is to take multiple gold standards under consideration (Niessen et al., 2000) or create a human-targeted reference translation by making use of post-editing (Snover et al., 2006), both of which are a key feature of HTER. Studies find that human-targeted metrics such as HBLEU and HTER are more correlated to human judgment than the conventional automatic metrics in various language pairs (Graham et al., 2016).

Of all the limitations that the automatic evaluation methods feature, their greatest challenge is a poor correlation to human judgment, which has long

been issued (Babych, 2014). Among various reasons that unpin this argument, many claim that the majority of popular automatic methods poorly manage a word order issue because they rely on a lexical level (Birch et al., 2009; Echizen-ya et al., 2009; Isozaki et al., 2010). In relation to Korean in particular, H.Chung (2018) performed automatic evaluation of MT for a *de-ko* translation in an attempt to explore the possibility of BLEU being a substitution of human methods for the given language pair. The result showed that the correlation between the two metrics was too marginal to replace one from the other. Albeit few, the previous studies show that there is a good chance that the automatic methods do not represent human judgment properly, and the situation is nothing but worse in terms of distant language pairs, as the difference between the two languages is more concerned with linguistic aspects.

3.4. Human Methodology

All things considered, desirable MT evaluation is the one that judges a human parity of the system translation by employing methods that fit the best to the circumstances i.e., evaluation purpose, engine, domain, language pair and so on. Although the contribution of the automatic evaluation methods to the development of the MT field is noteworthy in general and their role is of utmost importance to the MT evaluation, their applicability should be carefully decided. It is believed that for a linguistically dissimilar pair like the *es-ko* pair, the reliability of automatic methods is dubious, to date. As such, this chapter addresses the most widely-used human evaluation methods —fluency scoring (Chapter 34.1), adequacy scoring (Chapter 3.4.2) and segment ranking (Chapter 3.4.3)— that are applied to the experiment of the current thesis.

3.4.1. Fluency Scoring

Fluency is one of the most popular evaluation criteria of human evaluation. The concept of fluency was first discussed and applied in the famous ALPAC report as “how intelligible and clear a translation is” under the name of *Intelligibility* (ALPAC, 1966). Such a simple concept was further detailed by LDC (2002) who termed a fluency score. The condition of a hypothetic translation to be considered *fluent* is that the sentence is as well-formed as possible in the target

language and natural to understand the meaning of the sentence in the context as if it was written originally in the intended language (Han et al., 2016). The definition of fluency that TAUS advocates is in the same context:

> "It captures to what extent the translation is well-formed grammatically, contains correct spellings, adheres to common use of terms, titles and names, is intuitively acceptable and can be sensibly interpreted by a native speaker." (Görög, 2014)

The process of the evaluation of fluency scoring is to ask a panel of judges to assign an *n*-point Likert scale on each sentence following the description of the scales and to compute an average score. Some use a mean average. The Advanced Research Projects Agency (ARPA) suggests a formula of *fluency* (White et al., 1993):

$$Fluency(P) = \frac{(J-1)}{(S-1)*N} \tag{3.1}$$

where *J* stands for judgment point and *S*, the total pile of scale and *N*, the total number of sentences in the passage. The fluency ratio P is a division of judgment point (*J*) by a total number of scale (*S*) which is normalized by a total number of sentences (*N*). In the meantime, TAUS (2010) declares that a mode average is a more reliable indicator than a mean average "*since the distribution of results may not always be normal*", and so, the calculation is on an aggregated basis.

While there are various ways to compute the fluency score, the granularity of the scale also differs per study. When the intelligibility was first developed and applied in the famous ALPAC report in the '60s, they designed the scoring to be assessed on a 9 point rating scale, where Scale 9 represented a perfect level of intelligibility and clarity while Scale 1 was a "hopelessly unintelligible translation despite a thorough posterior study" (ALPAC, 1966). Each stage of intelligibility was computed with an equal distance —upper scales being largely centered in stylistic features and word choices whereas the lower scales being concerned about syntactic components— so that raters could make an even judgment without difficulty on all items, according to Carroll

(1966).

Such fine-grained scale, however, was criticized to have more chances to cause a low intra-correlation of the raters as it required humans to do a meticulous calculation of the errors such as differentiating between Scale 4 —“Masquerades as an intelligible sentence, but actually it is more unintelligible than intelligible”— and Scale 3 —"Generally unintelligible”— (ALPAC, 1966) or the raters had to stick to their evaluation standard at all times (Dorr, 2009). In the same manner, Way (2018) claimed that inter-rater consistency was much higher for coarse-grained scales.

It, of course, does not insinuate that a binary division —intelligible and unintelligible— is the best scenario. Such yes-or-no division does not give much information about which factor affects the degrading of human comprehension of the sentences, hampering a distinction of a completely unintelligible and partially unintelligible sentence. Thus, among the range of minimum 2 to maximum 9 piles of scale, it is a common tradition to employ 4-7 point scales.

In fact, a range of the scale that TAUS suggests is based on a 4-point scale where Score 4 stands for the highest degree of quality while Score 1 equates to the worst quality, as shown in Table 3.1. The numeric scale can be accompanied by a categorical description (4-flawless; 3-good; 2-disfluent; 1-incomprehensible) as shown in the instance.

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.

Table 3.1: Rating scale of the fluency score[25].

The benefit of the fluency scoring is that it allows a human judge to

[25] dqf.taus.net

perform direct evaluation of each sentence. That is, a more fluent sentence earns a higher score, and thus, it achieves a perfect correlation with human judgment. Secondly, it is relatively convenient in selecting annotators as the task does not necessarily require bilinguals. Monolinguals in a target language are sufficient for the job.

3.4.2. Adequacy Scoring

The evaluation of an adequacy score is also one of the most widely used human methods. In the ALPAC report, it was denominated as *fidelity, which was defined as:*

> "[…] the rater was asked to gather whatever meaning he could from the translation sentence and then evaluate the original sentence for its "informativeness" in relation to what he had understood from the translation sentence. Thus, a rating of the original sentence as "highly informative" relative to the translation sentence would imply that the latter was lacking in fidelity." (ALPAC, 1966)

The same concept has been developed into what is now known as *adequacy* that concerns how well the content is preserved from the source to target text (LDC, 2002). The definition of the adequacy score of TAUS is pretty much the same:

> "It captures to what extent the meaning in the source text is expressed in the translation." (Görög, 2014)

The process of the adequacy evaluation is identical to the fluency evaluation: a panel of annotators making an intuitive judgment on an *n*-point scale from Scale *n* - Everything, the most positive, to Scale 1 - None, the most negative as in Table 3.2. The score can be computed by the same formula that is devised by ARPA:

$$Adequacy(P') = \frac{(J' - 1)}{(S' - 1) * N'} \tag{3.2}$$

where a judgment point (*J'*) that is divided by a total number of scale (*S'*) is normalized by a total number of fragments (*N'*) (White et al., 1993). The sole difference from the fluency score is that the normalization is based on a fragment, not a sentence. It can be also computed with a mean or mode average.

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.

Table 3.2: A 4-point scale of the adequacy score of TAUS (TAUS, 2010).

As the process and the computation of the adequacy evaluation are very much identical to that of the fluency evaluation, the two measurements are often averaged together. Bojar (2011) is the first to suggest the unified score of the fluency and adequacy scoring. Such combination, however, is still controversial, questioning if they have equal value in essence. On the one hand, some studies insist that fluent sentences are prone to be adequate because it is hard to imagine, in theory, that the translation contains entirely different contents (Arnold et al. 1994). On the other hand, some argue that the essence of the adequacy evaluation is not in sorting out drastic fallacies but in finding subtle blemishes of the translation (Way, 2018). In fact, Way (2018) points out that the adequacy evaluation is an indivisible tool particularly for NMT evaluation on the grounds that it is easy for the panels to be confused by the system output that contains tactful translations that are impeccably fluent but deliver slightly different information.

Moreover, unlike the fluency score that does not require a bilingual as an annotator, it is prerequisite to hire a bilingual that can understand both languages because the score represents a relevance of the source text to the target text. In practice, however, monolinguals are sometimes permitted to

perform the task with the help of high-quality reference translation (Dorr, 2009).

3.4.3. Segment Ranking

The ranking-based method was first introduced in the Association for Computational Linguistics (ACL) Workshop on Statistical Machine Translation (WMT) in 2007 (Callison-Burch et al., 2007). The segment ranking evaluation is a relative assessment of MT engines on a segment basis. It inquires into a position of a subject MT engine in comparison to other similar or different MT systems. Evaluators are provided with a number of translations in a random fashion without the information of their origins. They are asked to make a judgment on their rankings (1 = best; *n* = worst), with the possibility of a tie rank. By randomly distributing the sentences, a guessed judgment is avoided. When performing the task, the evaluators are generally provided with a source text, but the reference translation is optional.

A ranking score is then computed. It shows a distribution of favored or disfavored sentences of an MT system. Among various ways to compute the score, Han et al. (2016) suggest the following formula:

$$\frac{\textit{Number of better pairwise ranking}}{\textit{Total number of pairwise comparison - Number of ties comparison}} \quad (3.3)$$

While such calculation is a relative assessment whose individual figure is only meaningful in the given context, the calculation of TAUS is based on an absolute assessment that compares the engines at the same time makes an absolute judgment on each sentence. For example, when three candidate translations are of poor quality, all of them are ranked as 3rd, which is equal to 1 point. The scores are calculated by adding weights on each ranking. That is, in a comparison of three systems, each system that is selected as the 1st, 2nd, and 3rd, the system marks 3, 2 and 1 rank respectively.

To describe in detail the way of computation of the absolute segment ranking, Table 3.3 is given as a hypothetical situation where three MT systems (M1, M2, M3) are evaluated on three translation sentences (S1, S2, S3). The rank score of M1 for the three segments are computed as follows:

$$\Sigma \frac{W_j}{N} = \frac{1+2+3}{3} \quad (3.4)$$

where W_j stands for a weight (W) in the j-th time and N stands for the total number of the segment. The system earned 2 rank score, which means that the system M1 has favored by 66.7%. In the same way, M2 and M3 obtain equally 2.3 scores so that their percentile score is 77.8%.

	M1	M2	M3
S1	3	2	1
S2	2	2	1
S3	1	1	3

Table 3.3: Exemplary calculation of the ranking score for three segments (S = system translation, M = MT engine).

According to such formula, in the ideal scenario where a system is awarded all 1st ranks, it will earn 3 raking score ((3 + 3 + 3)/ 3) and in the worst scenario, 1 ranking score ((1 + 1 + 1) / 3). The benefit of such an absolute ranking comparison evaluation is that the problem caused by a tie ranking for not being able to give meaningful information (Denkowski and Lavie, 2010) is alleviated to a great extent.

Compared to the score-based methods described in the previous sections —fluency and adequacy scorings—, the segment ranking method does not require the evaluators to consider any judgment scales. In that sense, it can be said that the segment ranking evaluation is a more direct way of performing the evaluation task. The benefit of applying such a direct method is that it is easier for a human to make a consistent decision than to have to always consider certain criteria of scores (Duh, 2008). Callison-Burch (2007) even claimed that the ranking-based methods earned a higher correlation to a human parity than the score-based ones in their experiment. The ranking approach becomes more valuable than the score-based especially when the output translations are of similar quality.

The human judgment on ranking, however, is challenged when too many

engines are compared. It can impede human judgment. The evaluation is also difficult when each translation has different types of errors so that an evaluator has to decide their relative importance (Denkowski and Lavie, 2010). Albeit some limitations, it is still one of the most commonly used methods, and attempts are made to make better use of it in various ways. Callison-Burch (2007) proposed *constituent-based evaluation* where judges ranked the systems by considering only specific linguistic features that were highlighted after the source text are parsed and aligned to the candidate translations. Duh (2008) strived to automatize the process of the segment ranking by way of applying a Machine Learning technique.

3.5. Post-Editing & HTER

This section traces back to the remote antiquity of post-editing that is known to many as a recently-created concept but the initial foundation of the concept of post-editing dates back to when the idea of automation of translation was first discussed. After briefing the history of post-editing in Chapter 3.5.1, we focus our attention on the role of post-editing as MT evaluation method in Chapter 3.5.2. Their role is further enhanced to be included in the course of automatic methods, which creates a semi-automatic method explained in Chapter 3.5.3.

3.5.1. A Brief History of Post-Editing

Post-Editing is not a new concept that has recently emerged with an influence of the rapid development of MT. The early pioneers of MT have already envisioned the future relations between machine and human that would be affected by MT and considered pre-editing and post-editing as one of the human roles (García, 2012). The editor of a journal *Mechanical Translation's*, Yngve (1964), declared that "MT was a dream, and a machine might never be able to produce perfect translation […] The wide use of imperfect but useful mechanical translation may actually increase the demand for human translators" (García, 2012).

The infamous ALPAC report in 1966, however, was as fatal to MT as to post-editing, opening a period of latency for another decade. The report clearly stated that post-editing was a time-consuming job that was not only difficult to

perform but also meaningless to do because the quality of the final translation was not any better than a translation from scratch (ALPAC, 1966). Such dooming statement shifted the application of post-editing from a military sector on a governmental level to a commercial sector on a corporate level. Various corporations started to develop software that would facilitate the post-editing task from the mid 1980s (Vasconcellos, 1987). While various organizations, i.e. the European Commission and the PanAmerican Health Organization, began to apply post-editing in the course of translation and many post-editing-related conferences and journals were held in the end of ’80s and the beginning of ’90s, the upsurge of the interest of translators and stakeholders towards post-editing was visible after the development of translation memory in the ’90s (Koponen, 2016).

3.5.2. Post-Editing as an MT Evaluation Method

The exponential growth of MT and increasing demands for post-editing cast a question in terms of the relation between MT quality and post-editing speed such as: How much post-editing is needed in the given MT output? And from there, another question as to: How much post-editing effort does the MT output require? Such questions built a theoretical groundwork that higher quality of MT output would benefit the efficiency of post-editing while poor quality of MT output would complicate the post-editing task (Koponen, 2016). Based on such a hypothesis, post-editing came to be used as a key indicator of MT performance, and the vast amount of research have strived to measure the efficiency and efforts of a post-editing task. This line of research is different from the study about the post-editing act itself that questions into the influence of post-editors’ profile, a type of text, a language pair on post-editing.

3.5.2.1. Post-Editing Productivity

The typical experiment design to measure the post-editing productivity, or namely post-editing ratio, is to compare the time taken for post-editing and for translation from scratch. In theory, however, the two indicators cannot be measured in the same dataset by the same group of post-editors because by performing one of the tasks the post-editors get biased for having previous

knowledge of the text. As such, some research compare the hourly throughputs of post-editing obtained from an experiment to a standard value of the average human translation throughputs that are publicly known. Pinnis et al. (2016), for example, stated that the average human translation throughput in a medical domain was about 800-900 tokens per hour. Based on such information, they computed the post-editing productivity and reported a 200% increase in post-editing productivity. Escartín and Arcedillo (2015) criticized the credibility of such a comparison in that the throughputs are not real. They unpinned the claim by launching an experiment of post-editing. The result showed that there was a 250% productivity gain when the standard value was considered whereas only a 24% gain was observed when real translation-from-scratch throughputs were calculated.

Others attempt to obtain an objective result by making a time interval between the post-editing task and translation from scratch so that they can use an identical dataset and post-editors (Blagodarna, 2018). Others divide a dataset into two groups and let the post-editors translate from scratch and perform post-editing on the respective dataset. The proportion of each data size can vary, i.e. 1:1 (Escartín and Arcedillo, 2015), 1:2 (Zhechev, 2014) or undefined (Plitt and Masselot, 2010).

The basic unit of a post-editing productivity is words per hour (WPH), which are used to compute the post-editing productivity. It represents a proportion of the number of post-editing WPH to that of translation-from-scratch WPH. The time spent for other tasks such as web searching or looking up a dictionary should be preferably eliminated in the equation. TAUS, for example, indicates that “while actual translation and post-editing speed/throughputs will be different in other workbenches (i.e. slower due to a terminology look-up), the post-editing productivity ratio in this workbench should be a reliable indicator independent of the environment” (TAUS, 2010). The productivity ratio is then displayed as a percentile gain that shows how much time is saved to perform post-editing in comparison to translation from scratch, following the equation below (Escartín and Arcedillo, 2015):

$$\left(\frac{PE\_Throughput}{T\_Throughput} - 1\right) * 100 \quad (3.6)$$

where *PE* represents post-editing and *T* represents translation from scratch. The WPH of post-editing are divided by the WPH of translation from scratch, and such a figure is subtracted by 1.

The range of the post-editing productivity or productivity gain differs from research to research, influenced largely by the performance of MT engines, a language pair, domain, individual ability of post-editors and so forth. Zhechev (2014) reported that there was an average productivity gain of 37.13% in *en-pl*[26], 44.94% in *en-de*, 58.02% in *en-zh*, 59.07% in *en-jp*, 59.43% in *en-es*, 63.23% in *en-pt*, 63.6% in *en-it* and 92.33% in *en-fr*. According to their experiment, there was an 81.93% productivity gain in *en-ko*. Plitt and Masselot (2010) reported on average a 74% gain in all language pairs in the experiment (*en-fr, en-it, en-de,* and *en-es*). In a low-resource language pair such as English and Latvian, the post-editing productivity gain ranged from 70% ~ 145% in SMT and 34% ~ 118% in NMT (Skadiņa and Pinnis, 2017). Unlike the positive outcome, Groves and Schmidtke (2009) obtained a small gain of 6.1% in *cz-en*, 8% in *sv*[27]*-en,* 14.7% in *en-nl*[28], 14.5% in *en-fr,* 20% in *en-pt* (Brazilian) and 28.7% in *en-da*[29] in user manual domain and 5.9% in *en-zh*, 6.6% in *en-fr* and 16% in *en-de* in technical documents. Similarly, Escartín and Arcedillo (2015) achieved an average of 24.09% gain in *en-es* pair, ranging from -1.23% to 56.34%. Such a result proves that the productivity gain depends on a considerable number of factors.

Additionally, the translation and post-editing WPH vary drastically per person, as well. Plitt and Masselot (2010) stated that the absolute post-editing throughputs ranged from 400 ~ 1800 WPH. Tatsumi (2009) claimed that the mean average of WPH in the *en-jp* pair numbered 1,084.8 ~ 2,182.2 WPH. In

[26] Polish

[27] Swedish

[28] Dutch

[29] Danish

the case of Escartín and Arcedillo (2015), the translation throughputs ranged from 473 ~ 1701 WPH while the post-editing throughputs differed from 739 ~ 1,804 WPH. Such a result was quite dissimilar from the study of Sanchez-Torron and Koehn (2016) in the same language pair that acquired 716 ~ 887 WPH from nine post-editors evenly.

3.5.2.2. Post-Editing Effort

Other line of research in MT evaluation that takes advantage of post-editing is to measure how much effort MT outputs require to be post-edited in three aspects: temporal, technical and cognitive efforts (Krings, 2001). The temporal efforts are a good indicator that includes both the technical and cognitive efforts, but it is not always easy to collect accurate time (Koponen, 2016). A common way to monitor the correlation of post-editing effort and time is to score each segment on a level of 1 to 4 where *Scale 1 = complete retranslation* and *Scale 4 = no modification performed*, for instance, and compare the score with the actual post-editing time (Specia, 2010; Sousa et al., 2011).

While there are relatively intuitive measurements for the temporal efforts, ways to measure the technical and cognitive efforts are not standardized, and at some point, they can be measured by post-editing time itself (Koponen, 2012). In the case of technical effort, the conventional method is to track down keystrokes that post-editors make (Carl et al., 2011) or measure types of edits introduced by post-editors. In addition, an edit distance can also serve as a good automatic method in that regard (Tatsumi, 2009). The edit distance on a sentence basis is calculated by monitoring the number of edits introduced on a scale of 0 to 10 when a post-edited version and a reference translation are compared. 0 stands for a 0 distance between the two translations, which means almost no edits are needed in a sentence because they are almost identical. A 10 implies that “there is no similarity between the two translations or that the post-edited version has been made from scratch” (TAUS, 2010). In that sense, TER and HTER are a good source to detect the relations between types of edits and time taken for post-editing or the relations between types of edits and sentence length. Interestingly, Koponen (2012) performed POS tagging on post-edited system translations in order to find relations between TER metrics and

POS types. One of their findings was that edits of noun tended to achieve higher TER scores, meaning more effort was taken in editing nouns.

In the case of cognitive effort, many assumptions have been made. As mentioned previously, the more time it takes to post-edit, the more cognitive effort the task requires (Koponen, 2012). Or a longer gazing duration can represent a level of cognitive efforts. Carl et al. (2011) and Doherty and O'Brien (2009) used an eye tracker to measure the gazing time and fixation time in translation and post-editing, respectively, to measure cognitive efforts.

3.5.3. A Combination of Post-Editing and Automatic Methods

The influence of post-editing has extended from a mere act of converting a raw MT output to a final product of affordable quality to an evaluation method that measures a level of usefulness of MT in terms of time and effort. Post-editing is useful not only as a human method but also in automatic evaluation, in that it creates the closest reference translation and resolves the sparsity issue of the automatic evaluation. This section first does brief research about the architecture of mainstream automatic metrics —BLEU, METEOR, WER, and TER— as background information. And then, it introduces HTER as a promising way of mediating the advantages and disadvantages of human and automatic evaluation methods.

3.5.3.1. Mainstream Automatic Methods

The automatic evaluation methods have driven the development of the MT research community by enabling immediate monitoring of the performance of MT engines, as well as a comparison of multiple MT engines (Babych, 2014). Their successful implementation has been extended to the MT quality evaluation and compensated the costly, inconsistent and inefficient human evaluation methods. Most of the evaluation measures are based on a calculation of *n*-gram similarities between a hypothetical sentence, or so-called system translation, and reference translation. The most widely adopted metrics to date is BLEU (Bilingual Evaluation Understudy) by Papineni et al. (2002). It is precision-oriented metrics that measure the accuracy of the system outputs rather than recall-oriented metrics that measure a reproductivity of the reference

translation in the system outputs (Snover et al., 2009). The calculation of BLEU is based on the number of matched *n*-gram (usually from unigram to 4-gram) with the following formula (Papineni et al., 2003):

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}. \quad (3.7)$$

where the number of counts (*Count$_{clip}$(n-gram)*) is divided by the number of total *n*-grams counts of the system translations (*Count (n-gram')*). The score is then penalized by the length of reference translation in relation to the length of system translation. Such penalty helps to avoid a case where a *word salad* will obtain a high BLEU score. One of the benefits of BLEU is that it can take into consideration multiple reference translations that can increase a chance to achieve a better human parity.

Similarly, METEOR (Metric for Evaluation of Translation with Explicit ORdering) by Banerjee (2005) is also favored by many pieces of researches for its good correlation with human judgment. Unlike BLEU, it is a recall-based measure that is defined in these metrics as the number of matched unigrams divided by the number of reference translation (Snover et al., 2009). A precision-based measure divides the figure with the number of system translation. The singular point of METEOR is that a match is comprised of three types —exact match, stem match, and synonym match—, which is a way to consider linguistic aspects.

3.5.3.2. From TER to HTER

Despite the robustness that the popular automatic methods such as BLEU and METEOR exhibit on a lexical level, one of their greatest limitations is that they do not consider word order of a sentence. To compensate for such issue, some automatic methods approach the calculation from a different point of view that makes use of edit distance. The initial model of such kind is WER (Word Error Rate) that is based on Levenshtein distance (Levenshtein, 1966). These metrics detect deletion, substitution, and insertion by aligning elements from a

candidate translation to a system translation in such a way that each item is defined as (Snover et al., 2009):

- **Deletion**: a reference translation is matched to none
- **Insertion**: a system translation is matched to none
- **Substitution**: anything that does not belong to deletion, insertion and match

The biggest challenge of WER is that it does not capture reordering. In that context, TER (Translation Edit Rate) mitigates the problem by taking a segment movement into consideration in the form of *shift* (Snover et al., 2006). That is, a block of words that move to a correct position is computed as one edit, and it has equal value to other types of edits (deletion, insertion and substitution). In such way, the automatic metrics can measure reordering. Snover et al. (2006) define this metrics as "the minimum number of edits needed to change a hypothesis so that it exactly matches one of the references, normalized by the average length of the references", following the computation based on the formula given below:

$$\text{TER} = \frac{\textit{Number of edits}}{\textit{Average number of reference words}} \tag{3.8}$$

They found some limitations, however, such as the fact that it could not consider multiple reference translations at once as BLEU can do, and that it did not consider synonyms like METEOR can do (Denkowski and Lavie, 2010). Such limitation is inevitable as the initial purpose of TER when proposed in the Global Autonomous Language Exploitation (GALE) program (Olive, 2005), is to simply count the number of edits in a human translation (Snover et al., 2006).

It is improved later under the name of *Human-mediated* TER, called HTER. HTER (Human-targeted Translation Edit Rate or Human-mediated Translation Edit Rate) is "human-in-the-loop evaluation" that involves a human revision of a system translation called *Targeted Reference* (Snover et al., 2006). The core difference from TER is the fact that HTER requires human annotators to generate the closest reference translation to the system translation. In

addition, the normalization of HTER is computed on the number of reference translations, unlike TER that computes on the number of the source translation. In this way, the new metrics can make finer evaluation by not being affected by linguistic features such as a different but correct linguistic structure. In fact, Snover et al. (2006) claimed that HTER reduced 33% of the edit rate computed by TER. In fact, such a concept of creating targeted reference translation is also applied to the conventional BLEU and METEOR so that they are denominated as HBLEU and HMETEOR. They are contrasted to HTER and Snover et al. (2006) reported on the superiority of HTER over the two metrics.

HTER is considered as a semi-automatic method as it requires human post-editing in the process. In that sense, the metrics are human-oriented and the results are more consistent than those of human methods (Snover et al., 2009). However, the necessity to obtain post-edited translations makes HTER a more expensive and time-consuming method than other automatic methods. Snover et al. (2006) acknowledge that it takes 3 ~ 7 minutes for an annotator to post-edit one segment. In anyhow, it is still a widely-used evaluation method for its meaningful score that tells a good translation from a bad one —HTER = 0 means a perfect match that does not need any edits while HTER = 1 means a translation of the poorest quality.

3.6. Previous Studies of MT Evaluation & Post-Editing in a Korean translation

We have offered a broad picture of MT evaluation from its rich history to methodologies. This section shifts the focus from general MT evaluation to MT evaluation in the Korean language. We strive to introduce Korean-related former studies in the given field. As mentioned in advance in Chapter 1.2, only a few studies are available, and most of them are recently published. The likely first work in regard to MT evaluation in Korean dates back to 2013. D.Choi (2013) monitored how Google Translate dealt with certain linguistic features. Meanwhile, the seemingly first work in relation to post-editing in Korean was published slightly earlier than MT evaluation, in 2010. M.Hong (2010) investigated the utility of Google Translate in combination with pre-editing and post-editing. In terms of the *es-ko* pair, no study has been found in any of the

aforementioned sectors. With that in mind, this section presents the most related studies that delve into MT evaluation and post-editing in Korean in chronological order.

3.6.1. MT evaluation in a Korean translation

In the MT evaluation field, H.Choi (2016) in her Ph.D. thesis reasoned out a set of evaluation standards for patent translation from Korean to English. Based on them, they evaluated a freely available MT system *K2E-PAT* whose baseline was unspecified. The MT engine was run by Korean Intellectual Property Rights Information Service (KIPRIS)[30] in the domain of semiconductor-related patent abstracts. The bottom line of her evaluation was a qualitative examination by dint of interviews and analysis of the type of errors. The study reported that the quality of the MT system was too poor to satisfy its proper function in the given setup.

A.Chang (2017) carried out a descriptive and comparative analysis of seven freely available NMT engines and one SMT engine that could translate Korean and Chinese for both directions —SMT and NMT version of Papago, Genie Talk (interpretation application) of Hancom Interfree[31], PNS (Pure Neural Server) and exTalky (interpretation application) of Systran International[32], Google Translate and Baidu. Five Chinese-to-Korean translation outputs and five Korean-to-Chinese translation outputs from a variety of domain were analyzed one by one by the author's intuitive judgment. The study asserted that the NMT engine of Papago had the best performance for the translation of Chinese and Korean in both directions.

Hyoeun Choi developed further the previous diagnostic analysis into a quantitative experimental study in the following year in H.Choi and J.Lee (2017) by applying automatic and human evaluation methods —BLEU, fidelity, and readability—, with the experiment design identically maintained to that of H.Choi (2016). From our knowledge, it was the only NMT evaluation study (that was officially published before 2019) with applying the standard methods of the field

---

[30] http://www.kipris.or.kr/khome/main.jsp

[31] http://www.interfree.com/if/main/main.do

[32] http://www.systransoft.com

in a Korean-included language pair. The system gained 22.90 BLEU score, which was quite below the acceptable level (above 35.0 scores). In the human evaluation, the engine obtained an average 2.375 in fidelity scoring and 2.275 in readability scoring when assessed on a scale of 1 to 5. All in all, the results again revealed that the performance of the aforementioned MT engine was not enough to tackle a patent translation from English to Korean professionally.

In a similar manner, S.Kim and H.Lee (2017) compared the performance of unspecified NMT and SMT systems in embedded sentences obtained from an *en-ko* translation in a movie script, from a descriptive and phenomenological standpoint. The data summed up to 179 embedded clauses composed of simple and complex sentences. The study stated that NMT had less syntactic errors than SMT while they tended to produce an out-of-context translation that had nothing to do with the original text.

Y.Ki (2018) sought to investigate the state-of-the-art of the performance of online machine translators based on NMT in a *ko-zh* translation. The assessment was performed according to a series of evaluation criteria developed personally by the author. It was based on a 4-point scale and put a penalty on errors by their type. The comparison of Google Translate and Papago with 580 simple sentences in a textbook showed that 60% of the translation outputs obtained favorable scores, among which Google Translate earned a mean average of 3.4 points. It was about 0.28 percentage point less than that of Papago.

B.Kang and J.Lee (2018) assessed three NMT-based online translators —Google Translate, Papago and Baidu— for Chinese-to-Korean translation in an attempt to understand the principle of NMT engine. The assessment was based on a comprehensive calculation of fidelity and accuracy by the author on a scale of 0 to 10, penalizing errors. In the first dataset of 170 sentences composed of colloquial expressions, Papago featured the best performance while Google ranked the lowest average score. In the rest of the dataset of 200 sentences extracted from written texts, on the other hand, Google yielded the highest rank, slightly over that of Papago with a 0.02 percentage point.

3.6.2. Post-Editing in a Korean translation

The most relevant study was Zhechev (2014) who evaluated Autodesk's in-house SMT system, which was built on open-source Moses on a basis of automatic evaluation and post-editing productivity for a translation from English to nine different languages together with Korean. In total, 18 directions of translation were studied. Notwithstanding some promising results obtained from their experiment, the evaluation of SMT system was not our major concern. Moreover, the study claimed that all data acquired for the *en-ko* language combination were presumably unreliable due to the fact that one of the participants made a mistake of translating all segments from scratch instead of practicing post-editing. There was an 81.93% productivity gain of post-editing over translation from scratch in this pair.

Sangbin Lee conducted a series of extensive evaluation studies of post-editing in an attempt to observe i) post-editing as a product, ii) perception of post-editors and iii) a post-editing process, respectively, from the pedagogical point of view. S.Lee (2017) analyzed error typology of NMT output (in a medical domain) of English to Korean that was post-edited by five university students in translation studies. He reported on major errors that novice post editors tended to commit, such as revision without having the context in mind and ungrammatical/unnatural expressions or terminologies that were left undealt with. In S.Lee (2018a), he carried out a phenomenological study of post-editing process based on data collected via interviews and mind maps of the same post-editing group. The study observed that the editors felt that although the qualification of a post-editor was in collusion with that of a translator, the post-editing task was rather challenging and that specialized training was necessary. He also claimed that the quality of NMT (Google Translate in this study) was much better than their expectation. His final research strived to investigate a cognitive process of post-editing by collecting simultaneous comments while the editors practiced post-editing and recorded their working screen. As the objective of this study was beyond our scope, the detailed discussion was referred to S.Lee (2018b).

J.Kwak and S.Han (2018) analyzed the error type of NMT outputs generated by three general-purpose NMT engines —Google Translate, Papago,

and Systran for English to Korean translation. The translation data of the study was the one previously built in ‘AI vs. Human Translation Competition’ held in February 2017 at Sejong University, Seoul. Eight graduate students of translation studies annotated the error typology on a dataset composed of 453 words from two texts —literary and non-literary— in accordance with the standard of ISO 18587: 2017. They also labeled a level of gravity of the errors in four stages —critical, major, minor, preferential. They excluded sentences that obtained inconsistent assessments and considered only 12 and 9 sentences from each text type valid where more than two evaluators’ judgments coincided. They found out that the major error of NMT in the *en-ko* translation was related to a semantic clarity and morphological accuracy. In detail, the system translation of the NMT engine exhibited untranslated segments and omissions. Such result was different from the error classes that were observed from human translations, usually an over-translation.

3.6.3. Survey: the State-of-the-art of MT & Post-Editing Usage in a Korean Translation

When Lommel and DePalma (2016) launched extensive research over the usage of MT and post-editing in the commercial hemisphere, most of the professional translators, in general, and freelance translators, in particular, showed their hostility against the idea of the machine taking away their job, with 83% of negative opinions. Such repulsion was still witnessed in the 2018 European Language Industry survey, with only 8% of independent professional translators maintaining a positive stance towards them. They were quite surprising statistics considering that 56% of the providers of post-editing were European countries and the most demanded languages in terms of post-editing were French, Spanish and Germany (Lommel and DePalma, 2016). Like it or not, the demand for MT and post-editing seemed to grow fast, just by seeing that 82% of the LSPs of less than one year experience offered post-edited MT (Lommel and DePalma, 2016).

While such a hot trend of MT and post-editing in the translation field was global, it seemed to be premature in Korea or Korean as the low number of related articles proved in Chapter 1.2. The number of published articles could

show the trend in the educational field, but no survey or study had been proposed as to the current usage of MT and post-editing in Korea or in Korean, in general. We, therefore, investigated the state-of-the-art of the usage of MT and post-editing in relation to Korean by launching a survey in both domestic and global markets. The objective of the survey was to obtain the most up-to-date information about the utility in the two sectors.

Two versions of the questionnaire were prepared: the first survey (Survey I) was concerned with the usage of MT and post-editing in the *es-ko* pair in a global translation market, and the second (Survey II) dealt with their usage in any language pair in the Korean translation market. The duration of the survey was 30 days and was launched in November 2018. Survey I under the title of "Machine Translation & Post-Edition for a Spanish-Korean language pair" was distributed to 100 LSPs ranked in the list of Top 100 LSPs worldwide in 2018 (by revenue) by Nimdzi (Nimdzi 100, 2018). Survey II, on the other hand, titled as "Machine Translation & Post-Edition in Korea" was distributed to 18 institutions and LSPs (including some multinational companies who had an office in Korea) in Korea. The small number of the target in Survey II was due to a different market structure in Korea where a small number of giant LSPs dominated the market.

The survey was composed of 15 questions in Survey I and 17 in Survey II. All contents are identical except the additional two questions in Survey II concerning a working source and target language. The questions of Survey I are as follows:

1) Age
2) Mother tongue
3) Years of experience in translation
4) Years of experience in post-editing
5) Please indicate the name of the institution/agency you belong to at the moment.
6) Please choose your main source and target language pair for post-editing.

7) How much volume does the post-editing account for in your workload?
8) In the case of the *es-ko* translation, which MT engine is mostly applied to your work?
9) In which domain do you usually perform post-editing for the *es-ko* pair?
10) Your impression of the accuracy of MT results: has it improved for the Spanish-to-Korean translation, compared to the past?
11) Your impression of the accuracy of MT results: has it improved for the Korean-to-Spanish translation, compared to the past?
12) In which direction is post-editing more required?
13) Which is the post-editing tool that facilitates your work for the *es-ko* pair?
14) How many edits do you think you perform for the *es-ko* pair in post-editing?
15) Among the three following edit types, what you think is the most frequent edit for post-editing the *es-ko* pair?

And the questions in Survey II are as follows:

1) Age
2) Mother tongue
3) Years of experience in translation
4) Years of experience in post-editing
5) Please indicate the name of the institution/agency you belong to at the moment.
6) Please specify your main language pair for post-editing.
7) Please specify your main source language (A language) for post-editing.
8) Please specify your main target language (B language) for post-editing.
9) How much volume does post-editing account for in your workload?
10) Which MT engine is mostly applied to your work?

11) In which domain do you usually perform post-editing?
12) Your impression of the accuracy of MT results: which direction is better?
13) Your impression of the accuracy of MT results: has it improved for the A-to-B translation, compared to the past?
14) Your impression of the accuracy of MT results: has it improved for the B-to-A translation, compared to the past?
15) Which is the post-editing tool that facilitates your work?
16) How many edits do you think you perform?
17) Among the three edit types pre-defined below, what is the most frequent edit for post-editing?

The questionnaires were prepared in Google Forms and were distributed by e-mail. With the link of the survey, they were notified that the purpose of this survey was to investigate the current usage of MT and post-editing of active professional translators in Korea, in general, and in the *es-ko* pair, in particular, and that the data obtained from the survey would be used for research purpose only.

Despite the ambitious goal it had, the response rate of the two surveys was considerably low. In Survey I, only 57 companies out of 100 (57% response rate) reacted. In Survey II, 11 bodies of 18 (61% response rate) answered to the e-mail. In addition to the low response rate, their answers were unfortunately negative. On the one hand, the majority of the responses for Survey I was that they were irrelevant to this survey because they did not provide service for the *es-ko* language pair. Only one company from Spain and one from Taiwan participated in the survey. On the other hand, most of the responses for Survey II were that they offered neither MT nor post-editing service, that their translators were not particularly instructed to apply MT in the course of work, nor they were paid to perform post-editing.

Considering the number of participants, the survey could not make any conclusion about the usage of MT and post-editing in the given pair. However small the response rate was, it still gave information about the terrain of Korean-related market. Firstly, in a global market, the *es-ko* pair was not a widely-

demanded language combination. As the contacted companies were mostly based in North America, Asia, and Europe, however, there was a chance that this distant language pair might be more demanded in Latin America. There was also a positive chance of its demands in Spain as only two Spanish LSPs were included in the Top 100 list. The demands for the *es-ko* pair in Spanish-speaking countries needed a further investigation. Secondly, in a domestic market, the fact that 60% of the translation-related leading bodies claimed that they were irrelevant to this survey implied that MT and post-editing were still premature in not only academic but also in the commercial hemisphere in the Korean-related pair, on the premise that they told the truth. All in all, a further investigation was necessary to confirm the usage of MT and post-editing in relation to the Korean-related pair.

3.7. Chapter Summary

Chapter 3 provided a theoretical background of MT evaluation in general and post-editing in particular. In Chapter 3.1, we defined the concept of equivalence in the field of MT evaluation that was dissimilar to that in translation studies: a human parity. That is, a human translation was an ideal scenario that a system translation could achieve in MT evaluation. With that aim, various evaluation methods were devised, which was described in Chapter 3.2. According to a traditional typology of glass box and black box, an MT system could be evaluated in a preset or general environment. In a broad sense, human and machine could be a performer of the evaluation, which was denominated respectively as human/manual evaluation and machine automatic evaluation. The up-to-date evaluation methods of each type were detailed accordingly by the typology suggested by Han et al. (2016).

In Chapter 3.3, we discussed the advantages and disadvantages of the human and machine evaluations. While the human evaluation was comprehensive and a best way to achieve a human parity, the results could be inconsistent and subjective. In reverse, the automatic evaluation was partial and lack of correlation to a human parity, but the results were consistent and objective. From the discussion, we concluded that the automatic evaluation was not employed for this thesis for its particular weakness in distant pairs. To

enhance the balance of our evaluation, semi-automatic evaluation, HTER, was presented.

In Chapter 3.4, three human evaluation methods that were applied in our experiment were detailed —fluency scoring, adequacy scoring and segment ranking— from their definitions and history to their systems. The also manual evaluation, post-editing, was separately assigned in Chapter 3.5 along with HTER to stress its importance in our experiment. In Chapter 3.5, we clarified that the concept of post-editing already existed when the idea of MT was first devised. We approached post-editing from a perspective of its role as an MT evaluation method. We described how post-editing time and effort could be measured and could contribute to the evaluation result. The remainder was dedicated to HTER that made use of post-editing in the loop of automatic evaluation. With a brief information about the general automatic evaluation, we showed that HTER was a promising method for this thesis.

We took on research in Chapter 3.6 about the previous studies of MT evaluation and post-editing in a Korean-related language pair. Albeit few, they demonstrated where the interests were going in this regard. As an additional work, we initiated a survey inquiring into the current usage of MT and post-editing in any Korean-related language pair in a Korean market and in the *es-ko* pair in a global market, which was reported in Chapter 3.7. The survey was not successful in that the respondents rejected either because they did not service MT or because they did not provide service for such language pair. Although the result of the survey could not achieve the intended purpose, it reconfirmed that MT evaluation and post-editing were in a nascent stage in the Korea-related language pair.

[blank page]

PART II. EXPERIMENT

Having been equipped with the background information about NMT and MT evaluation, Part II deals with a pilot study that is designed to facilitate a well-prepared experiment. In Chapter 4, a detailed procedure of the pilot study and its result are discussed. Based on the feasibility confirmed from the pilot study, Chapter 5 illustrates the main NMT evaluation experiment, which is composed of fluency & adequacy scoring, segment ranking, and post-editing. As a semi-automatic method, we also report HTER scores.

The two chapters are designed almost identically. The parameters of the dataset, evaluators' profile, and workbench are detailed. And then, we explain how the pilot study or the main experiment are conducted in chronological order. They are composed of a training session, the first evaluation task of post-editing, and the second evaluation task of fluency & adequacy scoring and segment ranking. The experiments end with collecting feedback from the participants.

[blank page]

4. PILOT STUDY

> Pilot studies play a key role in the development or refinement of new inventions, assessments, and other study procedures. Commonly, results from pilot studies are used to support more expensive and lengthier pivotal efficacy or effectiveness studies … The fundamental purpose of conducting a pilot study is to examine the feasibility of an approach that is intended to ultimately be used in a larger scale study. This applies to all types of research studies … (Leon et al., 2011).

With the same purpose defined above in Leon et al (2011), the current thesis carries out a pilot study to perfect the actual experimental procedure of NMT assessment. In this chapter, an elaborate description of a pilot study preparation, experiment conduct, and analysis of the obtained data will be delivered.

In Chapter 4.1, the aim of the pilot study is enumerated from both the general and particular point of view. In Chapter 4.2, the design of the pilot study composed of three evaluation types —fluency and adequacy scoring, segment ranking and post-editing— are listed, from evaluation criteria to the dataset, volunteers' profile and their training. Having clarified the design of the experiment, Chapter 4.3 reports on the conducting of the study in a chronological way from day 1 (Chapter 4.3.1) to day 2 (Chapter 4.3.2). The results acquired systematically from the workbench are described method by method in Chapter 4.4. In the final Chapter 4.5, the objectives discussed in Chapter 4.1 are verified to see if they are accomplished.

4.1. Overview

The aim of the pilot study, in general, is to clarify the objectives of introducing the preliminary stage to the experiment and define the parameters that will be tested in the principal experiment. The motivation of implementing the pilot study comes from the intention of checking the variables set *a priori* for the experiment so that the final results are unbiased and reliable.

- It is a meta-objective to verify the properties of the experiment. The study sees if each variable serves its role properly by carrying out a small-scale experiment with the identical experiment design.
- In addition to verifying the pre-set parameters, the pilot study inspects minutely if there are some default variables that are unexpectedly harming other settings. The study also monitors if there are some properties that have not been taken into consideration.
- The study observes the participants' degree of understanding of each task so as to improve the guidelines and training session.
- The study tests the workbench if it works without errors in all stages.

To this aim, this section details each element that will be verified in the pilot study. The system in question is Google Translate of version 2018. The scope of verification covers four categories in all evaluation tasks: methodology, dataset, workbench, and training.

Methodology

- In fluency and adequacy scoring, this study sees if the evaluators distinguish the concept of fluency and adequacy well, as they are distinct methods but carried out in one platform.
- In segment ranking, this study sees if the evaluators understand the absolute score of the ranking score and award scores in an appropriate manner. The absolute score system, compared to a relative one, requires that if the three candidates are equally poor, they are marked by the lowest score, which is in this experiment Score 3, instead of the highest one. So, 1-3-3 or 2-2-2, for example, are a possible scenario.
- This study monitors if the order of the sentences is randomized.

	- In post-editing, this study sees if the post-editors follow the provided guideline. - This study confirms if half of the dataset is provided for post-editing and the other half for translation, randomly. - This study detects hindrances to the working environment.
Dataset	- This study examines the appropriacy of the domain and text type to the experiment.
Workbench	- This study observes the stability of the system in all stages; such as if the server is connected without errors or if the data is recorded securely. - This study monitors the interface if it causes confusion to the users.
Training	- This study monitors if the guideline is comprehensible and well explains the core idea of each task. - This study collects feedback from the participants to modify and improve the training session.

4.2. Study Setup

Under the aforementioned objectives of each part, the pilot experimentation was organized on the identical conditions to the main experiment. This section describes the detailed features of the pilot study. Chapter 4.2.1 indicates methodologies of the pilot study —fluency and adequacy scoring, segment ranking, and post-editing— with a brief description of each method. Chapter 4.2.2 gives information about the dataset of the pilot study, which is a transcription of an audiovisual script in a radio program. Chapter 4.2.3 introduces a profile of the participants and Chapter 4.2.4 presents a training session organized for the participants.

4.2.1. Evaluation criteria

This pilot study is composed of four (or three depending on a viewpoint) kinds of manual evaluation methods: fluency scoring, adequacy scoring, segment

ranking and post-editing. The specific properties of each method are referred to Chapter 3, and the core idea is briefed in this section.

Fluency & Adequacy Scoring The *fluency* and *adequacy* scoring is one of the most common methods in MT evaluation. Even though the two scoring types seem like a single method, they should be calculated as a separate criterion in this pilot study. The definition of *fluency* in TAUS is how "grammatically well-formed" a target text is, and *accuracy* is "how much the meaning of a source text is preserved in a target text" (Görög, 2014). The process of the evaluation is to ask a panel of judges to assign an *n*-point Likert scale on each sentence and to compute an average score on an aggregated basis. The calculation of TAUS is based on a 4-point scale of 1 being negative and 4 being positive, which is in *fluency,* 4 = Flawless; 3 = Good; 2 = Disfluent, 1 = Incomprehensible while in adequacy, 4 = Everything; 3 = Most; 2 = Little, 1 = None.

Segment Ranking A relative comparison of the performance of three translations will be carried out, between Google Translate, Kakao i and human translation. In the preliminary stage of the pilot study, the initial idea was to use Papago[33] instead of Kakao i because Papago was one of the few well-known neural-based free MT engines created by a Korean company Naver Inc. This engine could handle 32 translation directions with the *es-ko* translation included. The motivation came from multiple pieces of research claiming that Papago was considerably robust in Korean-related translations in both direction (Korean from/to a foreign language) because the engine had a Korean-centered approach (H.Lee et al., 2016). Having gone through a feasibility study, however, it was concluded that the performance of Papago was, in fact, too poor in the *es-ko* language pair to introduce to the current study mainly because this engine was developed for interpreting (H.Lee et al., 2016). Therefore, Google Translate is contrasted to the system translation of Kakao i and human translation in an anonymous way and on an absolute scale of 1 to 3. The absolute ranking lets the comparative ranking score serve as an absolute score

[33] https://papago.naver.com

as well. For instance, if the three translations are equally poor, an evaluator will mark 3-3-3 instead of 1-1-1.

Post-Editing The time taken for post-editing and translation from scratch are measured to monitor if post-editing is more productive than translation from scratch from the perspective of speed and throughput. A full post-editing is required to the evaluators, a quality that is similar or equal to human translation. The definition of full post-editing in TAUS is “being comprehensible, accurate, stylistically fine, though the style may not be as good as that achieved by a native-speaker human translator” (Görög, 2014). The detailed information is presented in Chapter 5.3.2. The participants will translate half of the dataset from scratch and post-edit the rest half. The productivity ratio is computed based on words per hour (WPH) of the two tasks.

4.2.2. Pilot Dataset

As the main experiment attempts to test the performance of Google Translate on the newswire domain, the pilot data is collected from the same domain type. The dataset is extracted from a book “시사 스페인어 (Sisa Spanish)” where an original Spanish radio script is transcribed and translated in Korean. The script belongs to a Spanish radio program named *Cinco Continentes*[34] of the Incorporation of Spanish Radio and Television (Corporación de Radio y Televisión Española, S.A., RTVE) that deals with global issues in plentiful sectors. The text type of such data is informative as well as narrative. The reason for having newswire as a domain is in line with the characteristics of freely available online MT engine of being intended to the lay public. It has come to our attention that the major challenge of employing news texts is that available parallel corpora are scarce between the *es-ko* pair. In that sense, the above-mentioned book is an incredibly invaluable asset to this study, in that an elaborated official translation is already prepared in the given domain. The dataset for this pilot study is excerpted from a political section, and the topic of the text is Ebola disease. Despite the thematic characteristics, the contents of

[34] http://www.rtve.es/alacarta/audios/cinco-continentes/

Newspaper	Domain/Genre	Size (w)	Size (s)	Source
RTVE	News/Politics	1,517	98	“Sisa Spanish”

Table 4.1: Description of the dataset of the pilot study.

Number of Sentences		98
Number of Words	ST	1,517
	GT	935
	KT	860
	RT	859
Minimum Sentence Length		1
Maximum Sentence Length		48
Average Sentence Length		15

Table 4.2: Size of the dataset of the pilot study. (ST = source text, GT = system translation of Google Translate, KT = system translation of Kakao i, RT = reference translation).

the text are taking a political stance, not medical. In a nutshell, the dataset of this pilot study is a newswire concerning Ebola disease in a political section. The summarized information of the dataset is provided in Table 4.1, and the further analysis of the dataset including the size of the system translations and reference translation, maximum and minimum sentence length and average sentence length is given in Table 4.2.

As stated in advance, the reference translation is publicly available. The system translation of Google Translate and Kakao I was obtained on October 13, 2018. For reference, the definite architecture of the two NMT engines is confidential and so, it cannot be explained in this thesis.

4.2.3. Volunteers’ Profile

Four university students of various grades are willing to participate in this mini-study without any financial compensation. The number of evaluators is smaller than the main experiment as usually is a pilot study because it is a preparation step for larger experimentation (Leon et al., 2011). Due tot a budget limitation and scarcity of professional translators in the given language pair unlike the *en-*

ko pair, the participants are recruited from the Spanish department of one university. Although the level of bilingualism is varied, all of them are bilinguals of Spanish and Korean.

In comparison to the principal experiment setup, there are minor variations across all properties of the participants including their social position. The volunteers in this pilot study, for example, are majoring in Spanish unlike those who are in an unrelated field of study in the main experiment. The year of working experience cannot be counted but all of the students have previous experience in translating (for the *es-ko* language pair). One salient contrast is that a male evaluator is participating and that some *simultaneous* bilinguals[35] of Spanish and Korean are included. It is, in fact, an intriguing point how the level of bilingualism of both the source and target languages would influence on the evaluation result. However, the standpoint of the present study towards NMT evaluation is to take it as a stepping stone to infer the performance of a system, not to investigate the nature of the post-editing itself. Thus, the impacts of the level of bilingualism and their linguistic or translational background, not to mention other qualities, are not of the core interest of this study.

The most crucial part of their profile in this phase is rather the fact that the volunteers are not overlapped with those in the main experiment. Not a single person from the pilot group participated in the main experiment. All of them are novice post-editors. All of the volunteers were introduced to this pilot study by a mutual acquaintance. As a matter of convenience, they are denominated as V1, V2, V3, and V4. The details of the evaluators are described in Table 4.3.

	School Year	Relevance of the major	Gender	Mother Tongue
V1	3rd year	Related	Female	Spanish, Korean
V2	4th year	Related	Female	Korean
V3	2nd year	Related	Female	Korean
V4	2nd year	Related	Male	Spanish, Korean

Table 4.3: Profile of the four volunteers in the pilot study.

[35] The term simultaneous bilingualism follows the definition of Liddicoat (1991) that the two languages are innate.

4.2.4. Evaluator Training

As the evaluators are introduced by a mutual acquaintance, a face-to-face training session is organized. It is assumed that such simultaneous communication will be of utmost help to design the main experiment. Considering that this pilot study is carried out on-site, most of the details are explained verbally at the same time a brief pamphlet containing all information about the project —a purpose of this pilot study, a definition of post-editing, general post-editing rules, MT evaluation metrics, and the dataset— was distributed. Some of the core information of the hand-out is given below:

Post-Editing Guideline

Notion of Post-Editing	"Post-editing (or postediting) is the process whereby humans amend machine-generated translation to achieve an acceptable final product. A person who post-edits is called a post-editor." (Wikipedia)
Post-Editing basic rules from TAUS	• "Aim for grammatically, syntactically and semantically correct translation. • Ensure that key terminology is correctly translated. • Ensure that no information has been accidentally added or omitted. • Use as much of the raw MT output as possible. • Basic rules regarding spelling, punctuation, and hyphenation apply. • Ensure that formatting is correct." (Görög, 2014)
Post-Editing additional rules	• Do not hesitate too long over a problem. • Do not embark on time-consuming research. • Do not edit spacing.
Parameters	• Speed of translation • Speed of Post-Editing • Number of edits

Data Information	• Domain/Genre: News/Politics • Theme: Ebola diseases • Size: 1,517 words / 98 segments • Source Language: Spanish • Target Language: Korean

The definition of post-editing is extracted from Wikipedia[36] and the post-editing basic rules are from a TAUS guideline (TAUS, 2010). While the basic rules emphasize the way of post-editing, the additional rules are designed to give some constructive advice.

The editors are allowed to question the task as much as they need before initiating the project. Furthermore, the first five sentences of each evaluation categories are considered to be a sample segment where their understandings of the task will be monitored. The simultaneous feedback about the overall experiment setup inclusive of post-editing rules is attained on those sentences.

4.3. Procedure

Under the configuration, the pilot study was carried out on November 1, 2 of 2018 in a quiet place with a stable Internet connection. The task was scheduled for two days: one day for post-editing evaluation (Chapter 4.3.1) and the other day for fluency and adequacy scoring and ranking comparison (Chapter 4.3.2).

4.3.1. Day I: Training & Post-Editing

The four participants and the project manager were gathered on-site with one's own laptop prepared. Prior to the initiation of the work, the volunteers were asked to go over issues such as a stability of the Internet connection, power of the laptop or preparation of Internet dictionaries in a separate window if needed. A special request was made to turn off their cellphones in order to prevent any disturbance from this work.

Having arranged an appropriate working environment, the general purpose

[36] https://en.wikipedia.org/wiki/Postediting

of this study was explained. They were aware of the fact that it was a pilot study for the main experiment of a doctoral thesis and that the data from the pilot study would be used for the purpose of the dissertation. Upon their agreement, a guideline for post-editing was distributed. A Q&A session lasted for a couple of minutes after they read through the guideline. At the same time reading the written guideline, they were allowed to ask questions. Many of the volunteers inquired about what post-editing meant, saying that they had not heard about this term (or work) before. Some of them declared that the detailed examples of the post-editing level let them have a good grasp on the task. The training session did not include a warm-up session as the direct communication facilitated active communication, and the post-editing task began immediately.

The first five sentences were reviewed by the project manager. The frequent mistakes were detected due to the confusing interface of the workbench. As an instance, despite the modern design of the platform, the evaluators tended to regard the previous or next segment as the current one. Adding fuel to the fire, they habitually pressed the Enter button, which led them to skip to the next segment without giving an option of coming back. It seemed like it was a general problem happened to every member of the group. After a few trial and errors and the simultaneous feedback, the volunteers got some ideas about the task.

During the experiment, it was found out that the speed of the Internet was quite influential on the performance because an unstable connection disrupted their concentration on the task. Moreover, facilitating quiet space was another crucial factor for the task. All of the participants stated, after the pilot study, that it was surprising how NMT could process the *es-ko* translation with such accuracy. One of them highlighted that despite some errors, the system translation gave her an impression that the engine might consider contextual information and that NMT took into account a syntactic structure of sentences.

4.3.2. Day 2: Fluency/Adequacy Scoring & Segment Ranking

On the second day, the same group of volunteers gathered in a different place with a better Internet connection. Similar to the previous day, a condition of the working environment was previously confirmed, and the second guideline was

disseminated. It contained a general description of the two tasks (fluency and adequacy scoring and segment ranking) and an instruction on how the evaluation scale was composed of and how the workbench looked like. Some of the volunteers requested more information about the difference between fluency and adequacy scoring. Besides that, everyone seemed to have understood the task well, and the study was initiated subsequently.

The fluency and adequacy scoring was launched. The volunteers were asked to feel free to pace themselves as no time constraint existed in this type of evaluation. Some expressed confusion when setting their judgment criteria of the segment right, but it seemed to be resolved in no time. One volunteer expressed that the task was challenging because he was not satisfied with the wording of the system translation but could not come up with a proper replacement either. In many cases, it was observed that their judgment got indecisive for long sentences and complex structures.

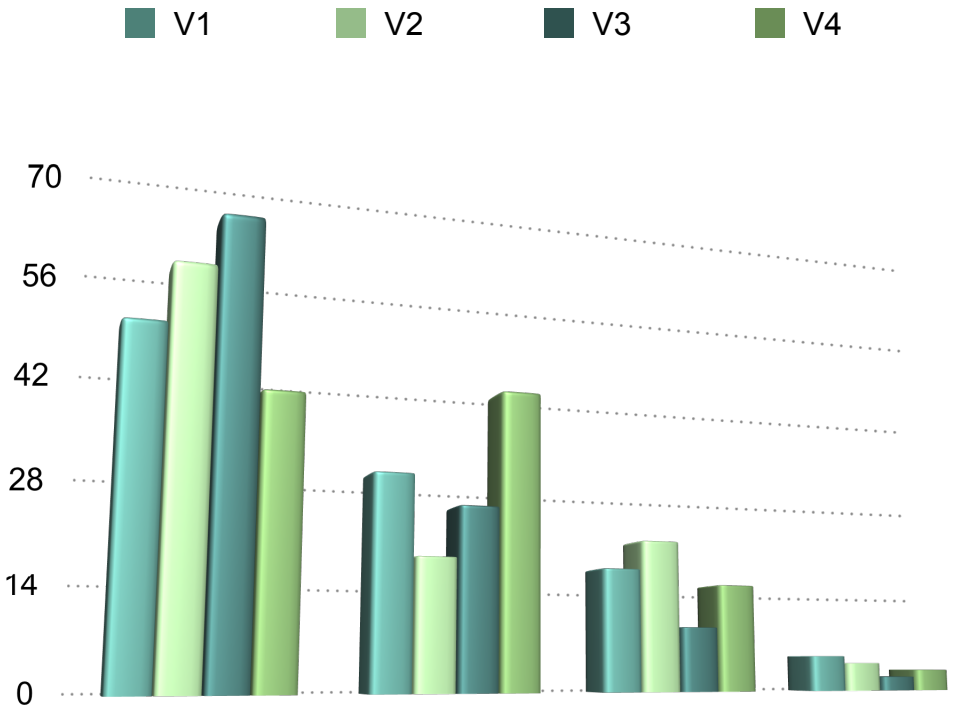
A 20-minute break was given before the segment ranking comparison. The project manager reviewed if the data was collected securely on the database and sent an email for the third evaluation task to the participants. They were not provided with any advance information about whose translations they ranked. Although the reference translation was not provided when they ranked the segments, everyone seemed comfortable with the task because of the previous two-fold evaluation on the same dataset. They stated that the user interface of TAUS DQF got comfortable after going through it in advance.

4.4. Result

It is not of our interest to analyze the result of this pilot study. Nevertheless, it is important to observe what kind of data can be obtained from the experiment. As such, this section reports on the result of the pilot study in a quantitative and qualitative way. The majority of the analysis was obtained systematically from the workbench. Each evaluation method is described in Chapter 4.4.1 (fluency scoring), Chapter 4.4.2 (adequacy scoring), Chapter 4.4.3 (segment ranking) and Chapter 4.4.4 (post-editing). HTER scores are also described in Chapter 4.4.4.

4.4.1. Fluency Scoring

The mean average fluency score of Google Translate in the *es-ko* translation is 3.39 of 4 and the mode average is 4 of 4. The result shows that the outputs of NMT are 84.75% well-formed and grammatically correct in the given setup. The distribution of the fluency scores per volunteer is presented in Table 4.4 and Figure 4.1. About 51.8% of the total segments (98 sentences) are considered as flawless (of Scale 4). With the disfluent (of Scale 3) segments included, about 80.6% of the dataset is nearly perfect. There are still 3.6% of the segments which are incomprehensible (of Scale 1) to human, but the overall fluency score seems consistently high.



	Flawless-4	Good-3	Disfluent-2	Incomprehensible-1
V1	47	29	17	5
V2	55	18	21	4
V3	62	25	9	2
V4	39	41	15	3
Avg	***50.75***	***28.25***	***15.5***	***3.5***

Figure 4.1: Distribution of fluency scores per volunteer and their average. (unit: %)

Table 4.4: Distribution of fluency scores per volunteer and their average. (unit: %)

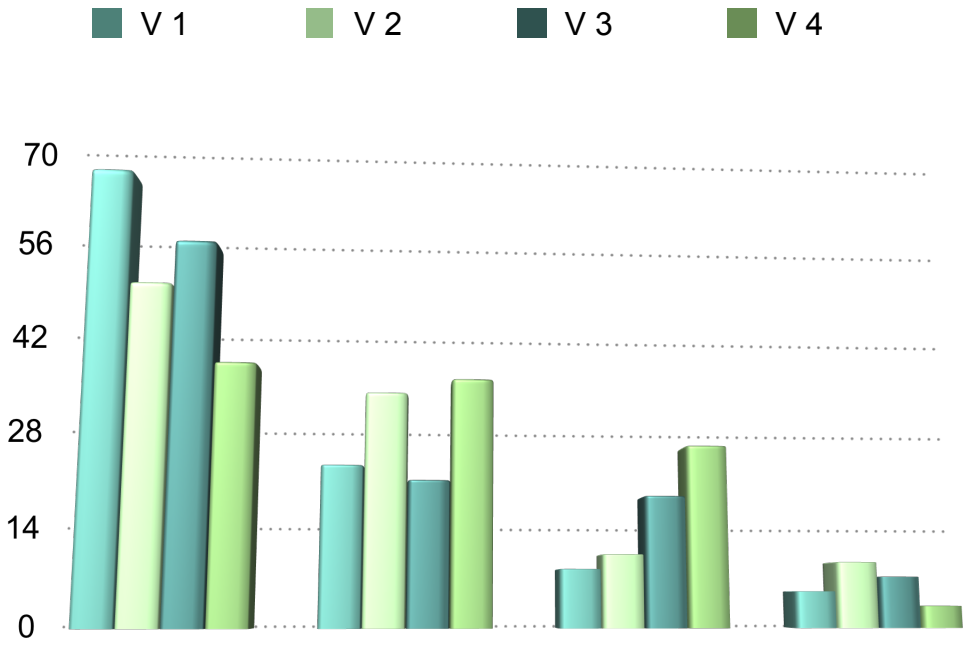
From the perspective of the volunteers, the one who gave the most favorable score is V3, with 88.8% of the segments regarded as fluent enough (of Scale 4 and Scale 3) and 67.3% being perfectly fluent (of Scale 4). The most critical evaluator, on the other hand, is V1, with 77.6% of fluent enough segments. The overall tendency, however, is that the biggest pie belongs to Scale 4 - Flawless (exception is made by V4 who assigns more scores on Scale 3.) while the smallest pie goes to Scale 1.

4.4.2. Adequacy Scoring

The Google NMT obtained a mean average of 3.23 of 4 and a mode average of 4 of 4 in the translation of Spanish to Korean. It is found that the result of adequacy score (80.75%) is as positive as that of fluency score (84.75%) but slightly lower. It shows that the NMT engine preserves 81% of the contents in the source text. The distribution of the score per volunteer is presented in Table 4.5 and Figure 4.2. Note that about half of the translations (50.8%) are assessed as perfectly adequate to the source (of Scale 4). The total proportion of Scale 4 - Everything and Scare 3 - Most numbers 78.3%, which is slightly lower than the case of the fluency score (80.6%). The proportion of the segments that contain irrelevant contents (Scale 2) or no contents at all (Scale 1) is 6.1%.

Figure 4.2 exhibits a resembling tendency of the performance of the engine to Figure 4.1; the graph is inversely proportional to the score as shown in the fluency evaluation. The most favorable evaluator is V1, with 86.7% of positive scores (Scale 4 - Everything and Scale 3 - Most) and 64.3% of the highest score (Scale 4). In the meantime, V4 carries out the most captious judgment with the highest score occupying only 36.7% and the positive scores numbering 71.4%.

It can be inferred from the fluency and adequacy scoring evaluation that the outputs of the given NMT engine are quite linguistically fluent in Korean (the target language) and convey most of the contents in Spanish (the source text).



	Everything-4	Most-3	Little-2	None-1
V 1	63	22	8	5
V 2	47	32	10	9
V 3	53	20	18	7
V 4	36	34	25	3
Avg	***49.75***	***27***	***15.25***	***6***

Figure 4.2: Distribution of adequacy scores per volunteer and their average.
Table 4.5: Distribution of adequacy scores per volunteer and their average.

4.4.3. Segment Ranking

The data acquired from this evaluation is approached in three ways: selected as the best engine, distribution of rankings by MT engine and distribution of rankings by ranking choice.

Selected as the Best Engine The result of the ranking comparison is displayed in Figure 4.3 in a pie chart. The human translation occupies the biggest proportion with 44.6%, followed by Kakao i (30.2%) and Google Translate (25.2%). From this comparative study, it can be inferred that human translation is 20% more favored by the participants against Google Translate in the *es-ko* translation (and including all circumstances). Even though GT is the

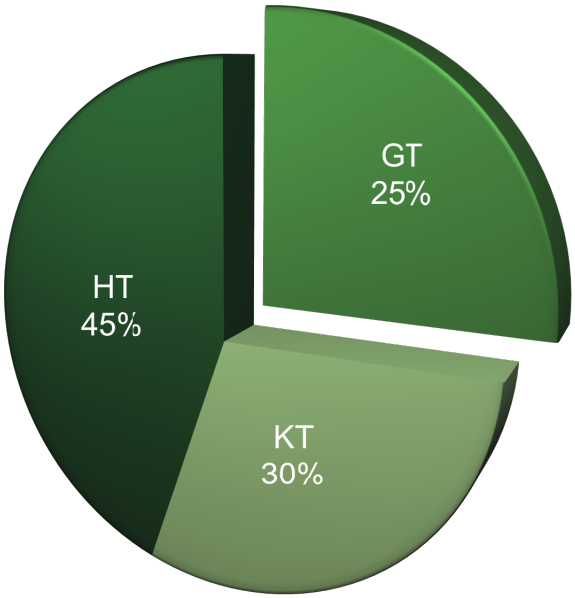


Figure 4.3: Proportion of selected-as-the-best engine of HT, GT, and KT. (GT = system translation of Google Translate, KT = system translation of Kakao i, HT = human translation).

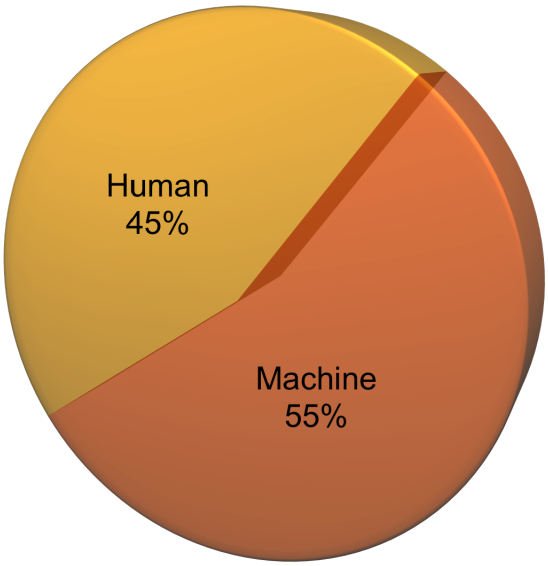
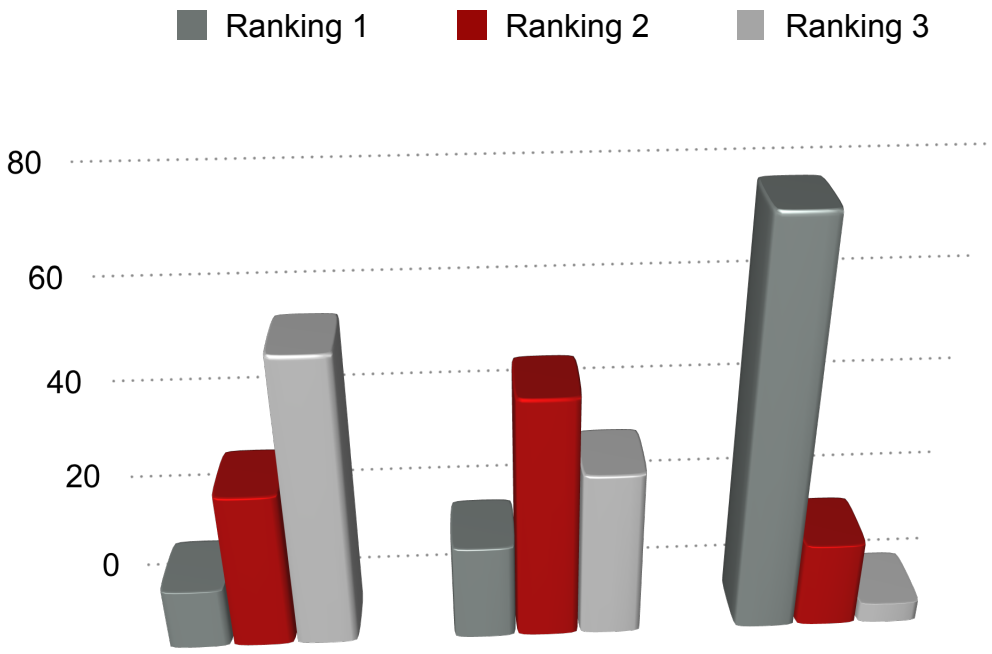


Figure 4.4: Proportion of selected-as-the-best engine of human vs. machine.

least favored translation, such a small gap between HT and GT shows a positive side of GT. In a comparison of human versus machine translation, more system translations are selected as the best engine than human translation by 10.8% points (see Figure 4.4). It can be that in spite that Google is ranked the lowest of the three, the system translation is about as good as human translation or that the evaluators somehow conceive that the system translation is a *better* form of translation. In any sense, further analysis should be carried out.

Rankings By MT Engine Table 4.6 and Figure 4.5 showed the detailed percentage of the distribution of rankings by MT engine. About 57% of the outputs of GT were ranked as the third, and most of the upper segments were marked lower than the rest of the two candidates. The proportion of the rankings in GT was noticeably leaning towards the lowest rank. The noticeable point was that there were segments of the system translation that were equal to or outperformed the human translation (Ranking 1): 12% in the case of GT and 19% in the case of KT. Moreover, what was curious was that 4% of HT were



	GT	KT	HT
Ranking 1	12	19	**79**
Ranking 2	31	**48**	17
Ranking 3	**56.99**	33	4

Figure 4.5: Average distribution of rankings (of four people) by MT engine (unit: %).

Table 4.6: Average distribution of rankings (of four people) by MT engine (unit: %).

given Ranking 3. It could be interpreted as either mistranslation or mistakes by the evaluators, but further investigation was required in this regard.

As such, one segment was selected from an anonymous volunteer who ranked HT as Ranking 3. It is shown below to question into why HT got the lowest rank, from a qualitative viewpoint, as well as a perspective of the machine.[37] The first point to take into consideration was that none of the Rank 1 and Rank 2 translated 'Let's start (*Comenzamos*)' while the human translation did (시작하겠습니다). It was noted that the two system translations translated 'the crisis (*la crisis*)' as if it was 'the crisis of Guinea, Sierra Leone and Liberia' while HT translated it directly to 'Ebola (에볼라로)' because the crisis in this context referred to the Ebola virus. Moreover, in the part where the literal translation of the main subject and verb was 'information has emerged (*han surgido*

[37] The reason we put an emphasis on the standpoint of a machine comes from the hypothesis that considering the nature of NMT, the error analysis of NMT has more sense when human strives to follow the trace of its footprints in its own way. It also helps to objectify the evaluation.

ST	Comenzamos, en los últimos días han surgido informaciones sobre una supuesta reducción de la epidemia del ébola en los países, en más castigados por la crisis Guinea, Sierra Leona y Liberia.	
KT	우리는 지난 며칠 동안 기니, 시에라 리온, 라이베리아 위기의 영향을 가장 많이받은 나라에서 에볼라 전염병이 줄어들 것으로 추정되는 정보를 얻었습니다.	*Rank 1*
GT	우리는 최근 기니, 시에라리온, 라이베리아 위기로 인해 더 많은 처벌을 받고 있는 국가의 에볼라 전염병 감소에 대한보고를 시작했습니다.	*Rank 2*
HT	시작하겠습니다. 최근 에볼라로 가장 피해가 컸던 기니, 시에라레온, 라이베리아에 에볼라가 감소하고 있다는 소식을 입수했습니다.	*Rank 3*

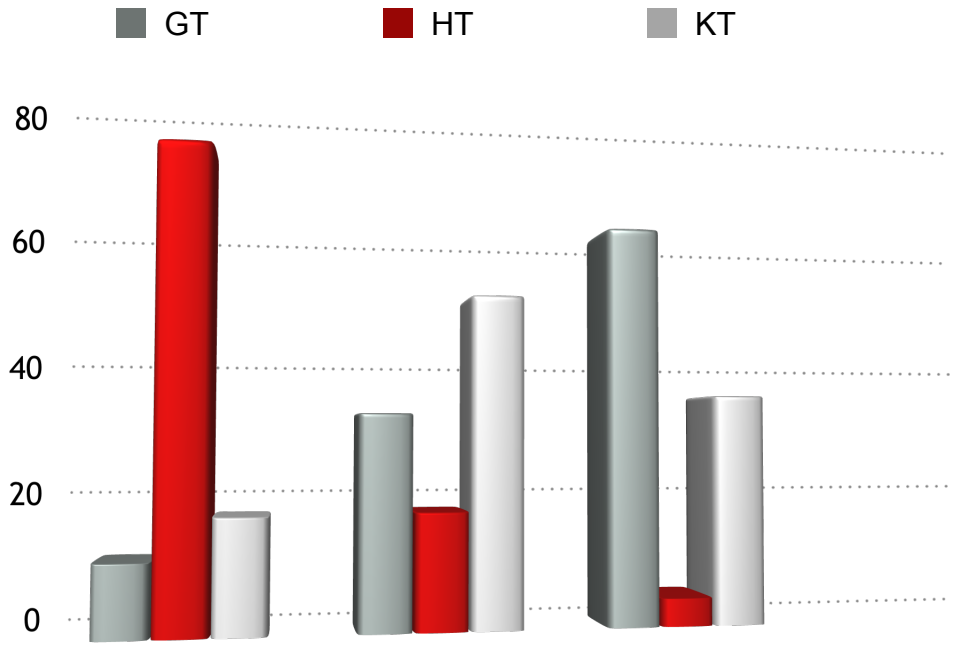
informaciones)', the translation of KT was 'we got the information (정보를 얻었습니다)' and that of HT was 'we got the news (소식을 입수했습니다)' while GT was 'we started to report on (보고를 시작했습니다)'. The translations of KT and HT made more sense than that of GT on the grounds that the source text stated that the information was acquired from an unrevealed reference while the translation of GT implied that they (in the context) reported the news.

From brief error analysis[38], it was detected that KT committed two errors (omission = 1, mistranslation = 1), GT had three errors [39] (omission = 1, mistranslation = 2) and HT had one error (omission = 1). The specifics of their errors were not going to be discussed, but the following analysis showed a possibility that the judgment of the evaluators could be biased or mistaken.
The appropriate ranking of the three systems, according to such analysis, should be HT - Rank 1, KT - Rank 2 and GT - Rank 3.

Rankings By Ranking Choice Table 4.7 and Figure 4.6 showed the distribution of rankings by ranking choice. HT obtained about 72% in Ranking 1 while 28% belonged to the system translations. In other words, the NMT engine translated almost 30% of the segments similar to or equal to a human translation in this pilot study. The largest proportion of Ranking 3 was occupied by GT with 60.63%.

[38] The discussion of an error taxonomy is exempt because the purpose of the error analysis in this pilot study is to justify that HT has fewer errors.

[39] Besides those, another spacing error has been detected in one part where two words 'about report (대한보고를)'are attached. Considering it or not, it is evident that Google Translate should be situated in the third ranking. The spacing error, however, is not discussed as the guideline requires the evaluators not to consider it as an error.



	Ranking 1	Ranking 2	Ranking 3
GT	10.9	32.29	**60.63**
HT	**71.81**	17.7	4.25
KT	17.27	**50**	35.1

Figure 4.6: Distribution of rankings by ranking choice (unit: %).

Table 4.7: Distribution of rankings by ranking choice (unit: %).

4.4.4. Post-Editing Productivity & Effort

The result obtained from post-editing will be discussed in this section from two perspectives: post-editing productivity.

Post-Editing Productivity The total time taken for translation from scratch and post-editing per volunteer is given in Table 4.8. Every volunteer performed the post-editing task faster than translating from scratch. The time spent for post-editing ranged from 23 ~ 30.77 minutes while translation from scratch needed 34.28 ~ 41.24 minutes. In total, the volunteers spent on average 1 hour (maximum 66.6 minutes) to complete the two tasks. The P ratio (post-editing productivity ratio) shows that there was on average 43% productivity gain of post-editing against translation.

		V1	V2	V3	V4	***Avg***
PE	Time (h)	0.47	0.43	0.38	0.51	***0.4475***
	WPH	1,429	1,561	1,767	1,316	***1,518***
T	Time (h)	0.57	0.66	0.69	0.60	***0.63***
	WPH	1,499	1,294	1,238	1,424	***1,364***
Total time		1.04	1.03	1.07	1.11	***1.0625***
P ratio		1.22	1.54	1.79	1.17	***1.43***

Table 4.8: Time (unit: hours) and words per hour (WPH) of post-editing versus translation from scratch, along with the total time (hour) of the two tasks and *P* ratio.

In relation to the individual result of the four volunteers, post-editing was the most effective to V3, with 79% productivity gain. V3 spent the shortest time in post-editing while the longest time in translating. Post-editing was the least effective to V4, with 17% productivity gain, meaning that by post-editing he saved 5 minutes. Be it that post-editing was slightly more productive than translation in the statistics, the difference would mean in the real world that post-editing was not worth it, in his case. In any case, one noticeable point was that the time and WPH were consistent throughout the volunteers.

Post-Editing Efforts To shift the focus of attention to the relations between time and throughputs of post-editing and sentence length, a graphical description was given in Figure 4.7 and Figure 4.8. The figures additionally showed the case of translation from scratch along with post-editing. Figure 4.7 showed the correlation between time and sentence length and Figure 4.8 showed the correlation between throughput and sentence length. In relation to the time and sentence length in Figure 4.7, in a range of 1 to 48 words in the dataset, it was hard to conclude that there was a direct proportion between the two variables due to some inconsistent cases such as the cases of length 22, 31 and 44 in translation and 31 and 35 in post-editing. However, it was observed that the shorter sentences required less temporal efforts in comparison to the longer sentences in both post-editing and translation. In any sense, no clear tendency was witnessed as to the level of temporal efforts in regard to the sentence length.

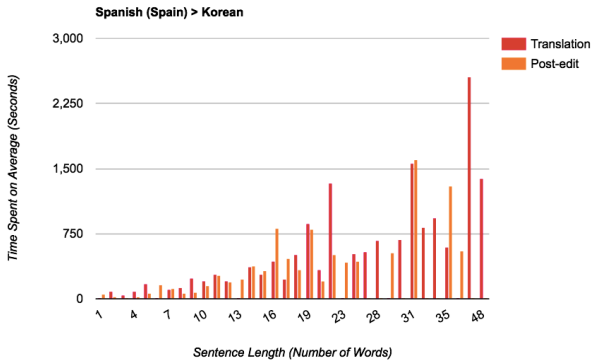


Figure 4.7: Average time spent for post-editing and translation from scratch by sentence length (unit: second).

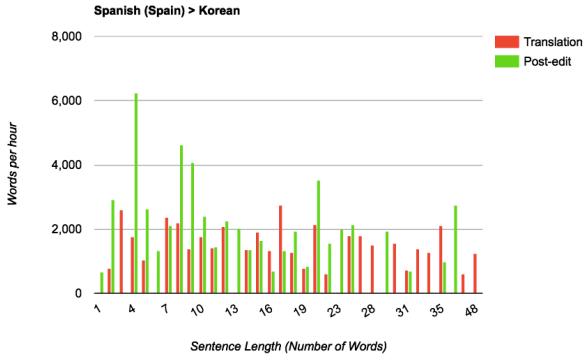


Figure 4.8: Average WPH of post-editing and translation from scratch by sentence length (unit: second).

In terms of WPH and sentence length, the correlation was much less clear than the former case, with considerable ups and downs (see Figure 4.8). It was crystal clear that the volunteers' throughputs did not increase or decrease in relation to the sentence length. In other words, they did not feel particularly easy when working on the shorter sentences. A further investigation would be necessary to reason out such phenomenon. A noticeable tendency, in fact, was witnessed from the comparison of WPH of post-editing and translation from scratch. The WPH of post-editing seemed much higher than those of translation from scratch in short sentences (up to length 10). More investigation was also required in this regard to suggest any speculations.

Additionally, the number of edits introduced by the volunteers were analyzed. The hypothesis was that more edits meant more human efforts when post-editing. Figure 4.9 gave an overview of the edit efforts across the half of the dataset, considering only the post-edited segments, based on an edit distance calculated by Levenshtein's algorithm normalized per segments (TAUS, 2010). This so-called edit distance score range from Distance 1 to Distance 10, where Distance 0 means no or hardly any edits (high similarity) were required while Distance 10 implied the sentence was almost translated from scratch. It turned out that 1/3 (31.6%) of the segments earned Distance 0. It implied that about 32% of the half of the dataset was perfectly

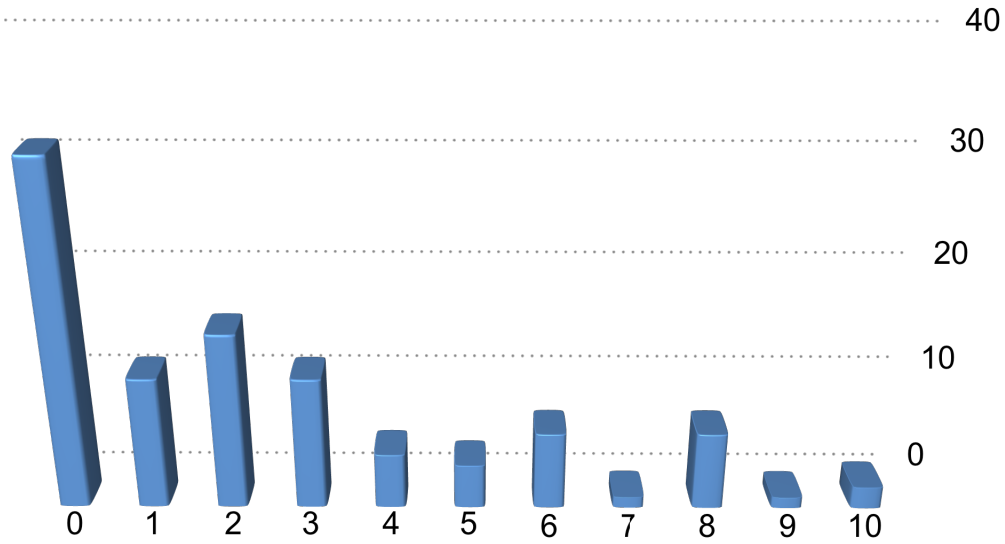


Figure 4.9: Edit distance across the dataset computed by Levenshtein's algorithm. (unit: %)

comprehensible to human. Furthermore, the top three scales —Distance 0, 1 and 2— accumulated 60.2% of the whole dataset. Such a result exhibited a positive aspect of GT.

To sum up, there was an on average 43% post-editing productivity gain over translation from scratch. No significant correlation was detected as to post-editing efforts (or translation effort) in relation to sentence length from the perspective of time and WPH, but the result left a room for discussion in regard to the low level of post-editing effort in shorter sentences. Lastly, the edit distance score showed that 32% of the post-edited segments did not require edits.

4.5. Contribution

The analysis of the result obtained from this pilot study delivered some of many interesting findings in relation to the performance of the NMT engine in the *es-ko* translation. While thesis findings would have to reconfirmed in a larger experiment setup, this pilot study proved its usefulness in this thesis. This section is dedicated to verifying it by comparing the objectives introduced in Chapter 4.1 and to estimate the feasibility of the main experiment.

- **Objective 1: to verify the properties of the actual experiment**

  > It was proven that all the variables served their roles.

- **Objective 2: to detect default variables that impair the setting**

  > No missing variables have been found.

- **Objective 3: to observe the volunteers' level of comprehension**

  > Many volunteers made questions about the definition of post-editing as they were a novice post-editor. They commented that more direct and specific examples would improve the guideline.

  > The on-site experimentation was useful in that the project manager could monitor the degree of their understanding and level her discourse to their perspective.

- **Objective 4: to test the performance of the workbench**

  > No error has been detected.

  - **Sub-Objective - Methodology 1: In fluency and adequacy scoring, the pilot study monitored the volunteers' comprehension of the dissimilar concept of the two methods.**

    > The comprehension of the volunteers got clearer as they were more engaged in the task.

  - **Sub-Objective - Methodology 2: In segment ranking, the pilot study monitored the volunteers' comprehension of the concept of an absolute ranking score.**

    > At first, some of the volunteers tried to perform the task with a relative scoring system, but after some feedback, they understood the concept. More stresses on the difference between the two systems were required in the training session.

  - **Sub-Objective - Methodology 3-1: In post-editing, the pilot study monitored post-edited outputs to check if the volunteers followed the guideline.**

    > The volunteers stated that they were not clear about the level of post-editing until they were engaged in the work and tried a couple of cases. More detailed guideline about the distinction between the light and full post-editing would be of help.

  - **Sub-Objective - Methodology 3-2: The pilot study confirmed if the sentences were randomly distributed to let the volunteers post-edit**

and translate.

> The randomization was confirmed.

- **Sub-Objective - Methodology 4: The pilot study detected hindrances to the working environment.**

  > The Internet connection was a significant factor in the result that could distract the volunteers.

- **Sub-Objective - Dataset: The pilot study examined the appropriacy of the domain and text type.**

  > The volunteers did not give a critical opinion about the domain. They, in fact, claimed that they felt the domain was suitable.

  > Some volunteers declared that the use of newswire as the domain reflect ed the real-world usage of Google Translate.

- **Sub-Objective - Workbench 1: The pilot study monitored the stability of the platform.**

  > The stability of the workbench was confirmed.

- **Sub-Objective - Workbench 2: The pilot study monitored the user interface of the workbench.**

  > All of the volunteers criticized the user interface of the workbench for provoking confusions on the job. The unclear classification of current, previous and next segments should be noticed in advance.

- **Sub-Objective - Training: The pilot study monitored the quality of the guideline.**

  > The guideline was helpful, but it could be improved by including the above-mentioned suggestions.

4.6. Chapter Summary

Chapter 4 reported on the result of the pilot study. In Chapter 4.1, we clearly stated that the main goal of the pilot study was to verify the parameters preset for the experiment. To that end, we specified sub-goals of each parameter from methodology to the dataset, workbench, and training session.

In Chapter 4.2, after briefly explaining the evaluation methods for the study —fluency & adequacy scoring, segment ranking, and post-editing—, we specified each parameter: the dataset, evaluators' profile, and their training. The dataset was composed of 1,517 words (98 sentences) in the newswire domain extracted from a book that contained a transcription of a Spanish radio script and its Korean translation. There were four evaluators in the pilot study who volunteered for the study. They were University students majoring in Spanish philologies. Three of them were females and one of them was a male. All of them were native Koreans and two of them were also native Hispanics. A face-to-face training session was carried out. We monitored the working environment and reaction of the volunteers towards the guideline and an interface of the workbench.

In Chapter 4.3, the process of the pilot study was stated in chronological order. The pilot study lasted for two days. On the first day, the volunteers were informed of the task and performed post-editing. For the first five sentences, we got instant feedback from the volunteers to enhance their understanding. On the second day, they performed the fluency & adequacy scoring and segment ranking. We could observe that the comprehension about the texts was greatly enhanced on the second day after performing post-editing on the same dataset in advance.

In Chapter 4.4, we reported on the result of the experiment. The NMT engine obtained 84.75% fluency level and 80.75% adequacy level. In segment ranking against an NMT engine of Kakao i and human translation, the volunteers selected GT as the best engine in 25% of the segments. In post-editing, there was 43% productivity gain over translation from scratch. Moreover, 32% of the segments did not require any edit efforts. The result analysis was limitedly provided as the main purpose of the pilot study was to monitor the parameters of the experiment.

In Chapter 4.5, we compared each and every goal we had for the pilot study and concluded that all goals were successfully met.

[blank page]

5. EXPERIMENT SETUP & CONDUCT

The experiment of the current thesis is expected to assess the performance of an NMT engine by means of the following four types of human evaluation methods:

- Fluency Scoring
- Adequacy Scoring
- Segment Ranking
- Post-Editing

The NMT engine is represented in this experiment by Google Translate of version 2018, and the performance of the engine is tested in the *es-ko* pair in newswire domain. While all evaluation is based on the translation output of the Google Translate, the segment ranking will contrast the given engine to Kakao i and human translation.

This chapter covers a preliminary study of the experiment from Chapter 5.1 to Chapter 5.3 and experiment conduct in Chapter 5.4. Chapter 5.1 gives information about the dataset of the experiment. Chapter 5.2 is dedicated to evaluators, or also called post-editors, detailing a recruitment process and their profiles. Chapter 5.3 introduces the workbench of the experiment, TAUS DQF. Having clarified the experiment design, Chapter 5.4 reports on the procedure of the experiment chronologically starting from a training session that is composed of a distribution of a guideline and a warm-up session.

5.1. Dataset

The domain of the experiment is newswire. Newswire is one of the most popular domains along with Europarl in WMT. The main reason for the selection of newswire is that it is a general-purposed text. It is in line with the objective of the thesis to assess the general performance of the system that is publicly available and free. It is believed that the purpose of a text would have an influence on the performance, and so a more general text is apt for the current experiment.

The dataset consists of a mixture of 11 articles concerning political issues in general and an election in particular. The articles are selected from three different Spanish and Korean newspapers although the nature of newspapers does not have much impact on the result. The main curiosity from the different nationality of the journals is how the engine will process out-of-vocabularies (OOVs) contained in diverse contents on a global and national level (of both Spain and Korea). These three major sources are ABC[40], El País[41] and KBS World Radio[42], two of the most popular newspapers in Spain and the Korean journal that publishes cultural, social and political issues in South Korea in 11 languages including Spanish. The details of the contents of the dataset are demonstrated in Table 5.1.

Topic	Date	Size (words)
Election in Adalucía	Nov 15, 2018	1,107
Brexit	Nov 12, 2018	513
Election in Florida	Nov 10, 2018	749
General election and Trump	Nov 4, 2018	836
Election in Ukraine	Nov 11, 2018	610
Mission 2019 of EU	Dec 29, 2018*	1,710
Election in the USA	Nov 6, 2018	173
Violation of election law	Sep 29, 2018	127
Low support for President Moon	Aug 16, 2018	197
Result of regional elections	Jun 14, 2018	212
Electoral recount	Jun 13, 2018	181
		6,415

Table 5.1: Topics of 11 newspaper articles. The asterisk mark * refers to the last-updated date. The date was obtained for the experiment in November.

The principal topic of the articles is elections held in various regions

[40] https://www.abc.es

[41] https://elpais.com

[42] http://world.kbs.co.kr/service/index.htm?lang=s

including Andalucía of Spain, the United Kingdom and Florida of the USA and general elections in the USA, Ukraine, European Union, and Korea. Note that the number of articles related to Korean issues is more than other articles due to a relatively small size of the words of each article. There is an imbalance of the size of the data per article and per journal, but as clarified in advance, it does not influence the quality of outputs considering the nature of NMT.

The dataset is pre-edited by the author to a minimum extent by correcting typos and restructuring each sentence on a sentence level to match them to a format required by the workbench. For the convenience of the evaluators, headlines and sub-headlines are appended with a title of '[Headline]' and '[Sub-headline]', but they are instructed not to translate them. While the total number of dataset is displayed as 6,415 words in Table 5.1, it does not include the headlines and sub-headlines. The total size of the dataset is presented in Table 5.2, along with a sentence length. The size of the dataset of this thesis is a total of 6,426 words or 253 sentences, and the sentence length of the source text ranges from 3 to 82 while the average sentence length is 32.08.

Number of Sentences		253
Number of Words	ST	6,426
	GT	4,277
	KT	3,916
	RT	3,816
Minimum Sentence Length		3
Maximum Sentence Length		82
Average Sentence Length		32.08

Table 5.2: Information about the size of the dataset of the experiment. ST, GT, KT, and RT stand for a source text, Google translation, Kakao translation, and reference translation (human), respectively.

It has been detected from Table 5.2 that the size of the three translations (GT-4277, KT-3916, RT-3816) is much shorter than that of the source text (6,426). In the case of RT, it is 1.68 times shorter than the original text. In the case of the system translation that involves GT and KT, the sizes are 1.5 times and 1.64 times shorter than the original text. Such a phenomenon can be reasoned out in part by a different calculation rule of the two languages. While

Spanish is space-based, in Korean a spacing does not necessarily mean that a word delivers one grammatical unit because Korean is an agglutinative language. In fact, a Korean POS tagging represents such dissimilar aspect of Korean in a POS tag set[43]. For example, there is a subject postposition (_JKS) in the tag set that is attached to the subject in the following way:

ko 바람이 분다.
en Wind blows.

In the given sentence, _JKS refers to the colored part of the Korean sentence which forms one word together with the subject “wind” and does not have equivalence in the English sentence. As such, the POS tagging of these two sentences are as follows[44]:

ko 바람-이 분-다.
NNG JKS VV EF
en Wind blows.
NN VBZ

It clearly shows that although the two sentences are composed of two words, the Korean sentence has four POSs. Returning to the size of the corpus, the two languages, thus, cannot be contrasted just by the number of words.

RT of the dataset is prepared by a professional translator. He has been an active translator for the *es-ko* translation for more than seven years, a native speaker of Korean (the target language), and has experience in translation in the political domain. Together with RT, GT and KT are obtained in November 2018.

5.2. Evaluators

There are no internationally adopted standards as to what a pertinent profile of a judge in either MT evaluation or post-editing is. In MT evaluation, the most required ability is knowledge in a target language while knowledge in a source language is optional. In post-editing, the discussions are much more fierce.

[43] http://incredible.ai/nlp/2016/12/28/NLP/

[44] The Brown tag set is applied to the English sentence.

Some argue that a professional translator with many years of experience performs better as a post-editor (de Almeida and O'Brien, 2010). On the other hand, others claim that lay users of MT that have no experience in translation practice post-editing more efficiently without making stylistic changes and sticking to a guideline (Aranberri et al., 2014; Mitchell, 2015). More important factor than being a translator, according to them, is previous knowledge in the domain. It is, therefore, concluded that as the profile of MT evaluators and post-editors is not yet standardized, flexibility in the criteria of selection of the participants is allowed per study as long as it is not against common sense. Likewise, concerning the number of participants, more engagement assures the legitimacy of the experiment in theory. An effort has been made to hire as many participants as possible as long as the financial condition allows. The total number of evaluators for the experiment is six, and they participate in both MT evaluation and post-editing. Their details are provided in the following chapters, but for the sake of their anonymity, specific information is left out.

5.2.1. Recruitment

It is thought that a validate candidate for this experiment is who satisfies the following conditions:

- A Korean native speaker
- A professional translator for Spanish and Korean
- Who is currently active in the translation market
- Who passes a selection test

In relation to proficiency in a target language, Sánchez-Gijón and Torres-Hostench (2014) claim that a post-editor does not need to be a native in a target language for a good-enough post-editing. In this thesis, as the level of post-editing is "similar or equal to human translation", the top priority for the candidate is their mother tongue in Korean. In addition to such quality, this thesis requires that the candidate be a currently-working professional translator for the *es-ko* pair regardless of years of experience.

The recruitment was based on online platforms through diverse channels.

Firstly, recruiting platforms such as oDesk work[45] and Upwork[46] were approached. These websites connected employers and employees for translation worldwide. The employers could post a description of the project and those who were interested in the project could contact the employer by submitting their résumé along with his/her activity history in the given website. Albeit competent, most of the applicants were translators for the *en-ko* pair who had knowledge in Spanish or who were currently based in Hispanic countries. There was one applicant who satisfied all suggested qualification, but she denied to take the selection test.

Having seen that freelance translators in such platforms were mostly *en-ko* translators, the scope has been narrowed down to graduate schools of translation in Korea (that train professional translators) or online Korean communities in Spain. Brief information about this project was posted on the board of each channel, and over 30 applicants showed their interest in this project. For those who met the first three qualifications, a selection test was prepared to ensure that they were qualified not only in paper but also in practice.

To this aim, a small part of an article about an election in the newswire domain was extracted. The detailed description of the text is provided in Table 5.3. The text was composed of 163 words and belonged to the identical topic and domain to the dataset of the experiment. The selection test was to translate this text.

Newspaper	Topic	Date	Size (word)
El País	Elections in Brazil	Nov 19, 2018	163 (7 sentences)

Table 5.3: Description of the text of the selection test.

The reference translation was prepared by the author in advance for an objective evaluation. The selection criteria of this test were to:

- Confirm their linguistic level of Spanish and Korean
- Confirm their knowledge in the given topic

[45] https://www.odeskwork.com/

[46] https://www.upwork.com/

- Their devotion to the task

5.2.2. Profile

After the selection test, a total of six applicants were invited to this experiment. However, during the experiment two of them gave up the task for financial reason and the level of intensity of the task. Two more evaluators joined the project in the middle of the experiment, but they all went through the same process that the existing evaluators experienced from training to completing the task.

	Educational Background	Relevance of the major	Work Experience	Gender	Mother Tongue	Test Score*
EV1	BA	unrelated	3	female	Korean	High
EV2	Ph.D.	related	5	female	Korean, Spanish	Medium
EV3	MA	unrelated	5	female	Korean	Low
EV4	BA	related	2	female	Korean	Medium
EV5	Ph.D.	unrelated	1	female	Korean	Low
EV6	BA	unrelated	4	female	Korean	High

Table 5.4: Profile of the six evaluators. *The score of the selection test was calculated based on the subjective standard of the organizer. The three levels (high, medium, low) are relative scores among the six people, in such a way that those who awarded a low rank still outperformed other unselected applicants.

The detailed information about the final six evaluators are given in Table 5.4, but for their anonymity, their private information is not provided. For convenience sake, they are named as EV1, EV2, EV3, EV4, EV5, and EV6. Their educational background varies from BA to Ph.D. in translation studies or other fields of study. The work experience in the *es-ko* translation ranges from 1 to 5 years. Note that the relative test score is provided in the last column for

readers' information. These profile features show that each evaluator has her strengths and weaknesses and that each feature does not interfere with one another. For example, those who earned a higher score in the selection test (EV1 and EV6) did not study translation as a major.

After getting verbal confirmation from each participant, written consent was requested for all participants, officially acknowledging that they participated in the project and agreed to provide information for research purpose only. Having reconfirmed their agreement on the suggested payment, they were assumed to be ready for the project.

5.3. Workbench: TAUS DQF

The main workbench of this experiment is TAUS DQF[47]. It offers an incomparable experience for MT evaluation of adequacy scoring, fluency scoring, segment ranking and post-editing with features adaptable to the design of the experiment. This chapter describes why this TAUS DQF is selected as the main workbench (Chapter 5.3.1), how the project is organized (Chapter 5.3.2) and

5.3.1. Why TAUS DQF?

TAUS, established in 2004, is a "language data network where users can share knowledge and data of translation" through their online evaluation platforms of TAUS Dynamic Quality Framework (DQF) and TAUS Dashboard and offline conferences. While there are some beneficial software for the NMT evaluation task such as PET (Aziz et al., 2012), the current study adopts a web-based workbench at TAUS for the sake of its comprehensiveness, user-friendliness, and flexibility. The TAUS workbench is comprehensive in that it provides a complete package of human evaluation methods including fluency scoring, adequacy scoring, segment ranking and post-editing. It also facilitates thorough analysis of the result along with TER scores. Optionally, error analysis is available, as well. In addition to the comprehensive service it offers, the platform is very user-friendly in that users can access to the workbench without installing the software. This is especially helpful to those who do not have technical skills.

[47] https://www.taus.net/

As such, any previous knowledge about programming is required for both the organizer and the evaluator. This workbench is also very flexible when organizing an experiment, tuning every feature suitable for the circumstances of the experiment, i.e. a language pair, domain, level of post-editing, and use of translation memory in the process.

5.3.2. Features of Workbench

It is very simple to create a project on DQF. Firstly, specific information about the project is set up: project name, project type, company name, content type, industry type, the source language, and the target language. The project name and company name are a blank space where one can type in his/her own information. These variables do not affect to the interface of the workbench. The most important category is the project type where one chooses one of the three types of evaluation — Comparison (segment ranking), Productivity (post-editing) and Quality Evaluation (fluency and adequacy scoring and error analysis). Upon selection of the project type, new options are provided: For Comparison, one can choose between a quick comparison and a ranking comparison. The quick comparison asks evaluators to choose the best translation among (up to) three candidate translations while the ranking comparison let them award a rank of each translation in relation to one another. For Productivity, a level of post-editing quality should be set between ‘good enough’ quality and ‘similar or equal to human translation’ quality. The notion of these two qualities is equivalent to the conventional definition of *light* post-editing and *full* post-editing, respectively. The exact definition of the two levels of post-editing is extracted from Görög (2014) and given below:

- **Light Post-Editing or Good-Enough Quality:** It is defined as “comprehensible (i.e. you can understand the main content of the message), accurate (i.e. it communicates the same meaning as the source text), but as not being stylistically compelling. The text may sound like it was generated by a computer, syntax might be somewhat unusual, grammar may not be perfect but the message is accurate. Aim for semantically correct translation. Ensure that no

information has been accidentally added or omitted. Edit any offensive, inappropriate or culturally unacceptable content. Use as much of the raw MT output as possible. Basic rules regarding spelling apply. No need to implement corrections that are of a stylistic nature only. No need to restructure sentences solely to improve the natural flow of the text." (Görög, 2014)

- **Full Post-Editing or Similar-or-Equal-to-Human-Translation Quality:** It is generally defined as "being comprehensible (i.e. an end user perfectly understands the content of the message), accurate (i.e. it communicates the same meaning as the source text), stylistically fine, though the style may not be as good as that achieved by a native-speaker human translator. The syntax is normal, grammar and punctuation are correct. Aim for grammatically, syntactically and semantically correct translation. Ensure that key terminology is correctly translated and that untranslated terms belong to the client's list of *Do Not Translate* terms. Ensure that no information has been accidentally added or omitted. Edit any offensive, inappropriate or culturally unacceptable content. Use as much of the raw MT output as possible. Basic rules regarding spelling, punctuation, and hyphenation apply. Ensure that formatting is correct." (Görög, 2014)

After the level of post-editing is decided, one of the five types of post-editing evaluation presented in Table 5.5 should be selected. For Quality Evaluation, three evaluation types —fluency, adequacy and typology errors— are suggested to choose which type will be included in the evaluation. Multiple options are allowed.

The content type includes 15 domain of texts such as legal text, user material, audiovisual content, etc. Likewise, the industry type requires to choose among various sectors the organizing field of study. In terms of languages, one can select one source language and multiple target languages upon necessity.

Having set the features of the experiment, the dataset should be

Procedure Type	Description	Measurement
Human	human translation	time spent on translation
MT+PE	post-editing on a system translation	time of editing, edit distance
TM+Human	human translation with the help of translation memory	time spent on translation
MT+PE+Human	comparison of post-editing and translation from scratch	post-editing time, human translation time, edit distance
MT+PE+TM+Human	comparison of post-editing and translation from scratch with the help of translation memory	post-editing time, human translation time, edit distance

Table 5.5: Five types of post-editing evaluation in DQF.[48]

uploaded in the platform in a format that DQF requires. The standard size of the dataset ranges from 250 segments to 5,000 segments. The source text and system translation are uploaded in Productivity and Quality Evaluation evaluation, and up to two more translations can be added to the existing file in Comparison evaluation. By putting evaluators' email address, the project gets ready. One can monitor the progress of the task of each evaluator in the platform and induce them to accelerate the speed when it is needed.

5.3.3. Technical Manual

Once the project is prepared, the evaluators will receive a link in their email through which they can access the DQF workbench. Along with the link, the email contains information about a language pair, a direction of translation, a name of the organizer and a brief manual of the workbench. Upon pressing the link, they access the workbench, and in the case of post-editing, the time is calculated. And so, it is hightailed in the guideline that they can always pause the task anytime they want by pressing a 'pause' button and that they should do when they do not concentrate on the task. They can continue the work by accessing the tool through the same link, and it is connected to the last

[48] The description is referred to as the following website: https://dqf.taus.net/projects?display=0&page=0.

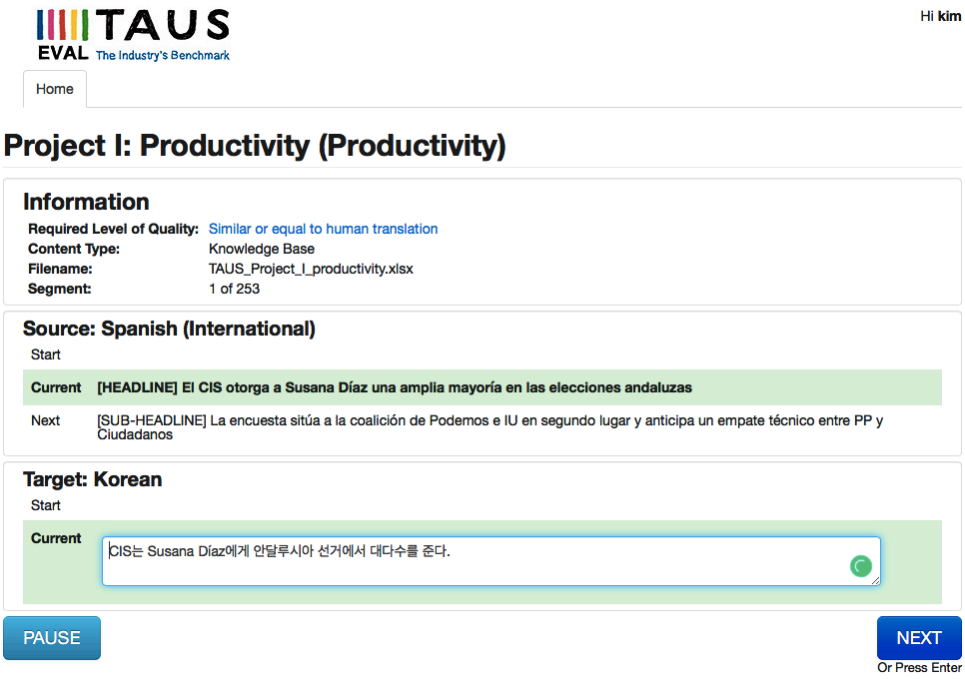


Figure 5.1: Interface of the post-editing evaluation.

segment they have paused. Not only the organizer but also the evaluators could monitor the progress of the task in the workbench. Upon completing the work, they receive an email that confirms the end of the work.

5.3.4. Interface of Workbench

This chapter provides concrete images of the interface of the TAUS DQF workbench. In the experiment, there are three types of human evaluation. The interface of each evaluation task will be shown.

In Figure 5.1, the interface of the workbench for post-editing is displayed. In a work process of MT+PE+Human, the current blank is provided to be empty when the evaluators translate a segment from scratch and the current blank is filled with the system translation when they post-edit. On top of the source segments, general information of the task is described, but again they do not have an influence on the task. The post-editors can refer to the previous and next segment along with the current segment, which facilitates better comprehension of the context. The time taken for the task is automatically measured once the post-editors enter the interface.

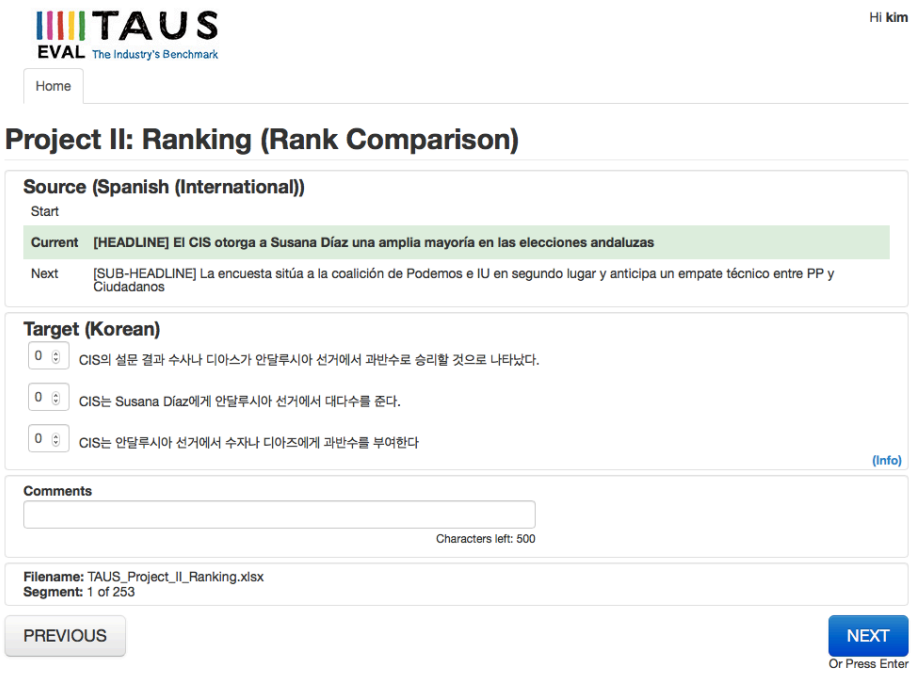


Figure 5.2: Interface of the segment ranking evaluation.

In Figure 5.2, the interface of the segment ranking evaluation is displayed. Unlike some studies that provide reference translation in the interface so that the evaluators can refer to, the current workbench does not offer such service. It is believed to be acceptable as the evaluators have previous knowledge over the dataset after post-editing in Project I. The overall working environment is identical with the previous, current and next segments exhibited on the screen. The only difference is that three hypothetical translations are randomly given without information on their translators. The order of the translations always changes so that the evaluators do not have a pre-judgment over the segments.

Figure 5.3 shows the interface of Quality Evaluation of fluency and adequacy scoring. Similar to the previous interfaces, the previous, current and next segments are displayed, with a special highlight on the current segment in green. When the evaluators make a judgment, they can always refer to the standards of the scale given in *More Info.* They can also leave comments on every segment when it is needed. The progress of the task is presented at the end of the line.

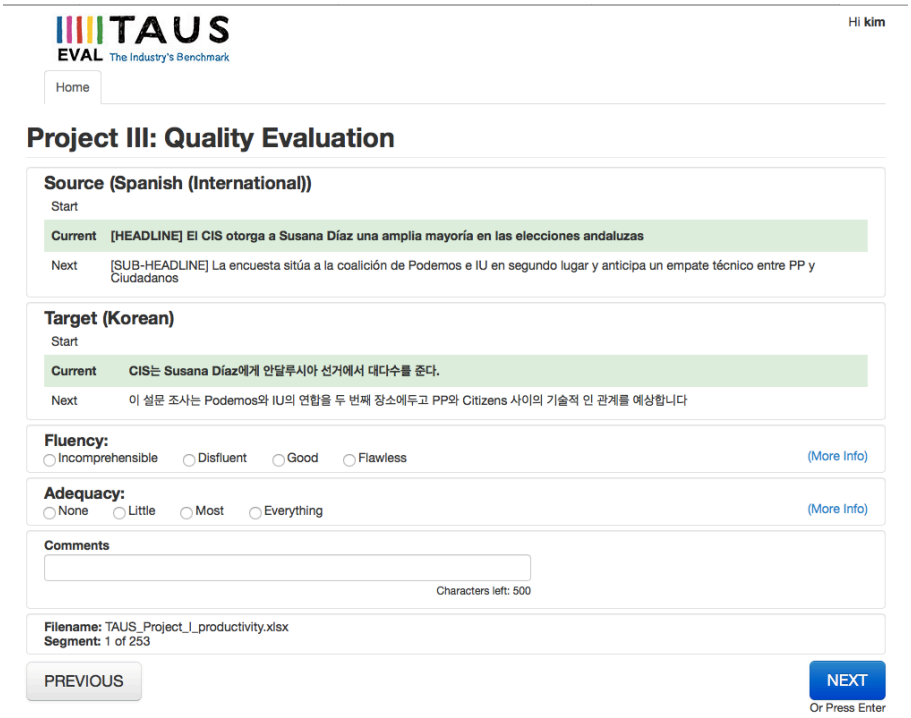


Figure 5.3: Interface of the fluency and adequacy scoring evaluation.

5.4. Experiment Conduct

This chapter describes how the experiment is prepared, from organizing the workbench (Chapter 5.4.1) and the project schedule (Chapter 5.4.2) to holding a training session before the experiment of distributing a guideline and having a warm-up session (Chapter 5.4.3).

5.4.1. Organizing the Project

The general category selection for the experiment of the current thesis is as follows:

- Company Name: UPF
- Content Type: Knowledge Base
- Industry Type: Undefined Sector
- Source Language: Spanish (International)
- Target Language: Korean

And the category selections of each evaluation type are as follows:

- Project Name: Project I: Productivity

- Project Type: Productivity - Similar or equal to human translation
- Process: MT+PE+Human

- Project Name: Project II: Ranking
- Project Type: Ranking Comparison

- Project Name: Project III: Quality Evaluation
- Project Type: Quality Evaluation - Fluency, Adequacy

The three types of evaluation were named as Project I, II and III following the order of evaluation conduct. It was important that post-editing was performed first so that the evaluators had no previous contact with the dataset. It was believed that for Ranking and Quality Evaluation the previous knowledge in the dataset will be of great help to them.

The type of content and industry were set to be *knowledge base* and *undefined sector* as there was no appropriate option that satisfied the condition of this experiment. Again, they were of no importance, at all events, to the result of the experiment because they existed for the organizer's information. In the source language, International Spanish was selected among some specified options of Spanish of Spain, Mexico, USA and so on, considering the different base of the newspapers.

In Project I, the post-editing process was set to be MT+PE+Human where the evaluators translated half of the dataset from scratch and post-edited the rest half of them that were previously translated by Google Translate. The level of post-editing was similar to human translation.

The source text and system translation of Google were uploaded in Project I and III, and the system translation of Kakao i and reference translation were added to the existing file in Project II. Having sent an email to each evaluator, the project was assumed to be ready.

5.4.2. Schedule

The project was scheduled for two weeks from November 29 to December 14 of 2018. The experiment was divided into a preparation stage (for training), Project

I, Project II and feedback session. The first two days were spent on distributing a guideline of post-editing and holding a warm-up session for post-editing. For the first week from December 1 to 7, the translators performed post-editing freely without time constraint. As post-editing was the key evaluation method for this thesis, it was categorized as Project I, and the rest two types of evaluation were performed altogether on the upcoming week upon completion of the first project. Before Project II, a guideline for Project II was delivered to the evaluators. Having confirmed that they completed the project, the evaluators were asked to give feedback in the form of a survey. In a period of seven days from the completion of all the tasks, they were offered a remuneration. All the communications between the project manager and the participants of the experiment were made online via emails on an immediate basis.

Upon selection of the six translators, the schedule of the project was sent to each of them to get an agreed consensus on the conduct of the experiment. One of the participants required to complete the tasks earlier than it was planned due to her schedule, and it was allowed as long as the tasks were carefully carried out. They were made clear that the information of the present project would be used for a doctoral dissertation, and signed written consent. Some of the translators lived in Korean while others in Spain, so the time difference was also considered for the deadline.

5.4.3. Training Session

The training session was held on November 29. The main purpose of this session was to deliver detailed information of the project, distribute the guideline of post-editing and let the evaluators be familiar with the post-editing as they are new to it. After describing the guidelines of the experiment in Chapter 5.4.3.1, how the evaluators understood the guidelines (Chapter 5.4.3.2) and how they are trained for post-editing (Chapter 5.4.3.3) will be explained.

5.4.3.1. Project Guideline

Based on the in-depth research on MT evaluation methodology in Chapter 3.4 and a theoretical framework in post-editing in Chapter 3.5, two guidelines for

the MT evaluation and post-editing were elaborated. Considering that the group of evaluators were native Koreans, it was believed that a guideline that was written in Korean would achieve a better understanding from them, but an English version was also prepared. Each evaluator was supposed to receive two guidelines; one for post-editing and the other for MT evaluation with some time interval. Each guideline is described subsequently.

Guideline of Post-Editing As post-editing was the first task of the evaluators, a complete procedure of the project was described beforehand to give them a whole picture. It included the objective of each evaluation task and what they were required to do in the three tasks. It also detailed information about the source language, target language, text type, domain, topic of the text, size of the dataset, workbench, and the number of evaluators. Having clarified the whole procedure of the project, a guideline for post-editing was explained. It mainly stated the definition of post-editing and level of post-editing. The explanation was as detailed as possible, considering that all evaluators did not have any experience in terms of post-editing. Considering some insights found in Aranberri (2017) that professional translators who were new to post-editing commented that some of the examples of potential errors in the guideline would be of great help to understand the task, each post-editing rule was provided with practical examples.

Basic Post-Editing Rules

1. Edit while maintaining the translation as much as possible.
2. Do minimum edits, but do not think deeply for the sake of making the best choice.
3. Do not edit for a stylistic reason.
4. Do not spend time on an extensive web search.

The four basic rules were provided as such, but sub-guidelines tailored especially for Korean and the purpose of this thesis were also given with some examples. The itemized examples are extracted from some of the most frequent errors found in the pilot study. They are as follows:

- Do not edit spacing. It is not of our interest to detect spacing, although

many spacing errors are observed. It is believed that they do not impede human comprehension of the text and that the correction will vary per editor as the spacing rule is complex even for Koreans. For example:

Item	안 녕	[Hel lo]
Desirable Correction	안 녕	[Hel lo]
Undesirable Correction	안녕	[Hello]

- Do not polish inconsistent terms as long as the translation is correct. In many Spanish texts, the same item tends to be expressed in various ways to avoid repetitive use of a word such as a job title while Korean texts prefer to stick to one identical expression. Although the correct translation, in theory, is to change them to one identical expression, the intention of this thesis is not to observe some translation skills of the engine. As such, an example of a correct post-editing is as follows:

Item	CEO, 사장, 대표 ...	[CEO, director, president …]
Desirable Correction	CEO, 사장, 대표 ...	[CEO, director, president …]
Undesirable Correction	CEO	[CEO]

- Edit untranslated words, except acronyms. An instance is a name such as the following case.

Item	Miguel Molleda
Desirable Correction	미겔 모예다
Undesirable Correction	Miguel Molleda

- Do not edit untranslated words that are inside brackets along with their translation, as long as they are erroneous or in other languages such as English.

Item	세계 보건기구 (WHO)	[World Health Organization (WHO)]
Desirable Correction	세계 보건기구 (WHO)	[World Health Organization (WHO)]
Undesirable Correction	세계 보건기구	[World Health Organization]

Advice For Post-Editing In addition to the guideline about the task, it was important to ensure that their working environment did not hinder the task in the process because time mattered for post-editing productivity. To yield an unbiased result, three recommendations were given:

- Prepare yourself for the work.
  -Check if you have a stable Internet connection.
  -Keep your computer plugged in so that you have enough battery.
  -Remove any distractions, i.e. smartphone, from you while working.
- You can pause whenever you want, but please do not pause too often or without pressing the pause button.
- Your concentration on the work is crucial for the result.

Guideline of MT Evaluation The guideline for the two evaluation tasks —fluency and adequacy scoring and segment ranking— was elaborated in one document. The core content of this guideline was to explain the purpose of the tasks and the procedure, to make sure that they understood the dissimilar concept of fluency and adequacy and to equip them with instrumental competence. With the instruction of each section, images of the interface were shown to let them familiar with the workbench.

5.4.3.2. Distribution of Guideline

Prior to Project I, a hand-out that contained the guideline of post-editing was delivered to the evaluators, and they were requested to go through it and ask questions. Those who understood the guideline and their role fully were supposed to reply back to confirm it. In this situation of online-based communication, a clear form of confirmation was crucial. EV1 asked in terms of translation from scratch if each segment should be translated within the context or be considered as an independent sentence; in terms of post-editing if the edit should be minimum as long as the meaning was conveyed or they could erase everything and translate the sentence; in terms of time if it was allowed to spend as much time as they could to make a perfect output. It seemed like the quality of post-editing and efficacy of the task were not crystal clear to her. In

the case of the rest of the editors, besides making sure the compensation of the project, they did not ask further questions and responded quickly. EV1 commented that she was interested in experiencing post-editing. After confirming that everyone was conscious of the objective, type of evaluation, process, task, as well as the notion of post-editing, post-editing rules, detailed instructions of post-editing, and TAUS workbench, the next session was launched.

5.4.3.3. Warm-up Session

Considering that the training was carried out at a long distance and that the translators did not have any experience in post-editing, a single distribution of the hand-out seemed to be insufficient to guarantee a reliable outcome. To enhance their comprehension, a warm-up session was scheduled on November 30. For their convenience, the same text that was used in the selection test was employed in this session. The detailed information of the dataset is provided in Table 5.3 in Chapter 5.2.1. With the dataset, a post-editing task (MT+PE+HT) was prepared on the workbench, and the post-editors were accessed to the platform via email. The procedure was identical to the main experiment as explained in Chapter 5.3.3 and 5.3.4. Upon completion, both of the parties received a confirming email.

As a final step, each of the participants got a PDF file where their inputs of post-editing and translating work were revised by the project organizer on a sentence basis. The image is displayed in Figure 5.1. Followed by the source text and system translation, their input texts were presented. The revision was made directly on each segment with red letters, and a desirable form of post-editing was given in a separate column. The calculation of post-editing and translation time and the number of edits were also provided for their information. Together with this feedback sheet, they were given a detailed correction sheet (see in Figure 5.2) where their inputs were corrected in detail to a desirable answer.

Editor: {name}

ID	Source Segment	MT Target Segment	Post-edited Target	Translated Target	바람직한 PE	Segment Length	Post-edit Time(ms)	Segment Origin	Edit Effort
1	Un nuevo ciclo democrático está a punto de iniciarse en Brasil.	새로운 민주주의주기가 브라질에서 시작될 예정입니다.	새로운 민주주의주기가 브라질에서 시작될 예정입니다.		새로운 민주주의시대가 브라질에서 시작될 예정입니다.	11	16180.0	El País	0
2	El pueblo eligió a sus representantes para el poder ejecutivo, el legislativo federal y el los de los distintos Estados para los próximos años.	사람들은 집행력, 연방 입법부 및 앞으로 다가오는 여러 국가의 대표를 선출했습니다.		국민들은 행정권, 연방 입법부 및 (missing) 앞으로 후년을 위한 여러 국가의 대표를 선출했습니다.	사람들은 행정부, 연방 입법부 및 각 주지사 등 앞으로 다가오는 임기의 대표를 선출했습니다.	24	10900.0	El País	22
3	Los brasileños acudieron tranquilamente a las urnas y participaron una vez más del rito de renovación de la democracia.	브라질 사람들은 침착하게 여론 조사에 참여했으며 다시 민주주의의 재건 의례에 참여했습니다.	브라질 사람들은 침착하게 여론 조사에 참여했으며 다시 민주주의의 재건 의례에 참여했습니다.		브라질 사람들은 침착하게 투표에 참여했으며 다시 또 민주주의의 재건 의례에 참여했습니다.	19	10350.0	El País	0
4	Gracias al sistema electrónico de votación, supimos el resultado de la elección presidencial en apenas dos horas y 18 minutos.	전자 투표 시스템 덕분에 우리는 불과 2 시간 18 분 만에 대선 결과를 배웠습니다.		전자 투표 시스템 덕분에 우리는 불과 2 시간 18 분 만에 대선 결과를 알게 되었습니다.	전자 투표 시스템 덕분에 우리는 불과 2 시간 18 분 만에 대선 결과를 알게되었습니다.	20	18230.0	El País	11
5	Eso no es poco en un país de las dimensiones continentales de Brasil, la cuarta mayor democracia del mundo, con sus 208,5 millones de habitantes y 147 millones de electores.	브라질의 대륙 차원의 인구가 2 억 8 천 5 백만명, 유권자 1 억 4 천 7 백만명으로 세계 4 위의 민주주의 국가에서는 충분하지 않습니다.	(missing) 인구가 2 억 8 천 5 백만명, 유권자 1 억 4 천 7 백만명으로 세계 4 위의 민주주의 국가인 브라질에서는 이것으론 충분하지 않습니다.		브라질의 대륙 크기의 국가 면적, 인구 2 억 8 천 5 백만명, 유권자 1 억 4 천 7 백만명, 세계 4 위의 민주주의 국가 등을 고려하면 이는 충분하지 않은 시간입니다.	30	23280.0	El País	27
6	Solo la competente y seria labor de las instituciones públicas nacionales destinadas a garantizar el régimen democrático permitieron llevar a cabo una votación tan tranquila.	민주적 정권을 보장하기위한 국가 공공 기관의 유능하고 진지한 노력 만이 그런 조용한 표결을 허용했다.		민주적 정권을 보장하기 위한 국가 공공 기관의 유능하고 진지한 노력만이 그런 침착한 표결을 수행하게 했습니다.	민주적 정권을 보장하는 국가 공공 기관의 유능하고 성실한 임무 수행 덕에 그런 안정된 투표가 가능했다.	25	16560.0	El País	18
7	Destacamos la Justicia Electoral, rama especializada del Poder Judicial brasileño responsable de organizar y realizar las elecciones, que, una vez más, demostró estar a la altura de la noble función que le fue atribuida.	우리는 선거 조직 및 실시에 책임이있는 브라질 법무부의 전문 분과 인 선거 재판을 강조하며 다시 한 번 선거에 기인 한 고귀한 기능까지 입증되었습니다.	우리는 선거 조직 및 실시에 책임이있는 브라질 법무부의 전문가들과 선거 재판을 강조하며 다시 한 번 선거에 기인 한 고귀한 기능까지 입증되었습니다.		우리는 브라질 사법부 내의 선거 조직 및 시행에 책임이 있는 선거 법원이 다시 한 번 본연의 고귀한 기능을 입증했음을 강조한다.	34	53230.0	El País	4

Figure 5.4: A feedback sheet for the warm-up session.

바람직한 포스트에디팅 방법:

새로운 민주주의주기 시대가 브라질에서 시작될 예정입니다.

사람들은 집행력 행정부, 연방 입법부 및 앞으로 다가오는 여러 국가의 각 주지사 등 앞으로 다가오는 임기의 대표를 선출했습니다.

브라질 사람들은 침착하게 여론 조사 투표에 참여했으며 다시 또 민주주의의 재건 의례에 참여했습니다.

전자 투표 시스템 덕분에 우리는 불과 2 시간 18 분 만에 대선 결과를 배웠 알게되었습니다.

브라질의 대륙 차원의 크기의 국가 면적, 인구가 2 억 8 천 5 백만명, 유권자 1 억 4 천 7 백만명,으로 세계 4 위의 민주주의 국가에서는 등을 고려하면 이는 충분하지 않은 시간입 습니다.

민주적 정권을 보장하기위한 는 국가 공공 기관의 유능하고 진지한 성실한 노력 임무 수행 만이 덕에 그런 조용한 안정된 표결을 투표가 허용 가능했다.

우리는 브라질 사법부 내의 선거 조직 및 실시 시행에 책임이있는 브라질 법무부의 전문 분과 인 선거 재판을 법원이 강조하며 다시 한 번 선거에 기인 한 본연의 고귀한 기능까지을 입증되었습니다.했음을 강조한다.

Figure 5.5: A correction sheet for the warm-up session.

The number of post-edits and translated words per editor is displayed in Table 5.6. In terms of post-editing outputs, many of the editors performed the task in such a diverse way that each version of post-editing had to be individually corrected. The striking point was that EV3 and EV6 performed under-editings while EV2 did over-editing. In terms of translation from scratch, most of the post-editors introduced translations of longer forms than was needed, but within an acceptable boundary. EV6 was advised to translate more carefully and post-edit more as her inputs were way below the average.

	Seg 1	Seg 2	Seg 3	Seg 4	Seg 5	Seg 6	Seg 7
EV1	7	73	10	38	21	46	15
EV2	3	84	44	51	22	69	75
EV3	0	54	0	53	10	50	27
EV4	7	70	0	46	40	44	25
EV5	0	64	0	46	30	71	36
EV6	0	22	0	11	27	18	4
RT	***10***	***58***	***10***	***42***	***27***	***64***	***64***

Table 5.6: The number of post-edits (in the white columns) and translated words (in the gray columns) per segment per editor in the warm-up session. The number of reference translation (RT) is also calculated.

From a qualitative point of view, most of the editors restrained themselves from changing the system translation as much as possible, in such a way that many segments were left to be syntactically unnatural or semantically inadequate and did not satisfy the full post-editing criteria. Especially the proper nouns such as job titles or names of an institute were left untranslated.

Some post-editors stated that the guideline gave the impression that the segment should be maintained as much as possible. Others declared that they could not help but leave the errors without edits, be it that they noticed them because they accidentally pressed the Enter button and going back to the previous segment was systematically prohibited. Having experienced the task and workbench in advance, many of them declared that the user interface of the workbench was so confusing that they often translated or post-edited the previous source segment, instead of the current one. EV2, in fact, had difficulty availing herself of the workbench, performing post-editing on one segment and skipping the next where translation should have been added. In any sense, the overall opinion was that after having gone through the warm-up session and got corrections, their understanding of the project and technical skills was enhanced considerably.

5.5. MT Evaluation

This chapter addresses the experiment conduct of Project I, II and III (as titled in Chapter 5.3.4. Project I is the post-editing evaluation which is launched prior to the other two types of evaluation in order to avoid a situation where the post-editors have previous contact with the dataset. The process of each evaluation task is described in Chapter 5.5.1 (post-editing), 5.5.2 (fluency and adequacy scoring) and 5.5.3 (segment ranking), respectively. In the case of Project II and III, the editors are allowed to begin with whichever they feel comfortable.

5.5.1. Post-Editing

The first project began on December 1 and ended on December 7. All the post-editors met the deadline. Out of 253 sentences, 127 sentences were post-edited and 126 sentences were translated from the beginning. The total number of post-editing and translation numbered 38,556 words, and the total time spent on the first project was 54.91 working hours (for six post-editors). The specifics of the time taken for post-editing and translation from scratch per editor is presented in Table 5.7. They spent on average 9.15 hours to complete the task, and post-editing was faster than translation from scratch throughout the post-editors. The more detailed analysis will be given in Chapter 6.

	EV1	EV2	EV3	EV4	EV5	EV6
PE	3.99	4.45	6.41	1.54	2.84	2.18
T	6.35	7.46	8.44	2.29	5.10	3.86
Total	10.34	11.91	14.85	3.83	7.94	6.04

Table 5.7. Total time of post-editing and translation from scratch per editor (unit: hour).

Figure 5.6 shows a fraction of the result sheet of post-editing prepared systematically from the workbench. The system calculated each segment length, post-editing and translation time and the edit efforts. Note that the calculation unit of the segment length is word-based. Moreover, the edit efforts stand for a TER score. The time was calculated in milliseconds.

ID	Source Segment	MT Target Segment	Post-edited Target	Translated Target	Segment Length	Time(ms)	Segment Origin	Edit Effort
1	[HEADLINE] El CIS otorga a Susana Díaz una amplia mayoría en las elecciones andaluzas	CIS는 Susana Díaz에게 안달루시아 선거에서 대다수를 준다.	CIS는 수사나 디아쓰에게 안달루시아 선거에서 대다수를 준다.		14	133980.0	El País	26
2	[SUB-HEADLINE] La encuesta sitúa a la coalición de Podemos e IU en segundo lugar y anticipa un empate técnico entre PP y Ciudadanos	이 설문 조사는 Podemos와 IU의 연합을 두 번째 장소에두고 PP와 Citizens 사이의 기술적 인 관계를 예상합니다		설문조사에 의하면 포데모스와 IU 연합이 2위를 차지할 것이고, 국민당과 시우다다노스가 비슷한 표를 얻을 것으로 예상된다.	23	242690.0	El País	73
3	La presidenta de la Junta de Andalucía, Susana Díaz, ganaría las próximas elecciones autonómicas con más de 10 puntos de ventaja, pero lejos de la mayoría absoluta (55 escaños de un total de 109), según la encuesta hecha pública este miércoles por el Centro de Investigaciones Sociológicas (CIS).	유다 데 안달루시아 (Susana Díaz) 의장은 다음 주 지방 선거에서 10 점 이상의 우위를 점할 것으로 예상했지만, 이번 수요일에 센터에 의해 공개 된 설문 조사에 따르면, 절대 다수 (109 개 가운데 55 석) 사회학 연구 (CIS).	안달루시아 위원회의 수사나 디아쓰 의장은 다음 지방 선거에서 10 점 이상의 우위를 점할 것으로 예상했지만, 이번 수요일에 사회 조사 기관(CIS)에 의해 공개 된 설문 조사에 따르면, 절대 다수 (109 개 가운데 55 석) 에는 미치지 못할 것으로 예상했다 .		47	233700.0	El País	34
4	La candidata socialista obtendría el 37,41% de los votos y entre 45 y 47 escaños.	사회당 후보는 37.41 %의 표와 45 - 47 표의 표를 얻는다.		사회주의 입후보자는 표의 37.41%와 45-47개 의석을 얻을 전망이다.	15	95780.0	El País	56
5	Esta última cifra coincide con el número de diputados que el PSOE tiene en la actualidad.	이 마지막 숫자는 PSOE가 현재 보유한 대리인의 숫자와 일치합니다.	이 마지막 숫자는 PSOE가 현재 보유한 국회 의원의 숫자와 일치합니다.		16	36360.0	El País	12

Figure 5.6: Example of the result sheet of one editor's post-editing acquired from TAUS.

5.5.2. Fluency & Adequacy Scoring

Having completed post-editing, the evaluators[49] received a guideline for the second and third evaluation tasks —fluency and adequacy scoring and segment ranking. The second project resorted to a sole guideline and no warm-up session was organized. The same group of people were asked to go over the guideline and to raise questions for one day. Upon confirmation, they were given two emails additionally for the fluency and adequacy evaluation and segment ranking evaluation. They were allowed to start with whatever they felt comfortable but within seven days. The second tasks ended on December 14.

A preview of the result sheet with the first few sentences is presented in Figure 5.7. It shows the source text and system translation together with the number of words of the source text and the system translation. Note that the counting system is both word-based. A comment section is granted to justify their assessment or claim some problems. The evaluation of fluency and adequacy scoring follows the Likert scale of 1 to 4, and the scale information is given on each segment. The total amount of assessment data of fluency and adequacy reached 3,036 (253 sentences by six evaluators).

[49] The post-editors are addressed as 'evaluators' or 'judges' instead of 'post-editors' or 'editors' in this section.

ID	Source Segment	MT Target Segment	Segment Origin	Comment	Number of Words (Source Segment)	Number of Words (Target Segment)	Fluency	Fluency Score	Adequacy	Adequacy Score
1	[HEADLINE] El CIS otorga a Susana Díaz una amplia mayoría en las elecciones andaluzas	CIS는 Susana Díaz에게 안달루시아 선거에서 대다수를 준다.	El País		14	7	Disfluent	2	Little	2
2	[SUB-HEADLINE] La encuesta sitúa a la coalición de Podemos e IU en segundo lugar y anticipa un empate técnico entre PP y Ciudadanos	이 설문 조사는 Podemos와 IU의 연합을 두 번째 장소에두고 PP와 Citizens 사이의 기술적 인 관계를 예상합니다	El País		23	16	Disfluent	2	Little	2
3	La presidenta de la Junta de Andalucía, Susana Díaz, ganaría las próximas elecciones autonómicas con más de 10 puntos de ventaja, pero lejos de la mayoría absoluta (55 escaños de un total de 109), según la encuesta hecha pública este miércoles por el Centro de Investigaciones Sociológicas (CIS).	유다 데 안달루시아 (Susana Díaz) 의장은 다음 주 지방 선거에서 10 점 이상의 우위를 점할 것으로 예상했지만, 이번 수요일에 센터에 의해 공개 된 설문 조사에 따르면, 절대 다수 (109 개 가운데 55 석) 사회학 연구 (CIS).	El País		47	36	Disfluent	2	Most	3
4	La candidata socialista obtendría el 37,41% de los votos y entre 45 y 47 escaños.	사회당 후보는 37.41 %의 표와 45 - 47 표의 표를 얻는다.	El País		15	11	Flawless	4	Most	3
5	Esta última cifra coincide con el número de diputados que el PSOE tiene en la actualidad.	이 마지막 숫자는 PSOE가 현재 보유한 대리인의 숫자와 일치합니다.	El País		16	9	Flawless	4	Most	3
	Con estos resultados, y si no									

Figure 5.7: Example of a result sheet of fluency and adequacy scoring evaluation.

5.5.3. Segment Ranking

As stated previously, the guideline for the current evaluation was provided altogether when the second part of the project began. The evaluators were allowed to begin with fluency and adequacy scoring or segment ranking, as they were already well aware of the dataset after the post-editing evaluation and both of the evaluation required previous knowledge on it. In this evaluation, the system translation of Google Translate was compared to human translation and a system translation of Kakao i. The evaluators were asked to leave comments upon necessity. In the result sheet (see Figure 5.8), all the three candidate translations are displayed with their ranking. In Figure 5.8, it is noticeable that the editor made comments on each segment, engaging actively in the job.

In Segment 4, for example, she pointed out that “the first evaluation standard was that Rank 1 translated ‘a Socialist candidate’ as ‘the Socialist Party’ while the others translated as it was, but both of the options sounded natural. As such, the judgment of the ranking in this segment was based on accurate handling of the number expressions.” To take a look at their translations, ‘vote (*voto*)’ and ‘seats (*escaños*)’ are uniformly translated to vote (표) in the system translation of Google Translate. On the contrary, Kakao i translated them to vote (표) and seats (석), respectively. From the comment, it is well justified why Kakao i outranked Google Translate on that segment. From

the ranking comparison, 4,554 evaluation data were collected.

ID	Source Segment	Segment Origin	Comments	Human Translation		Google Translate		Kakao i	
				Target Segment	Rating	Target Segment	Rating	Target Segment	Rating
1	[HEADLINE] El CIS otorga a Susana Díaz una amplia mayoría en las elecciones andaluzas	El País	CIS만으로는 설문 결과인지 알 수 없고, 대다수가 아닌 과반수가 맞는 표현이라 생각한다. 또한 이름과 지명을 모두 한글로 바꾼 것을 1순위로 주었다.	CIS의 설문 결과 수사나 디아스가 안달루시아 선거에서 과반수로 승리할 것으로 나타났다.	1	CIS는 Susana Díaz에게 안달루시아 선거에서 대다수를 준다.	3	CIS는 안달루시아 선거에서 수자나 디아즈에게 과반수를 부여한다	2
2	[SUB-HEADLINE] La encuesta sitúa a la coalición de Podemos e IU en segundo lugar y anticipa un empate técnico entre PP y Ciudadanos	El País	2위와 동률이라는 표현이 정확히 해석된 것에 1순위를 주었다. 다른 두 문장은 표현이 어색하다	여론조사 결과 포데모스와 IU 연대가 2위, PP와 시우다다노스가 동률 예상되다.	1	이 설문 조사는 Podemos와 IU의 연합을 두 번째 장소에두고 PP와 Citizens 사이의 기술적인 관계를 예상합니다	2	이 설문 조사는 Podemos와 IU의 연합을 2 위로 끌어 올리고 PP와 Citizen 사이의 기술적 동점을 예상합니다.	2
3	La presidenta de la Junta de Andalucía, Susana Díaz, ganaría las próximas elecciones autonómicas con más de 10 puntos de ventaja, pero lejos de la mayoría absoluta (55 escaños de un total de 109), según la encuesta hecha pública este miércoles por el Centro de Investigaciones Sociológicas (CIS).	El País	3 순위 문장은 동사가 없고 기관이름이 제대로 번역되지 않았다. 2순위는 이사회, 벗어날 것으로 나타났다 등 컨텍스트에 맞지 않는 번역이다	사회학연구센터(CIS)가 이번 수요일 공식 발표한 여론조사에서 안달루시아 의회의 수사나 디아스 의장이 차기 자치선거에서 10석 이상의 우위를 점할 것이지만 과반수(109석 중 55석)는 불가능할 것으로 나타났다.	1	유다 데 안달루시아 (Susana Díaz) 의장은 다음 주 지방 선거에서 10 점 이상의 우위를 점할 것으로 예상했지만, 이번 수요일에 센터에 의해 공개 된 설문 조사에 따르면, 절대 다수 (109 개 가운데 55 석) 사회학 연구 (CIS).	3	사회학연구센터(CIS)가 수요일 발표한 여론조사에 따르면 안달루시아 이사회 수잔라 디아즈 회장은 10점 이상 앞선 차기 자치선거에서 승리할 것이지만 절대 다수당(109석 중 55석)에서 벗어나는 것으로 나타났다.	2
4	La candidata socialista obtendría el 37,41% de los votos y entre 45 y 47 escaños.	El País	후보자라는 표현을 쓰지 않고 정당 이름만을 사용해도 번역이 자연스럽다. 단 표와 의석 수를 정확히 번역하는 것에 의미를 두었다.	사회노동당이 37.41%로 45 - 47석을 획득할 전망이다.	1	사회당 후보는 37.41 %의 표와 45 - 47 표의 표를 얻는다.	2	사회주의 후보자는 투표의 37.41 %와 45-47 석을 얻을 것이다.	1
5	Esta última cifra coincide con el número de diputados que el PSOE tiene en la actualidad.	El País	의원수가 맞다	이는 PSOE가 현재 보유하고 있는 의원 수와 일치한다.	1	이 마지막 숫자는 PSOE가 현재 보유한 대리인의 숫자와 일치합니다.	2	이 마지막 수치는 현재 PSOE가 보유하고 있는 회원 수와 일치합니다.	2

Figure 5.8: Example of a result sheet of segment ranking evaluation.

5.6. Feedback

For qualitative analysis of the data (in Chapter 6), feedback from the participants were gathered. The feedback took the form of a survey on Google Forms and the link is provided[50]. The survey has two separate sections of general feedback and suggestion for improvement. The evaluators responded in Korean, and the translation of their feedback is presented in the following section.

1) **General feedback**: Tell us about your opinion regarding the overall evaluation task; How do you feel about the performance of Google Translate in the Spanish-to-Korean translation? How was the post-editing work? Was the system translation helpful to post-editing? How was the working environment (i.e. the user interface of the workbench)?

   - “I was very surprised at the quality of GNMT. It was much more

[50] https://docs.google.com/forms/d/e/1FAIpQLScgdkrGmd64yYIlZD9kGye2X6IFUu5gwbOK2DIF1ttuFm_4vg/viewform

accurate than I expected. Even though some unnatural expressions were noticed, it seemed like the engine considered the context and inferred the right meaning for each word from it without failure. The most impressive point was that the engine knew the syntactic structure of a very long and complicated sentence and translated it right. The biggest weakness of the NMT, however, was that despite the marvelous success, once the system selected wrong translation for a word, the whole sentence went in a totally different direction. The aim of the experiment is to edit the machine-translated text. And so, I had difficulty revising the sentence even though I felt like it would be much more effective if I translated from scratch. I guessed it would not be a problem in the real world. Besides that, I felt that post-editing was more efficient than translation. Moreover, in my case, if I was assigned to a translation job, what I did first was to do a preparatory investigation of the topic by reading similar texts as well as going through the whole source text in advance, but in the experiment I had to translate or revise the sentence without knowing the whole context. That was very challenging to me, and I assumed that the quality of my output would be degraded for such reason."

-Editor 1

- "I thought MT was not bad if you want to acquire brief information of a text. The fact that the proper nouns such as names of an organization or of a person were translated in a Korean way gave me a pleasant shock. In terms of PE, the fact that the workbench measured the time pressured me and led me to be in a hurry and make mistakes, i.e. pressing the Enter button for nothing."

-Editor 2

- "It was the first time for me to practice post-editing. What I felt from the task was that for a short and simple sentence, the post-

editing was very effective, just revising for a correct semantic equivalence. In the case of a complex and long sentence, on the other hand, the structure had to be transformed due to the wrong word order, so post-editing was not worth it. Especially, the Spanish way of writing many sentences in succession by putting commas did not work for Korean. I should punctuate the sentences, but the word order of the system translation was too complicated to just revise them."

-Editor 3

- "I think MT serves its role when one needs a rough version of translation to just get the idea of the text or to translate something against the clock. In other words, MT is very efficient but has a large number of errors that without post-editing the output cannot be served as a finished product. Some had stylistic problems that could be improved with a simple touch whereas others were heavily mistranslated or even left untranslated when the segment was syntactically complex. In general, however, I felt that post-editing was very effective due to the fact that most of the machine-translated translations were more or less well structured and did not depart from the usual way of Korean syntax (be it that the experiment ignored the spacing rule). I guess half of the post-editing work that I did was a minor revision."

-Editor 4

- "The performance of MT was very good for short and simple sentences, but it was of no help to long and complex sentences except that the machine prepared a rough draft. I felt like MT was weak at selecting words including nuance. When practicing post-editing for long sentences, it was rather more challenging for me to figure out the structure of the system translation than to translate from scratch. In relation to the working environment, I felt something was wrong that I could not go through the whole text.

Displaying only the previous and next sentence was not enough for the proper work. Moreover, the separation was so vague that I referred many times to the wrong sentence to translate or post-edit."

-Editor 5

- "As a system translation, the quality of the translation was just enough to understand what the sentence tried to imply. It is also true, however, that many errors were detected and they needed to be revised. In relation to the tool, the user interface was convenient."

-Editor 6

2) **Suggestion for improvement**: Tell us how the experiment can be improved.

- "It would be nice if the post-editing can be done in a paragraph or document level so that the post-editors can take the context into consideration. I assume that the context is very important in translation and in the real world, post-editors will have to embark on the task with the context in mind. I also wish to have an offline experiment where I can make questions and get simultaneous feedback."

-Editor 1

- **"**Regarding the user interface of the workbench, I suggest that you click the 'Next' button when moving to the next segment, not pressing the Enter on keyboards. I do not know if it is my personal habit, but I think many people would press the Enter by accident. Moreover, the current segment is not well highlighted, and it is so easy to translate/post-edit the previous or next segment. The workbench has to be improved."

-Editor 2

- “In the workbench, the system skipped to the next segment when pressing the Enter button or clicking the Next button. It was very annoying for me that I could not press the Enter to move to the next line because I was so used to it. It was painful that you knew there were mistakes but there was nothing you can do. It would be better if they delete that function.”

  -Editor 3

- **“**I think post-editing will be much easier if the performance of MT gets better.”

  -Editor 4

- “I guess the result of the experiment will be a bit different from that of the real world because there are many constraints such as you cannot go back to the previous sentence. It would be realistic if they allow the workers to go to the previous sentence at least once. That makes more sense, considering that with just one Enter key the segment vanishes. I think it needs to be improved.”

  -Editor 5

- No comment.

  -Editor 6

5.7. Chapter Summary

Chapter 5 reported on the parameters and process of the main experiment. In Chapter 5.1, we detailed the dataset of this experiment: 6,246 words (253 sentences) in the newswire domain. In Chapter 5.2, we provided detailed information about the six evaluators (post-editors) who were female native Koreans and professional translators in the *es-ko* pair. They were hired online and officially contracted for the task with the appropriate payment. In detail, their educational background differed from BA to Ph.D. Two of them majored in translation studies and the rest were from different majors. They had 1 ~ 5 work experience in the given language pair. In Chapter 5.3, our workbench TAUS

DQF was described. We complimented the user-friendly interface of the workbench and explained how to set up a project for the three types of MT evaluation: fluency and adequacy scorings, segment ranking, and post-editing. We also illustrated how each interface looked like.

In Chapter 5.4, we reported on how the experiment was carried out on chronological order. The experiment lasted two weeks. On the first week, post-editing was carried out, and on the second week, the fluency & adequacy scoring and segment ranking were performed. Prior to the experiment conduct, we had a training session which was composed of distribution of a guideline and a warm-up test. In Chapter 5.5, we stated how the results were collected by illustrating the result sheets of each evaluation type. In Chapter 5.6, we described feedback obtained from the evaluators about the perception of MT and post-editing.

[blank page]

PART III. ANALYSIS

Part III plays a key role in this thesis by reporting on findings of MT evaluation and error analysis. Firstly, we carry out general analysis of MT evaluation in a quantitative and qualitative way in Chapter 6. The main purpose of such analysis is to measure the performance of the NMT engine —Google Translate— in the *es-ko* language pair. We propose a fluency score, adequacy score, ranking score, post-editing time/productivity, and post-editing effort. We also compute HTER scores.

Secondly, in Chapter 7, we carry out error analysis in a binary way: first on a raw system translation and second on its post-edited version. For the task, we design an error classification adapted to the *es-ko* pair. The fine-grained error analysis gives in-depth ideas about the type of errors of the NMT engine and its behavior towards such errors. The second error analysis finds error types perceived by the post-editors.

[blank page]

6. EVALUATION ANALYSIS

This chapter reports on findings from the experiment carried out in Chapter 5. Many interesting findings are discovered from the experiment after quantitative and qualitative analysis. Such a dual approach will be given in each MT evaluation method of fluency and adequacy scoring (Chapter 6.1), segment ranking (Chapter 6.2) and post-editing (Chapter 6.3). The chapter ends with a chapter summary in Chapter 6.4.

6.1. Fluency & Adequacy Scoring

This section reports on the results of the fluency and adequacy scoring evaluation. The fluency score refers to “what extent the translation is well-formed grammatically” while the adequacy score stands for “what extent the meaning in the source text is expressed in the target text” (Görög, 2014). Each score is based on a 4 point scale and is calculated on 253 sentences (6,426 words) by six evaluators (titled as *EV*).

6.1.1. Fluency Scoring

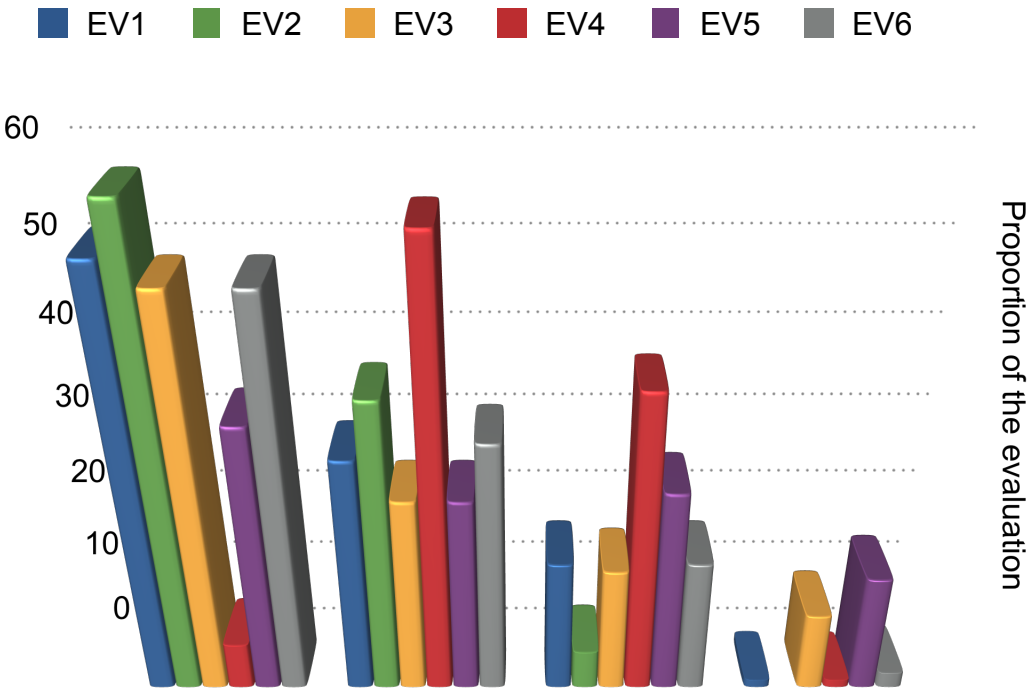
After reporting the fluency score of GNMT in Chapter 6.1.1.1, the fluency score was analyzed in terms of its correlation with sentence length in Chapter 6.1.1.2. Such quantitative analysis was further detailed in Chapter 6.1.1.3 with qualitative analysis.

	EV1	EV2	EV3	EV4	EV5	EV6	**mode**	**mean**
F	3.33	2.78	3.13	2.66	3.29	3.53	**4**	**3.12**
%	83.25%	69.50%	78.25%	66.50%	82.25%	88.25%	**100**	**78.00%**

Table 6.1: Fluency score of GNMT (*F* = Fluency).

6.1.1.1. Fluency Score

The mean average score of GNMT was 3.12 of 4, equivalent to 78% of fluency. TAUS also suggested referring to a mode average as a more reliable indicator



	4-Flawless	3-Good	2-Disfluent	1-Incomprehensible
EV1	52	30	17	1
EV2	58	37	5	0
EV3	49	25	16	10
EV4	6	55	38	1
EV5	34	25	26	15
EV6	49	32	17	2
Avg	**41.3**	**34.0**	**19.8**	**4.8**

Figure 6.1: Distribution of fluency scores per evaluators and their average. (unit: %)

Table 6.2: Distribution of fluency scores per evaluators and their average. (unit: %)

because it was calculated on an aggregated basis (TAUS, 2010). The mode average of Google Translate in this experimental setup was 4 of 4 (100%). In Table 6.1, the evaluation scores of each evaluator were displayed. They ranged from a minimum of 66.5% to a maximum of 88.3%. Such score claimed that the fluency of target language (Korean) of the given engine averaged 78%, with the possibility of going up to 88.3%.

The score distribution of all editors was detailed in Figure 6.1 and Table 6.2. First of all, the graph showed that the proportion of evaluations was higher in the positive scores. For instance, Score 3-Good and 4-Flawless occupied on

average 87% of all segments, showing that 87% of the sentences in the dataset were *almost* perfect Korean sentences from a grammatical perspective. Meanwhile, Score 1-Incomprehensible obtained the lowest proportion throughout all evaluators, and moreover, EV2 did not mark any sentences with Score 1. It could be interpreted as a positive sign.

As much interesting as the result was the proportion of the scores of EV4, which had a noticeably lower proportion of Score 4 but a higher proportion of Score 3, compared to other evaluators. Considering that she had drastically different criteria of judgment, an average was additionally computed excluding her scores, which was given in Table 6.3. The exclusion of EV4 resulted in a quite dissimilar average, with about 7% point higher score proportions in Score 4. According to this new average, it could be inferred that 48% of the dataset was a *perfect* Korean sentence. However, although EV4 seems to have more strict standards of evaluation than others, this thesis acknowledges that such phenomenon is entirely natural and the judgment can vary.

	Score 4	Score 3	Score 2	Score 1
Avg	41.3	34.0	19.8	4.8
Avg*	**48.4**	**29.8**	**16.2**	**5.6**

Table 6.3: New average fluency score that excluded EV4's score. (unit: %)

Contrary to a positive side of the results addressed previously, there was still a long way for the given engine to go considering that on average 4.8% of the dataset was considered as *incomprehensible*, which was equivalent to 11 sentences. The judgments of the evaluators could vary, but the fact that up to 15 sentences (according to EV5) were entirely meaningless raised a serious question about the reliability of the given engine.

6.1.1.2. Correlation with Sentence Length

The fluency score was further analyzed to detect if there was a correlation between the fluency score and a sentence length. The fluency score of all evaluators was averaged on a sentence basis and organized by the sentence length in Figure 6.2. Figure 6.2 described the proportion of each score, along

with the correlation of the two parameters in a logarithmic line. Although the individual scores did not exhibit a clear-cut tendency, it seemed that shorter sentences tended to earn a higher score. Moreover, the graphical distribution manifested that the high-scored sentences were not detected in the sentences of over 45 words.

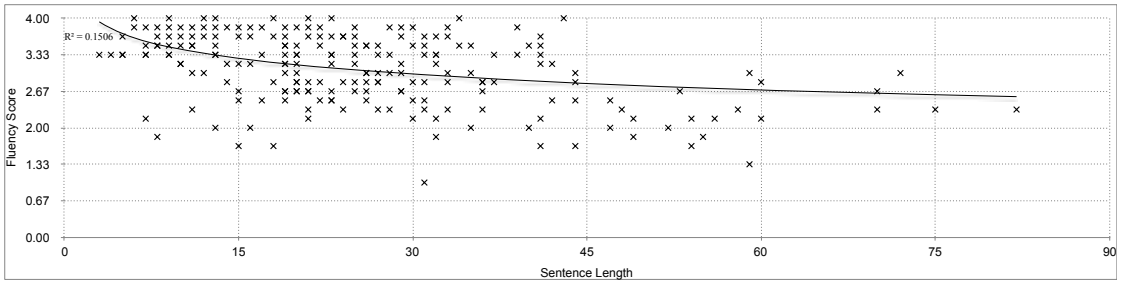


Figure 6.2: Correlation of fluency score and sentence length graphically shown with a logarithmic line.

6.1.1.3. Qualitative Analysis

With the quantitative result of 78% of fluency at hand, the data were analyzed qualitatively to look into some of the best and worst cases. The most *fluent* sentence was defined as a sentence with Score 4-Flawless throughout all six evaluators. The most *incomprehensible* sentence, on the other hand, was defined as a sentence with Score 1-Incomprehensible throughout all six evaluators. There were 10 sentences where all evaluators agreed that they were flawless (Score 4), which were given below. Each one was discussed one by one in terms of possible errors in the sentence that were regarded as *not an error* from the evaluators. (ST = a source text, GT = a system translation of Google Translate, BT = a back translation of GT to English)

Sentence 6-1

ST “No puedes ganar.

GT "당신은 이길 수 없습니다.

BT *"You can not win.*

Sentence 6-1 was the shortest sentence of the 10 sentences, composed of only 3 words. No errors had been detected, and it surely was a perfect translation

including generating the correct subject that usually hidden in the verb in Spanish.

Sentence 6-2

ST Y tampoco puedes abandonar el juego”.

GT 그리고 게임을 떠날 수도 없다. "

BT *And you can not leave the game. "*

Sentence 6-2 was composed of 6 words. Unlike Sentence 6-1, the subject was omitted, but it did not affect the sentence as Korean tended to omit a subject. In fact, considering the previous sentence (Sentence 6-1), it sounded more natural without the subject. Except that the endings of the two verbs in Sentence 7-1 and 7-2 were inconsistent (없습니다 vs. 없다), they were taintless.[51]

Sentence 6-3

ST «Pero el populismo no puede ser la solución a estos problemas», advierte.

GT "그러나 포퓰리즘은 이러한 문제에 대한 해결책이 될 수 없다"고 그는 경고했다.

BT *"But populism can not be a solution to this problem," he warned.*

Sentence 6-3 was composed of 12 words. This sentence was more linguistically challenging than the previous two sentences in that it included quotation marks which were in a different form («, ») from the Korean marks (“, ”) and that the sentence was a complex sentence. In fact, it was very surprising how a machine could deal with such complexity with so much accuracy, not to mention the fact that there were any spacing errors.

Sentence 6-4

ST El estado de Vermont fue el primero en abrir las puertas de sus colegios electorales, a las 5 de la madrugada, hora local.

GT 버몬트 주 (州)는 현지 시간으로 오전 5시에 처음으로 투표소 문을 열었습니다.

BT *Vermont has opened its polling place for the first time at 5 am local time.*

Sentence 6-4 was composed of 23 words, which was longer than the previous

[51] Note that it was our intention not to expect a consistency of the verb ending from Google Translation.

ones. Translating this sentence could be challenging considering that the two commas interfered the sentence. However, GT exhibited any errors and looked perfect in the eyes of a Korean linguist. The noticeable point was that the system even produced a Chinese character (州) of *estado* (state) as additional information to clarify the meaning.

Sentence 6-5

ST La campaña electoral finalizará el viernes 30 de noviembre.

GT 선거 운동은 11 월 30 일 금요일에 끝납니다.

BT *The campaign ends on Friday, November 30th.*

Sentence 6-5 was composed of 9 words. There were spacing errors when processing numeric expressions such as a date (11-월 30-일 > 11월 30일), but it did not impair human comprehension. A rather serious issue was that the source sentence had a future tense (finalizará : will finish) while the translation was present tense. Strictly speaking, it was not a perfect translation, but the evaluators somehow considered that there was no problem in meaning.

Sentence 6-6

ST El sondeo presenta un margen de error de 3,1 puntos porcentuales y un nivel de confianza del 95%.

GT 설문 조사의 오차 범위는 3.1 % 포인트이고 신뢰 수준은 95 %입니다.

BT *The error range of the survey is 3.1% and the confidence level is 95%.*

Sentence 6-6 was composed of 18 words. GT used such perfect expressions for the source sentence that it did not feel like an MT raw output. Some spacing errors were detected from the numeric expressions (3.1-% > 3.1%; 95-%> 95%), as well, but again, it did not impair the judgment.

Sentence 6-7

ST Mientras en Georgia, la demócrata Stacey Abrams ha prometido tomar acciones legales para garantizar que todos los votos se cuenten correctamente.

GT 조지아 주에서 민주당의 스테이시 에이 브람스 (Stacey Abrams)는 모든 득표가 올바르게 집계 될 수 있도록 법적 조치를 취할 것을 약속했습니다.

BT *Democrat Stacey Abrams in Georgia promised to take legal action to ensure that all votes are counted correctly.*

Sentence 6-7 was composed of 21 words. GT was a perfect Korean sentence except for some spacing errors when processing the name Abrams (에이-브람스> 에이브람스) and the omission of Mientras (while).

Sentence 6-8

ST Eso significa que de los 42.900.000 ciudadanos con derecho de sufragio, 25.840.000 votaron.

GT 투표권이있는 42,900,000 명의 시민 중 25,840,000 명이 투표했습니다.

BT *Of the 42.9 million citizens with voting rights, 25.84 million voted.*

Sentence 6-8 was composed of 13 words. As previously witnessed, there were spacing errors with the numeric expressions (42,900,000-명> 42,900,000명; 25,840,000-명> 25,840,000명) and normal expression that required a spacing (투표권이있는> 투표권이-있는). More importantly, while it was a complex sentence, one part of the sentence was entirely omitted: Eso significa que (It means that). However, all evaluators decided that such error did not affect the translation.

Sentence 6-9

ST En 2018, según el informe de Bresso y Wieland, se ha asignado un total de 32,44 millones de euros para subvenciones a los partidos políticos europeos dentro del presupuesto de la UE, y 19,32 millones de euros para subvenciones a fundaciones políticas europeas.

GT Bresso와 Wieland 보고서에 따르면 2018 년 EU 예산 내에서 유럽 정당 보조금으로 324.4 백만 유로, 유럽 정치 재단 보조금으로 1,932 만 유로가 할당되었습니다. .

BT According to the Bresso and Wieland report, within the EU budget in 2018, Euro-party subsidies amounted to Euro 324.4 million and European Political Foundation subsidies amounted to € 19.32 million. "

Sentence 6-9 was composed of 43 words, the longest of all. While the fact that such long sentence could be flawless proved the strong performance of the given engine, it was, in fact, an interesting case where some obvious errors were displayed but the evaluators did not take it into consideration. Firstly, two proper nouns were not translated: Bresso and Wieland. Secondly, the numeric expression was incorrect: 32,44 millones. While GT processed it as 324.4 백만 유로, which was neither correct (being translated to 324.4 million) nor natural (a

correct form was 3억2440만), the correct translation of the source segment was 3,244만 유로 (32.44 millions). Thirdly, there were two periods at the end of the sentence. The speculation was that the evaluators did not take much attention to the numeric values and stylistic errors such as the periods.

Sentence 6-10

ST	Según la Comisión Nacional Electoral, 78 mil contadores de votos, 12 mil observadores, 105 mil administradores y 2.500 máquinas para el recuento automatizado de papeletas han sido movilizados para el escrutinio a nivel nacional.
GT	전국 선거위원회 (National 선거위원회)에 따르면 78000 명의 투표자, 12000 명의 옵서버, 105000 명의 관리자 및 2,500 대의 자동 투표기가 전국 조사를 위해 동원되었습니다.
BT	*According to the National Election Commission, 78,000 voters, 12,000 observers, 105,000 managers, and 2,500 automatic voting machines were mobilized for national surveys.*

Sentence 6-10 was composed of 34 words. While no particular errors were detected except the usual spacing errors with numeric expressions (78000-명> 78000명; 12000-명> 12000명; 105000-명> 105000명; 2,500-대> 2,500대). An erroneous part was that when processing the proper noun la Comisión Nacional Electoral, it was duplicated and a part of it was translated in English.

All in all, the noticeable point from the in-depth analysis of the 10 best sentences was that all sentences were in a perfect word order, which could be challenging in the *es-ko* pair. Furthermore, the successful performance of the given engine was witnessed in the sentences irrespective of a sentence length. Although some problems were discussed in terms of spacing errors that did not impede the quality or normal errors that the evaluators ignored, those sentences surely proved that NMT was robust.

In the case of the worst sentences in the fluency score that were marked with Score 1-Incomprehensible throughout the evaluators, 7 cases were detected. The sentence length of those sentences ranged from 8 - 30, and the most disfluent sentence obtained 1.33 score (of 4), displaying a 33.25% fluency. The rest of the six sentences earned identically 1.67 score (41.75%). The sentences were analyzed one by one from a structural point of view, as considerable errors were expected. The in-depth analysis of all errors is

prepared in Chapter 7.

Sentence 6-11

ST	La última pésima noticia ha sido la dimisión, el viernes pasado, del ministro de Transportes, Jo Johnson, autor de esa diatriba entre vasallaje y caos, es decir, entre el sometimiento a las reglas del mercado único sin voz ni voto de Londres (la oferta de May rechazada por todos) o la desastrosa salida sin acuerdo el 29 de marzo.
GT	최악의 소식은 지난 금요일, 교통 장관, 조 존슨 (Jo Johnson)의 사퇴, 즉 매매와 혼돈 사이의 비난 즉, 목소리 나 런던의 투표가없는 단일 시장의 규칙에 대한 제출 사이다. 3 월 29 일에 합의없이 모든 사람이 거부 할 수 있음) 또는 비참한 출발.
BT	*The worst news was last Friday's minister of transportation, Jo Johnson's resignation, the blaming between trade and chaos, the submission of a single market rule without a voice or London vote. Everyone can refuse without consent on March 29) or a miserable departure.*

Sentence 6-11 was composed of 59 words and the sentence earned 1.33 fluency score. The observable feature of this sentence was that a couple of new information were added with commas and an additional phrase was inserted in brackets so that the sentence was highly complicated to be structured. Most of all, GT was composed of two sentences: one was incomplete in meaning and the other was incomplete in structure. To backtrace how the engine processed the source text, the translation was quite smooth until it reached the brackets where it regarded the segment as not related to the previous segments and cut it out in a new sentence even ignoring the range of the brackets. Firstly, the reason the first segment was not complete was that it literally processed entre el sometimiento while it needed to be further explained in Korean. Secondly, the other sentence was incomplete because the range of the brackets was ignored and the translation was mixed. In anyhow, it was speculated that if the original sentence had been trimmed in a favorable way to this engine, the result *could* have been improved.

Sentence 6-12

ST	Quienes mantienen un empate técnico son PP y Ciudadanos, los primeros con un 18,66% de voto y el segundo con un 18,55%, que se traduce en idéntico número de representantes en el Parlamento andaluz: una horquilla que va de los 20 a los 22.

GT	기술적 인 관계를 유지하는 사람들은 PP와 시민이며, 18.66 %의 투표권을 가진 첫 번째 사람과 18.55 %의 사람인 두 번째 사람은 안달루시아 의회 의원과 같은 수의 사람으로 번역됩니다. 머리말은 20 22시에
BT	*The people who maintain the technical relationship are the PP and the citizen, the first person with 18.66% of the vote and the second person with 18.55% will be translated into the same number of people as the Andalusian parliamentarians. The preamble is at 20*

Sentence 6-12 was composed of 44 words, and it earned 1.67 fluency score. This sentence was also an identical case to Sentence 6-11 where one source sentence was processed in two target sentences, one of which was structurally incomplete. While Sentence 6-12 also had a complicated structure, this time the main obstacle seemed a colon (:). The machine could not connect the segments before and after the colon. Moreover, the low quality of the output was also due to the part: "[…], los primeros con un 18,66% de voto y el segundo con un 18,55%, […]", where the translation could not identify los primeros (the first) with PP and el segundo (the second) with Ciudadanos, but more importantly, it was not smoothly involved in the whole sentence as much information of the source text was alluded and omitted, i.e. con un 18,55% (de voto), meaning "with 18.55% (of the vote)" or los primers (están) con 18.66%, meaning "the first (party is) with 18.66%". Thus, it was speculated that omitted information could give critical damage to the output.

Sentence 6-13

ST	En los primeros dos años de presidencia de Trump, las dos cámaras del Congreso tenían mayoría republicana, lo que no ha sido suficiente para que Trump impulsara puntos centrales de su agenda, como el desmantelamiento de la reforma sanitaria de Barack Obama o la financiación del muro con México, la estrella de su campaña.
GT	대통령 트럼프의 첫 2 년 동안, 의회의 두 집은 트럼프 충분하지 않은 공화당의 대부분은, 예를 들면 건강 개혁 버락 오바마의 해체 또는 벽 자금으로, 그의 의제의 중심 포인트를 재촉했다 그의 캠페인의 스타 인 멕시코와.
BT	*During the first two years of President Trump's two houses of Congress, Trump has urged the central point of his agenda, most of the Republicans not enough, such as the breakup of the health reform Barack Obama or wall funding, Wow.*

Sentence 6-13 was composed of 54 words (fluency score = 1.67). Although to find a reason for errors in NMT was known to be impossible, this sentence had

a highly complex structure as the previous sentences did, with new information attached to the main sentence with commas. The first challenge happened before […], lo que […] (, which was). The second challenge was to capture the relations of the segments among como el desmantelamiento de la reforma sanitaria de Barack Obama and the parallel structure of el desmantelamiento […] and la financiación del muro con México. The third challenge was the scope of the last segment […], la estrella de su campaña. This sentence was a case where GT clearly failed to capture the right range of each segment.

Sentence 6-14

ST	desde la profunda división del país en dos bloques al bandazo del Brexit duro al blando y la inestabilidad política, pasando por la voladura de las líneas rojas exhibidas por Londres frente a Bruselas o la incertidumbre, letal para la economía.
GT	브뤼셀 앞에서 런던이 전시 한 빨간 선이 폭발하거나 경제에 치명적인 불확실성으로 인해 브리짓이 강경 한 브리짓 (Br Brexit)이 부드럽고 정치적으로 불안정한 상태에 이르기까지 두 블록에 걸친 국가의 깊은 분파에서 일어났습니다.
BT	*The explosion of London's red line in front of Brussels or uncertainty in the economy caused Brigitte's brutal brigit to rise to a soft, politically unstable state in a two-block deep state of the country.*

Sentence 6-13 was composed of 41 words (fluency score = 1.67). It was, in fact, surprising that GT had a complete sentence while the source text itself was incomplete. In that sense, the possible reason for the low fluency score was due, in part, to the incompleteness of the source sentence. However, further investigation was necessary to verify this speculation. Meanwhile, the sentence featured many parallel segments whose relations were quite vague, not to mention some fixed expressions such as Brexit duro (hard Brexit), Brexit blando (soft Brexit) and líneas rojas (red lines). The literal translation of those items led to poor comprehension of the sentence, as well.

Sentence 6-15

ST	La popularidad del presidente mejora cuando deja los exabruptos de lado y consigue resultados, como la bajada de impuestas o la confirmación de Brett Kavanaugh como juez del Tribunal Supremo.
GT	부시 대통령의 인기는 그가 버드 대금을 제쳐두고 떠날 때 개선되고 세금 감면 또는 대법원 판사 브렛 카바 너 (Brett Kavanaugh)의 확인과 같은 결과를 얻는다.

BT Bush 's popularity gets improved when he steps away from Bud, resulting in tax reductions or confirmation by Supreme Court Justice Brett Kavanaugh.

Sentence 6-15 was composed of 30 words (fluency score = 1.67). The sentence length was relatively short, and the sentence structure was quite easy to understand from the human perspective. The first challenge was to recognize the relations of the two phrases before and after […], como […], which was used to give examples of the previous phrase of consigue resultados but translated as a coordinate clause of *and*. The incorrect word choice also influenced on the poor comprehension of the sentence, i.e. exabruptos (버드 대금: Bud payment > 공격적인 발언 harsh remarks) and confirmación (확인: confirmation > 임명: nomination).[52]

Sentence 6-16

ST De forma divertida, el partido más euroescéptico extremista húngaro es antieslovaco y el eslovaco, antihúngaro;

GT 재미있는 방법으로, 대부분의 Eurosceptic 극단적 인 헝가리 당은 반대로 슬로바키아어이고 반대로 헝가리 인 슬로바키아어이다;

BT Interestingly, most Eurosceptic extreme Hungarian parties are Slovak, which is anti-Slovak and anti-Hungarian;

Sentence 6-16 was composed of 15 words, the shortest one (fluency score = 1.67). This sentence was a good example of how originally omitted segments were hard to be generated in the given engine. The complete form of the phrase [...] y el eslovaco, antihúngaro (the Slovak, anti-Hungarian) was "y el (partido más euroescéptico extremista) eslovaco (es) antihúngaro (the most Eurosceptic extreme Slovak party is anti-Hungarian)". Secondly, the polysemy of nationality adjectives was problematic. GT considered all such adjectives in the aforementioned segment as a noun: translating antieslovaco, eslovaco and antihúngaro to 반대로 슬로바키아어 (Slovak language in reverse), 슬로바키아어 (Slovak language) and 반대로 헝가리인 (a Hungarian in reverse).

Sentence 6-17

ST la investigación del supuesto complot de la campaña de Trump con Rusia, cuyos resultados no tardarán en llegar.

[52] The alternative translation suggested in this paragraph is one of many possible expressions.

GT 주장 음모 캠페인 트럼프와의 조사를 그의 결과는 오래 가지 않을 것이다.

BT *The results of the investigation with Trump and the assassination plot campaign will not last long.*

Sentence 6-17 was composed of 18 words (fluency score = 1.67). The main obstacle seemed that the source text was not complete, and it affected the performance because the structure of GT was correct until it reached [...], cuyos [...]. GT regarded the segment before cuyos as the objective and left alone without connecting to the following clause, which resulted in awkward word order.

6.1.2. Adequacy Scoring

After reporting the adequacy score of GNMT in Chapter 6.1.2.1, the adequacy score was analyzed in terms of its correlation with sentence length in Chapter 6.1.2.2. Such quantitative analysis was further detailed in Chapter 6.1.2.3 with qualitative analysis.

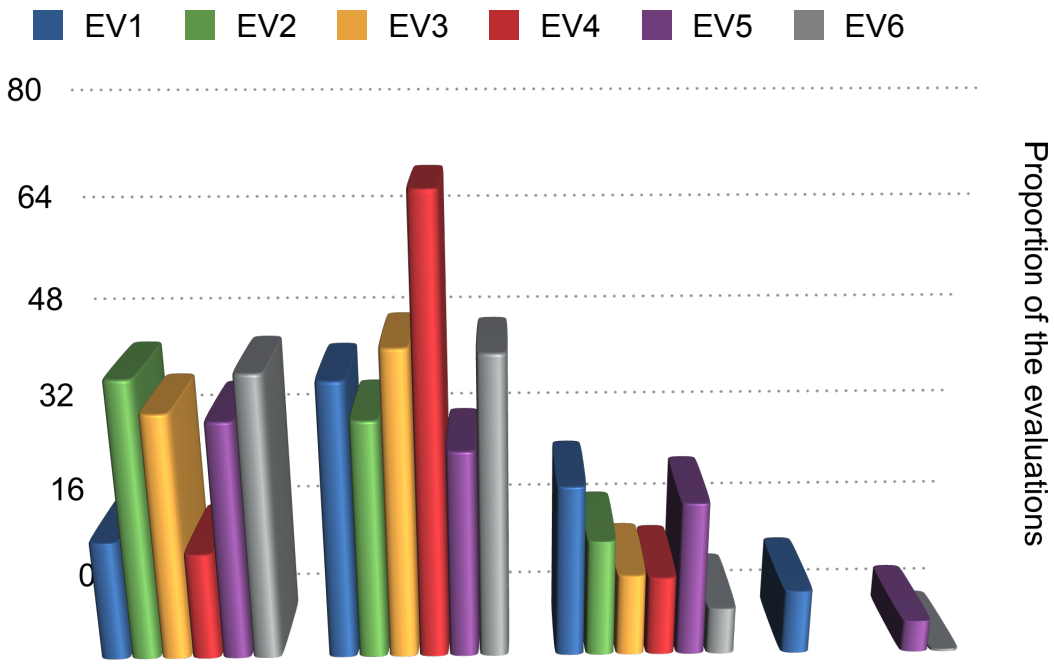
6.1.2.1. Adequacy Score

	EV1	EV2	EV3	EV4	EV5	EV6	**mode**	**mean**
A	2.71	3.03	3.26	3.04	3.36	3.25	**3**	**3.11**
%	67.75%	75.75%	81.50%	76.00%	84.00%	81.25%	**75**	**77.75%**

Table 6.4: Adequacy score of GNMT (*A* = Adequacy).

A mean average of an adequacy score of GNMT was 3.11 of 4, equivalent to 77.75% of adequacy while a mode average scored 3 (75%). It was a slightly lower level than the fluency score (78% and 100%). Table 6.4 displayed the adequacy score of the evaluators. It ranged from a minimum of 67.75% to a maximum of 84%. Such a score meant that about 77.8% of the contents of the source text was conveyed in the translation outputs. In other words, about 23% of the contents were missing in GT.

The score distribution of all editors was detailed in Figure 6.3 and Table 6.5. First of all, the graph showed that the proportion of evaluations was also higher in the positive scores, like that of fluency scoring did. For instance, Score



	4-Everything	3-Most	2-Little	1-None
EV1	18.97	43.48	27.27	10.28
EV2	43.87	37.55	18.58	0.00
EV3	38.74	48.22	13.04	0.00
EV4	17.00	70.36	12.65	0.00
EV5	37.55	32.81	24.51	5.14
EV6	44.66	47.43	7.51	0.40
Avg	**33.47**	**46.64**	**17.26**	**2.64**

Figure 6.3: Distribution of adequacy scores per evaluators and their average. (unit: %)

Table 6.5: Distribution of adequacy scores per evaluators and their average. (unit: %)

3-Most and 4-Everything occupied on average 80.11% of all segments, showing that about 80% of the sentences in the dataset *almost* contained all meaning of the source text in the translation. Meanwhile, Score 1-None obtained the lowest proportion throughout all evaluators (on average 2.64%), and moreover, EV2, EV3, and EV4 did not mark any sentences with Score 1. It was a positive sign.

As previously witnessed, the judgment of EV4 was largely different from the others in adequacy scoring, as well, with a noticeably lower proportion of Score 4 but a higher proportion of Score 3. Therefore, the average of the adequacy score for five evaluators without EV4 was again additionally computed, which was given in Table 6.6. The exclusion of EV4 resulted in a

dissimilar average, with about 3% point increase in Score 4 and 4% in Score 3. According to this new average, it could be inferred that about 37% of the dataset maintained the contents of the source text *perfectly* in Korean sentences.

Contrary to a positive side of the results addressed previously, there was still a long way for the given engine to go considering that on average 2.6% of the dataset was evaluated as conveying *none of the source text*, which was equivalent to 6.6 sentences. The judgments of the evaluators could vary, but the fact that up to 26.8 sentences (according to EV1) deal entirely different contents raised a serious question about the reliability of the given engine.

	Score 4	Score 3	Score 2	Score 1
Avg	33.47	46.64	17.26	2.64
Avg*	**36.76**	**41.9**	**18.18**	**2.64**

Table 6.6: New average adequacy score that excluded EV4's score. (unit: %)

6.1.2.2. Correlation with Sentence Length

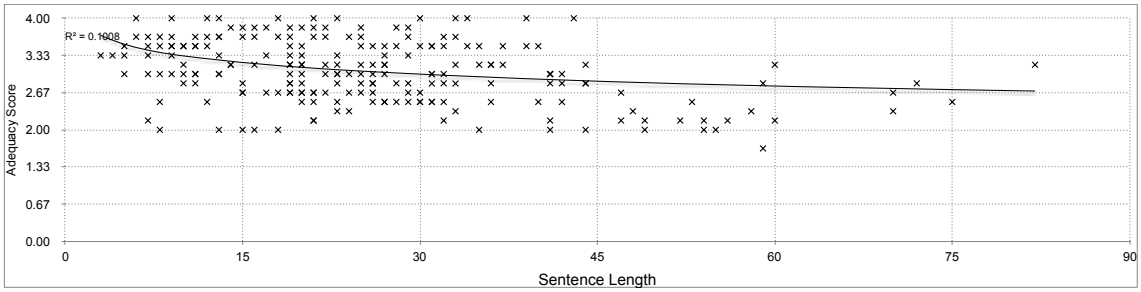


Figure 6.4: Correlation of adequacy score and sentence length graphically shown with a logarithmic line.

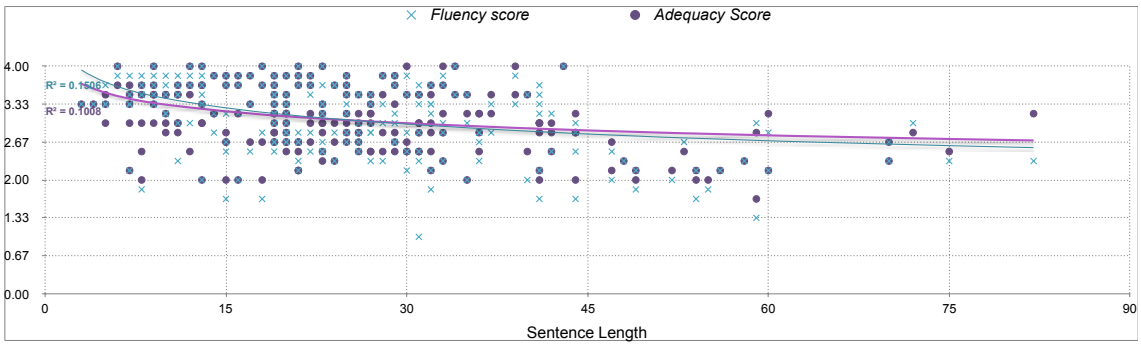


Figure 6.5: Distribution of the fluency and adequacy scores in relation to the sentence length.

The adequacy score was further analyzed to detect if there was a correlation between the adequacy score and a sentence length. The adequacy score of all evaluators was averaged on a sentence basis and organized by the sentence length in Figure 6.4. Figure 6.4 described the proportion of each score, along with the correlation of the two parameters in a logarithmic line. Although the individual scores did not exhibit a clear-cut tendency as did the fluency score, it also seemed that shorter sentences tended to earn more higher scores than longer sentences. Moreover, the graphical distribution manifested that the high-scored sentences were not detected in the sentences of over 45 words. The tendency of the fluency and adequacy score seemed quite similar. The exact match of the two scoring systems was compared in Figure 6.5.

6.1.2.3. Qualitative Analysis

With the quantitative result of 77.6% of adequacy at hand, the data were analyzed qualitatively to look into some of the best and worst cases. The most *adequate* sentence was defined as a sentence with Score 4-Everything throughout all six evaluators. The most *inadequate* sentence, on the other hand, was defined as a sentence with Score 1-None throughout all six evaluators. There were 14 sentences where all evaluators agreed that they were perfectly adequate (Score 4). Some of the examples were given below. Each example was discussed one by one in terms of possible errors in the sentence that were regarded as *not an error* from the evaluators. Such error included not only linguistic matters but also translational matters. (ST = a source text, GT = a system translation of Google Translate, BT = a back translation of GT to English)

Sentence 6-18

ST	El estado de Vermont fue el primero en abrir las puertas de sus colegios electorales, a las 5 de la madrugada, hora local.
GT	버몬트 주 (州)는 현지 시간으로 오전 5시에 처음으로 투표소 문을 열었습니다.
BT	*Vermont has opened its polling place for the first time at 5 am local time.*

Sentence 6-18 was composed of 23 words. The contents of the source sentence were all preserved in GT, except that there was a slightly unnatural segment in the translation. Such issue was raised in the relations of the subject

and the verb in: “The state […] opened the door […]”. Although it was a literal translation of the source text and conveyed all senses, it did not make sense as a Korean expression because an inanimate object could not act. The desirable translation would be “In the state […] the door was opened”. However, it was a problem of *fluency*, not *adequacy*.

Sentence 6-19

ST	Se prevé una tasa de votación extraordinariamente elevada, pues para el pueblo estadounidense es la ocasión de evaluar los primeros dos años de Donald Trump como presidente.
GT	미국 국민의 경우 도널드 트럼프 (Donald Trump)의 첫 2 년을 대통령으로 평가할 기회이기 때문에 매우 높은 투표율이 예상됩니다.
BT	*For the American people, a very high turnout is expected because it is an opportunity to evaluate the first two years of Donald Trump as president.*

Sentence 6-19 was composed of 27 words. The contents were all delivered. The sentence was quite fluent except a slightly awkward position of a phrase: 대통령으로 (as a President). In GT, the order of the translation was “the first 2 years as a President” and it seemed acceptable in English, but the bond of this segment was stronger with Donald Trump than to evaluate. Thus, a desirable position of the segment was before next to Donald Trump as in 도널드 트럼프를 대통령으로써 […] (Donald Trump as a Present) or 대통령으로써 도널드 트럼프를 […] (as a President Donal Trump). However, this was a matter of *fluency*, again.

Sentence 6-20

ST	Pero aparte de la cuestión de tener una regulación adecuada para evaluar el uso indebido de los fondos europeos para actividades políticas, el punto principal es poder analizar de manera eficiente la forma en que utilizan los Fondos de la UE.
GT	그러나 정치 자금에 대한 유럽 자금의 오용을 평가할 적절한 규제가있는 문제는 별개로 EU 자금 사용 방법을 효율적으로 분석 할 수있는 것이 주요 포인트입니다.
BT	*However, apart from the issue of adequate regulation to assess the misuse of European funding for political funds, apart from that, the main point is to be able to analyze efficiently how to use EU funds.*

Sentence 6-20 was composed of 41 words. On a hypothesis that the adequacy score should be affected by mistranslation, this sentence seemed inappropriate

to be judged as a top-scored sentence in the adequacy scoring. Firstly, there were some expressional issues, i.e. aparte de (apart from) whose translation was 문제는 별개로 > 문제**와**는 별개로. More importantly, there was a mistranslated segment: actividades políticas (political activities) whose translation was 정치 자금 (political funds). It was, thus, surprising that all evaluators marked the highest score on this sentence despite these errors.

All in all, the noticeable point from the in-depth study of the 14 best sentences was that all sentences were in perfect word order (except the previous two cases of Segment 18 and 19), which could be challenging in the *es-ko* pair. Furthermore, the successful performance of the given engine was witnessed in the sentences irrespective of a sentence length. Although some problems were discussed in terms of awkward expressions and mistranslation, somehow they were not reflected in the evaluation score.

In the case of the worst sentences in the adequacy score that were marked with Score 1-None throughout the evaluators, 13 cases were detected. The worst case obtained 1.67 score (of 4), displaying a 41.75% adequacy. The rest of the sentences earned identically 2 score (50%). Some of the most distinctive sentences were analyzed one by one focusing on the omitted contents. The extensive error analysis is prepared in Chapter 7.

Sentence 6-21

ST	«No existe un solo partido euroescéptico, existen veintitantos, cada uno con su sensibilidad.
GT	"유로 이탈리즘 당이 하나도 없으며, 감성이있는 스무 명의 사람들이 있습니다.
BT	*"There is not one Euro break, and there are 20 people with emotions.*

Sentence 6-21 was composed of 13 words and it achieved the lowest adequacy score of 1.67. Firstly, the overall meaning of the sentence was misguided because while the source sentence meant that “there is not one but many”, GT read as “There is no such party, and […]”. Secondly, GT did not generate properly the omitted words after veintitantos (twenty-something): GT was “twenty people”, but considering the omitted information (partido euroescéptico), a desirable translation was “twenty-something parties”. Secondly, sensibilidad was mistranslated to 감성 (sensitivity or with emotion) while the meaning of the

source segment was more like 성향 (*political* inclination or opinion). Lastly, the part of cada uno (each) was not preserved.

Sentence 6-22

ST No puedes empatar.

GT 당신은 묶을 수 없습니다.

BT *You can not bind.*

Sentence 6-22 was composed of 3 words, the shortest case (adequacy score = 2). The sole error was a wrong word choice of empatar (to tie). It is assumed that the influence of the main verb in such a short sentence had a great impact on the judgment.

Sentence 6-23

ST O sea, juegas al Brexit.

GT 즉, Brexit을 연기합니다.

BT *In other words, we play Brexit.*

Sentence 6-23 was composed of 5 words (adequacy score = 2). The translation of this sentence seemed to be largely dependent on the context, and that was why the adequacy score was low. Firstly, the main verb was selected incorrectly: while GT was “to act”, a desirable translation would be “to play”. Secondly, it was an imperative sentence, but GT was a plain sentence with no visible subject. Lastly and more importantly, more information was required to express the hidden meaning behind the sentence. That is, a better option than just putting it as “Play the Brexit” was necessary by reading between the lines.

Sentence 6-24

ST Entonces, ambos condados utilizaban papeletas en las que el votante hacía un agujero dentro de un círculo.

GT 그런 다음 두 카운티는 유권자가 원안에 구멍을 뚫은 투표 용지를 사용했습니다.

BT The two counties then used voter ballots punched through the circle.

Sentence 6-24 was composed of 17 words (adequacy score = 2). This sentence was quite distinctive in that there were no errors in a one-on-one comparison. The only issue was that the relations between the main clause and the

subordinate clause (en la que) were miscalculated. According to the source sentence, the original meaning was that “both counties used ballots and the voters should punch on them”, but in GT, the subordinate clause modified papeletas so that the sentence meant that “ballots that the voters punched on”. However great the impact of such error might be, it was dubious that this sentence could be scored as bad as Sentence 6-20, for example, that had multiple errors.

Sentence 6-25

ST	Así, la libra ya ha perdido el 12% de su valor y Reino Unido será el país europeo que menos crezca en 2019 y 2020. un agujero dentro de un círculo.
GT	따라서 파운드는 이미 가치의 12 %를 잃었으며 영국은 2019 년과 2020 년에 성장하지 않는 유럽 국가가 될 것입니다.
BT	*So the Pound has already lost 12% of its value and the UK will be a European country that does not grow in 2019 and 2020.*

Sentence 6-25 was composed of 42 words (adequacy score = 2). The overall meaning was well conveyed except two parts. Firstly, an awkward expression was observed in the first phrase of la libra ya ha perdido el 12% de su valor. While GT produced a literal translation of this sentence, the inanimate subject “pound” could not lose in a Korean expression because it could not act. Thus, it was more natural to say “the value of the pound has dropped”. Secondly, there was a mistranslation of menos. While the original sentence meant “The UK will be the least developed […] country”, GT meant that “The UK did not grow in such period”. Again, compared to other sentences with multiple errors, this sentence featured a small number of errors.

Sentence 6-26

ST	«No creo que el Parlamento Europeo esté ahora mejor preparado porque el sistema siempre ha funcionado.
GT	"나는 시스템이 항상 일 때문에 유럽 의회는 지금 더 잘 준비라고 생각하지 않습니다.
BT	*"I do not think the European Parliament is better prepared now because the system always works.*

Sentence 6-26 was composed of 16 words (adequacy score = 2). In this case, the translation committed many errors compared to a simple structure the

source sentence had. The source sentence was composed of a main clause and a subordinate clause of porque (because). However, GT translated until siempre in one clause and left out the verb part ha funcionado (has worked). Influenced by such a mistake, there were two subjects in GT, which caused a misunderstanding. Moreover, there was a wrong expression when translating esté preparado (is prepared) to 준비라고 (is preparation). It could be revised as 준비가 됐다고.

Sentence 6-27

ST	la investigación del supuesto complot de la campaña de Trump con Rusia, cuyos resultados no tardarán en llegar.
GT	주장 음모 캠페인 트럼프와의 조사를 그의 결과는 오래 가지 않을 것이다.
BT	The results of the investigation with Trump and the assassination plot campaign will not last long.

Sentence 6-27 was composed of 18 words (adequacy score = 2). To read the target sentence alone, it felt like a word salad. BT did not represent such errors, but in GT, many nouns were listed separately without any postpositions until the first five words[53] so that it sounded like: argument plot campaign investigation with Trump. Not to mention that the part of con Rusia (with Russia) was omitted, there was a mistranslation of no tardarán en llegar (they will not take long to come out): 오래 가지 않을 것이다 (they will not last long). In many aspects, the translation featured many errors.

Sentence 6-28

ST	Todo explica que Trump apenas se haya quitado el traje electoral este otoño.
GT	이번 가을에 트럼프가 그의 선거를 막 제거했다고 모든 것이 설명됩니다.
BT	Everything is explained that Trump has just removed his election this fall.

Sentence 6-28 was composed of 13 words (adequacy score = 2). The structure of this sentence was deformed due to a mistranslation of the expression Todo explica que (everything explains that). It led to an incorrect connection between

[53] The word in this context means an item counted by a spacing. The Korean counting system is not applied here. More information about the different Korean counting system is provided in Chapter 5.1.

the two clauses in GT. Furthermore, the source segment said se haya quitado el traje electoral (has taken off the election costume), which was translated into “removed his election”.

Sentence 6-29

ST	En términos de ingresos fijos, cada eurodiputado recibe un salario de alrededor de 6.600 euros netos al mes, 4.400 para gastos de oficina sin justificar y alrededor de 24.000 para asesores.
GT	고정 수입 측면에서, 각 MEP는 월 6,600 유로의 순이익을, 정당화되지 않은 사무 경비는 4,400, 고문은 약 24,000의 급여를받습니다.
BT	*In terms of fixed income, each MEP receives a net profit of 6,600 euros per month, unwarranted office expenses of 4,400 and torture of approximately 24,000 salaries.*

Sentence 6-29 was composed of 31 words (adequacy score = 2). GT was inconsistent with numeric expressions such as not specifying the unit “euros” after the figures: 4,400 > 4,400유로; 24,000 > 24,000유로. A great mistranslation was thought to be the subject of the sentence, eurodiputado, to MEP. MEP, according to Collins Concise English Dictionary, was a British abbreviation of Member of the European Parliament.[54] And the Korean equivalence was EU 국회의원. Not knowing who this sentence was dealing with provoked a confusion, according to EV2. Another mistranslation was the translation of asesores. GT translated it as “The advisers get a salary”. Secondly, there was a structural error: in the source text, the three figures (6600, 4400, 24000) constituted fixed salaries (ingresos fijos), but the way GT organized the segment seemed that only 24000 euros were the fixed salaries. Thirdly, there was an expressional error: there was no equivalent noun expression for salario neto (net earnings), which caused unnaturalness. While GT translated it as 순이익 (net profit), one possible option for this expression was 세전 급여 (pre-tax salary). Similarly, another expressional error was that the verb justificar (to justify) was not used as in a dictionary (정당화하다) in Korean in this context. It seemed more natural to use 증빙이 필요 없는 (that does not require evidential documents).

[54] https://www.wordreference.com/definition/MEP

Sentence 6-30

ST	[SUB-HEADLINE] La encuesta sitúa a la coalición de Podemos e IU en segundo lugar y anticipa un empate técnico entre PP y Ciudadanos
GT	이 설문 조사는 Podemos와 IU의 연합을 두 번째 장소에두고 PP와 Citizens 사이의 기술적 인 관계를 예상합니다
BT	*This survey puts the coalition of Podemos and IU in the second place and expects a technical relationship between PP and Citizens*

Sentence 6-30 was composed of 22 words except for the sub-headline mark (adequacy score = 2). The structure of the translation followed the source text well, but the main reason of the low adequacy score seemed to be due to wrong choices of segundo lugar (the second position) and empate técnico (technical tie). Firstly, GT translated “the second position” as “in the second place” in a spacial sense. The correct meaning of the segment would be “the second rank”. Secondly, the term ‘technical tie’ was a fixed expression used in boxing, but GT translated it literally and incorrectly: 기술적인 관계 (technical relationship). Additionally, the Spanish party Ciudadanos was translated into English as Citizens.

In summary, many of the most inadequate sentences had multiple mistranslations. Some mistranslation seemed to affect critically on human comprehension such as a mistranslation of the main subject. It was also noted that many sentences were decomposed into separate segments and not connected properly to the sentence. While there were some cases that raised a question about the judgment of the evaluators, the qualitative analysis helped to give an idea about what an inadequate sentence was.

6.2. Segment Ranking

The results of a segment ranking comparison of GNMT (GT), an NMT system of Kakao i (KT) and human translation (HT) on a 3-rank scale was analyzed. The evaluation was carried out on 253 sentences by six evaluators. The analysis was approached from a quantitative and qualitative stance, which were distributed in each sub-section. Chapter 6.2.1 addressed a total ranking score of each engine and their distribution. Going further, Chapter 6.2.2 approached the result by machine type and Chapter 6.2.3 approached it by ranking choice.

6.2.1. Ranking Score

The ranking score was analyzed quantitative and qualitatively in Chapter 6.2.1.1 and 6.2.1.2.

6.2.1.1. Quantitative Analysis

	GT	KT	HT
R	1.8	1.92	2.67

Table 6.7: Ranking score (R) (in a range of 0 - 3) of the three engines.

From the segment ranking evaluation, the ranking score was computed and given in Table 6.7. The score was based on a calculation where each system that was selected as the 1st, 2nd, and 3rd, the system marked 3, 2 and 1 ranks respectively. With the score obtained, the system normalized the score and averaged by the number of sentences. GT obtained 1.8 of 3 ranking score, meaning 60% of preference by the evaluators. As previously mentioned in Chapter 3.4.3, the ranking comparison evaluation in this thesis was based on an absolute system. KT earned 64% and HT, 89%.

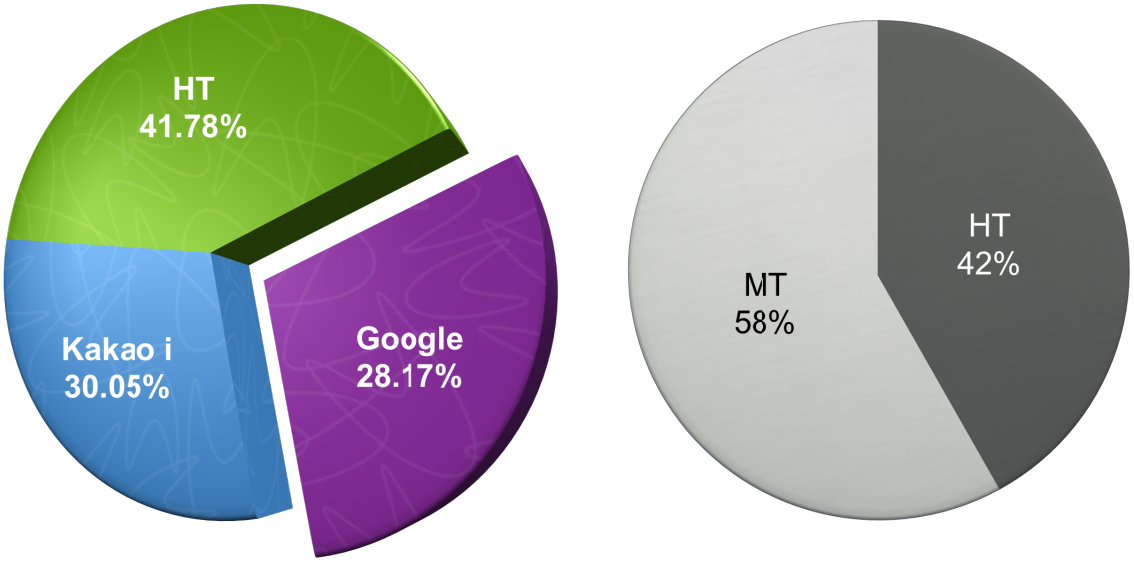


Figure 6.6: (A) Proportion of the three engines which were selected as the best engine in segment ranking. (B) Proportion of (A) distributed by MT versus HT.

To address the result relatively, Figure 6.6-A displayed that the very subject in question, GNMT, was ranked as the lowest candidate of all, with 28.17%. There was a minor gap between KT and GT of about 2%. HT was the first to be selected as the best engine[55] over the two NMT systems with 41.78%. The outperformance of HT might be an expected result. The unexpected part, however, was that if a distinction was drawn between human (HT) versus machine (MT), MT won over HT by 16.4%, as shown in Figure 6.6-B. Moreover, it came to our attention that the score of HT, which was 89% of preference, was quite lower than was expected. It was speculated that such phenomenon was due to a misjudgment of the evaluators or a mistranslation of the reference translation. However, an in-depth study was necessary to assure the relations. As such, the given issue was analyzed further in both quantitative and qualitative ways.

It was hypothesized that those segments in HT that were ranked as the 3rd caused the two previously-mentioned incidents: i) a low ranking score of HT and ii) a lower performance of HT over MT. Table 6.8 displayed the number of 3rd ranks of HT by the evaluators. On average 17 sentences were detected, which was equal to about 7% of the total dataset. Note that these segments did not necessarily mean that HT was the only candidate translation that was marked as the 3rd rank of the three.

	EV1	EV2	EV3	EV4	EV5	EV6	***Avg***
n	33	11	19	13	9	18	***17.17***
%	13.04	4.03	7.70	5.14	3.56	7.11	***6.76***

Table 6.8: Proportion of HT that was rated as the 3rd rank. (n = number)

6.2.1.2. Qualitative Analysis

Such a result showed that there was an inconsistency throughout the evaluators as EV5 had 9 cases while EV1 had 33. It was, therefore, important to approach those sentences from a qualitative perspective. There were 8 sentences that obtained the lowest rank from more than half of the evaluators ($N \geq 3$). Some of such cases were given below.

[55] We used the term ‘engine’ as it was a blind evaluation, and the evaluators were not aware that the human translation had been involved.

Sentence 6-31

ST Tuvimos que adaptarnos, sí, definitivamente estamos mejor preparados que antes».

GT 우리는 적응해야했습니다. 예, 우리는 이전보다 확실히 준비가 잘되어 있습니다.
G-BT *We had to adapt. Yes, we are definitely better prepared than before.*
HT **물론 적응해야 된다. 지금 우리는 옛날보다 훨씬 더 준비가 잘 되어 있다."고 밝혔다.**
H-BT *Of course you have to adapt. Now we are much better prepared than before. "*
KT 우리는 적응해야만 했다. 예, 우리는 이전보다 확실히 더 잘 준비되어 있습니다. "

Sentence 6-31 was evaluated as an incorrect translation by EV1 (in a comment), presumably because of the verb tense of the first segment. HT mistranslated tuvimos (we had) into “we have”. The rest of the segments were the same as other translations.

Sentence 6-32

ST «pero las reglas no deberían ser más estrictas para los representantes de determinado color político:

GT «그러나 어떤 정치적 색채의 대표자들에게는 규칙이 엄격 해져서는 안된다.
G-BT *«However, rules should not become strict for representatives of certain political colors.*
HT **"그러나 특정 유색정당 대표들에게는 더 엄격한 규정이 있어서는 안된다.**
H-BT *"But there is no stricter regulation for certain colored party representatives.*
KT 그러나 규칙은 특정 정치적 색의 대표자들에게 더 엄격해서는 안됩니다."

In Sentence 6-32, the main difference of HT from MT was a different structure of the sentence (규칙이 엄격해져서는 […] vs. 엄격한 규정이 있어서는 […]), which was a different way of expressing the same contents. And there was a different word choice of color político (political color) throughout the translations:

ST		color político
GT	정치적 색채	: political color
HT	유색정당	: a party with political colors
KT	정치적 색	: political color

The word choice of GT was a literal translation and correct expression, but that of KT seemed unnatural in the sentence, which was not distinguishable from its English explanation. The word choice of HT was not a literal translation, but from our personal judgment, it sounded more natural and more related to the context. It was dubious why EV1 commented that HT was an incorrect

translation. It could be her misjudgment or a low level of vocabularies.

Sentence 6-33

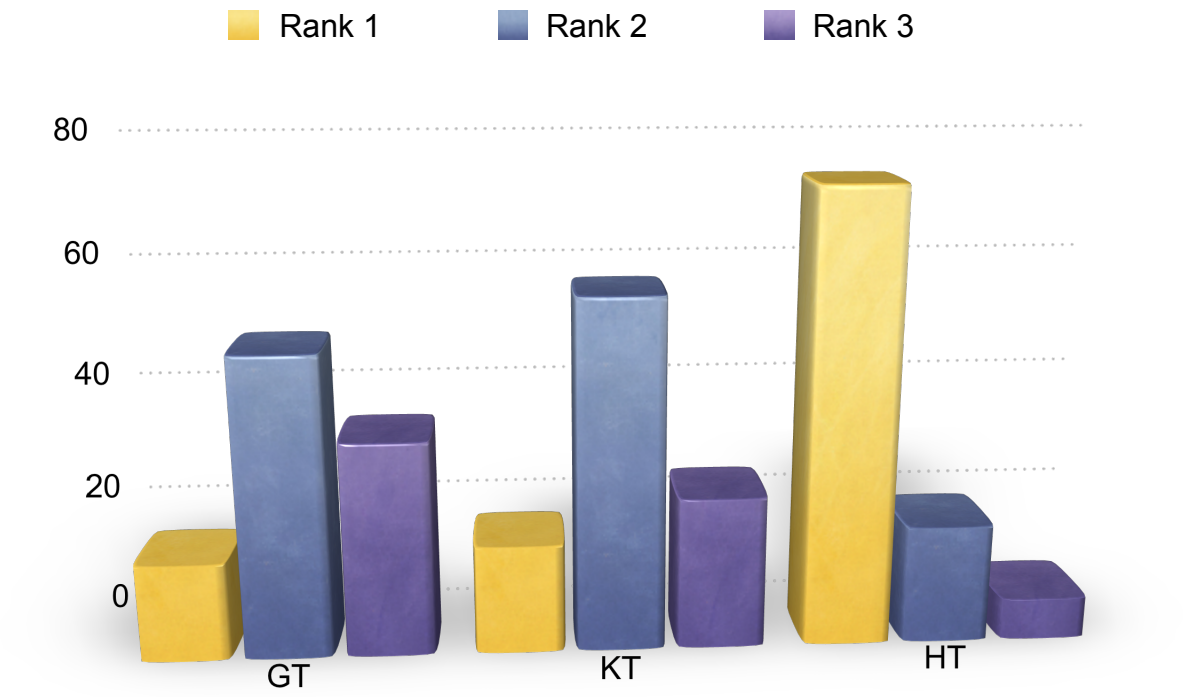
ST	Pero aparte de la cuestión de tener una regulación adecuada para evaluar el uso indebido de los fondos europeos para actividades políticas, el punto principal es poder analizar de manera eficiente la forma en que utilizan los Fondos de la UE.
GT	그러나 정치 자금에 대한 유럽 자금의 오용을 평가할 적절한 규제가있는 문제는 별개로 EU 자금 사용 방법을 효율적으로 분석 할 수있는 것이 주요 포인트입니다.
G-BT	*However, apart from the issue of adequate regulation to assess the misuse of European funding for political funds, apart from that, the main point is to be able to analyze efficiently how to use EU funds.*
HT	**유럽 기금이 정치활동에 불법적으로 사용되는 것을 판단하기 위해 적절한 규제안을 마련하는 것을 떠나서, 핵심은 EU 자금이 어떻게 사용되고 있는지 효과적으로 분석하는 힘이다.**
H-BT	*Apart from providing adequate regulation to judge that the European Fund is illegally used in political activities, the key is to analyze effectively how EU funds are being used.*
KT	그러나 정치 활동을 위한 유럽 기금의 오용을 평가하기 위한 적절한 규제의 문제 외에도 주요 요점은 EU 기금이 사용하는 방식을 효율적으로 분석할 수 있다는 것입니다.

Three evaluators ranked Sentence 6-33 as the lowest quality and EV1 and EV4 commented that HT was a mistranslation. A noticeable point of the three translations was that the structure of GT and KT was almost identical unlike the structure of HT. That is, HT tried to use structures and expressions that sounded more natural as Korean: considering that a nominalization was not preferred in Korean, it changed nouns into verbs such as 오용 (misuse) > 불법적으로 사용되는 (to be used illegally). It was dissimilar from the way MT produced an output, who adapted most of the features to the source sentence so that it looked like a literal translation. This trend was also visible in Sentence 6-30 and 6-31. It was speculated that the evaluators accepted such form of translation as a *correct* translation. In other words, the notion of equivalence was transformed by the influence of frequent use of MT. It was a very interesting observation that could have a great influence not only in MT but also in translation studies, but a more in-depth investigation was required to clarify the phenomenon.

From the three examples, a part of the reasons why HT achieved an underestimated result and why MT outperformed HT was explained. It could be due to errors of HT itself or misjudgment of the evaluators. More importantly, however, it was observed that the evaluators preferred a literal translation to a

liberal translation.

6.2.2. Distribution of Ranking by Machine



	GT			KT			HT		
	Rank 1	Rank 2	Rank 3	Rank 1	Rank 2	Rank 3	Rank 1	Rank 2	Rank 3
EV1	15.87	49.07	35.04	17.72	57.37	24.9	73.84	19.36	6.78
EV2	17.39	62.45	20.15	24.9	67.58	7.5	78.65	17.78	3.55
EV3	15.41	31.22	53.35	15.81	41.5	42.68	80.23	12.25	7.5
EV4	12.25	50.98	36.75	13.43	60.07	26.48	84.58	10.27	5.13
EV5	21.34	47.82	30.83	20.94	56.91	22.31	63.63	29.24	7.11
EV6	16.99	56.12	26.87	18.18	62.84	18.97	77.86	17.78	4.34
Avg	***16.54***	***49.61***	***33.83***	***18.50***	***57.71***	***23.81***	***76.47***	***17.78***	***5.74***

Figure 6.7: Distribution of an average ranking score by machine type.

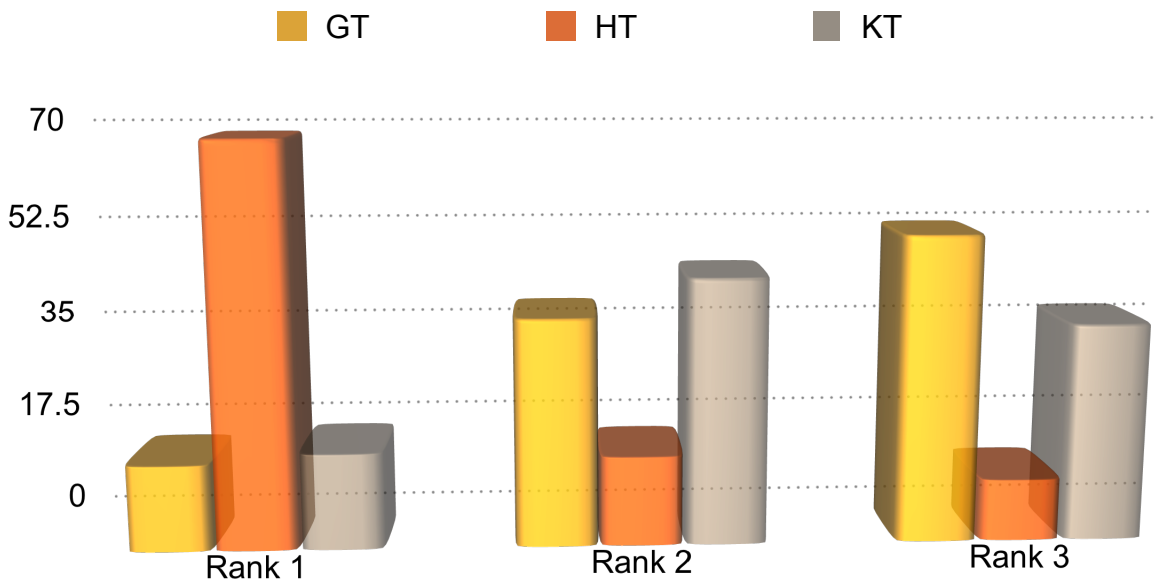
Table 6.9: Distribution of the ranking score by machine type.

The data obtained from the segment ranking evaluation was analyzed by machine type. The analysis manifested a relative position of GT in relation to KT and HT. In Figure 6.7, the averaged distribution of each candidate was graphically displayed. In GT, the highest proportion was Rank 2, with 49.61%. It could be interpreted that half of the output of Google Translate was ranked as

the 2nd. The 1st rank of GT was 16.54%. It was not a remarkable achievement compared to KT (18.5%) or HT (76.47%), but it still meant that GT achieved human parity in about 41.8 sentences.

Table 6.9 showed a detailed score of the ranking scores throughout the evaluators. In the case of GT, the proportion of Rank 1 was consistent throughout the evaluators, with the highest score observed in EV2 and the lowest one in EV1. In terms of Rank 2 and Rank 3, most of the evaluators designated Rank 2 more than Rank 3, except the case of EV3 where Rank 3 was more witnessed than Rank 2.

6.2.3. Distribution of ranking by ranking choice



	Rank 1			Rank 2			Rank 3		
	GT	HT	KT	GT	HT	KT	GT	HT	KT
EV1	14.28	70.00	15.71	35.25	22.18	42.55	48.63	15.00	36.36
EV2	14.37	65.03	20.58	42.24	12.03	45.72	64.55	11.39	24.05
EV3	13.82	71.98	14.18	36.74	14.41	48.83	51.52	7.25	41.22
EV4	11.11	76.70	12.18	42.01	8.46	49.51	53.75	7.51	38.72
EV5	20.14	60.07	19.77	35.69	21.82	42.47	51.31	11.84	36.84
EV6	15.03	68.88	16.08	41.04	13.00	45.95	53.54	8.66	37.79
Avg	***14.79***	***68.78***	***16.42***	***38.83***	***15.32***	***45.84***	***53.88***	***10.28***	***35.83***

Figure 6.8: Distribution of an average ranking score by ranking choice.

Table 6.10: Distribution of the ranking score by ranking choice.

The data obtained from the segment ranking evaluation was analyzed by ranking choice. The focus was shifted from how each candidate translation individually obtained the score in Chapter 6.2.2 to how the ranking was distributed to the candidates in Chapter 6.2.3. The results were given in Figure 6.8 and Table 6.10.

The 1st rank was mostly given to HT with 68.78% while GT obtained the smallest proportion (14.79%). In Rank 2, the biggest pie was taken by KT with 45.84% while GT was slightly behind it with 38.83%. In Rank 3, unfortunately, GT was the most marked candidate of all, with 53.88%. From the result, it was concluded that the largest proportion of Rank 1, Rank 2, and Rank 3 was occupied by HT, KT, and GT, respectively.

6.2.4. Qualitative Approach[56]

This qualitative analysis was mainly focused on the high-ranked and low-ranked cases of GT. Some examples of GT that earned the highest rank score and the lowest rank scores were analyzed below. There were five cases where all evaluators ranked GT as the best sentence and seven cases where they ranked it as the worst sentence.

Sentence 6-34

ST	“No puedes ganar.
GT	**"당신은 이길 수 없습니다.**
G-BT	*"You can not win.*
HT	“당신은 이기지 못한다.
KT	“이길 수는 없어.

Sentence 6-34 was composed of 3 words. The three translations were almost identical, except that KT had a colloquial style and GT used a formal language. The quality of the translation seemed perfect including spacing.

Sentence 6-35

ST	«Pero el populismo no puede ser la solución a estos problemas», advierte.

[56] As Chapter 6.2.2 and Chapter 6.23 deal with the same scores but approach with a different standard, the qualitative analysis of the two chapters is addressed in a lump in this section.

GT **“그러나 포퓰리즘은 이러한 문제에 대한 해결책이 될 수 없다"고 그는 경고했다.**

G-BT *"But populism can not be a solution to this problem," he warned.*

HT 그는 "그러나 포퓰리즘이 이 문제에 해결책이 될 순 없다" 라고 밝혔다.

KT “하지만 포퓰리즘은 이러한 문제에 대한 해결책이 될 수 없다.”.

Sentence 6-35 was composed of 12 words. The three translations had the same vocabularies and structures. The only difference was observed in dealing with the word advierte (warned):

ST	« […] », advierte.	
GT	“[…]“고 그는 경고했다.	: warned
HT	그는 “[…]” 라고 밝혔다.	: claimed
KT	“[…].”.	

From a structural point of view, both GT and HT were correct. However, in the case of KT, the verb was not translated and it had two periods, which was incorrect. From a lexical point of view, the word choice of GT was closer than that of HT to a desirable translation.

Sentence 6-36

ST Mientras en Georgia, la demócrata Stacey Abrams ha prometido tomar acciones legales para garantizar que todos los votos se cuenten correctamente.

GT **조지아 주에서 민주당의 스테이시 에이 브람스 (Stacey Abrams)는 모든 득표가 올바르게 집계 될 수 있도록 법적 조치를 취할 것을 약속했습니다.**

G-BT *Democrat Stacey Abrams in Georgia promised to take legal action to ensure that all votes are counted correctly.*

HT 한편 조지아에서는 스테이시 에이브럼스 민주당 후보가 법적 조치를 취해 모든 표를 올바로 세겠다고 선언했다.

KT 조지아에서 민주당 스테이시 에이브람스는 모든 투표가 올바르게 계산되도록 법적 조치를 취할 것을 약속했습니다.

Sentence 6-36 was composed of 21 words. GT seemed considerably detailed in every aspect. First, it put additional information on the proper noun (Stacey Abrams). Secondly, it processed Georgia as “Georgia State” while others did not. Thirdly, it translated votos (votes) to 득표, which was a more formal vocabulary than the choice of HT (표) or KT (투표). Lastly, it used a formal language in se cuenten (to be counted), translating it to 집계되다 (to total), while the others used 세다 (to count) or 계산하다 (to calculate). Although there was an omission of Mientras (while), the evaluators ranked GT as the best sentence for the above-

mentioned reasons.

Sentence 6-37

ST sigue siendo una democracia».

GT **그것은 여전히 민주주의입니다. "**
G-BT *It is still democracy. "*
HT 이것도 민주주의를 따른다."
KT 그것은 민주주의로 남아 있습니다.

Sentence 6-37 was composed of 4 words. It was a rather interesting case where the evaluators' judgment was out of the context. Although the sentence was short-lengthen and simple, it required a contextual comprehension to translate. Thus, the correct translation that was in line with the context was HT (who said that "it sticks to democracy"), but GT (: it is still democracy) was selected as the best translation.

Sentence 6-38

ST A cambio, Kiev debe conceder a los dos enclaves un sistema de autogobierno.

GT **그 대가로 키에프는 두 영토를 자치 체제로 만들어야합니다.**
G-BT *In return, Kiev must make both territories autonomous.*
HT 그 대가로 키예프 정부는 이 두 지역에 자치정부제도를 부여해야 한다.
KT 그 대가로 키예프는 두 영토에 자치 제도를 부여해야 한다.

Sentence 6-38 was composed of 13 words. This was also the case where the evaluators' judgment was dubious. From many aspects, HT and KT seemed more natural. First of all, in GT a spelling error was detected in Kiev: 키에프 > 키예프. Moreover, a desirable translation of Kiev in this context was "Russia" or "the government of Kyiv", but GT could not achieve it. Secondly, due to the characteristics of the verb as an inanimate subject, the structure of GT (: Kyiv should make […]) was incorrect. Thirdly, spacing errors were detected.

In summary, irrespective of the sentence length, GT proved an ability to process a translation on a human level, albeit few. Some segments were of exceptional quality that was even more accurate and detailed than HT and with a formal language. Others raised a question as to why GT was though to be the best option despite the errors.

In a negative side, there were seven sentences in GT that were considered as the worst translation. Most of the sentences were composed of more words than the best-ranked cases, with an average of 29.3 words, ranging from a minimum of 16 words to a maximum of 58 words. Some of the relevant examples are analyzed below.

Sentence 6-39

ST También se encaminan a recuentos las elecciones a senador por Arizona y a gobernador de Georgia.

GT **애리조나 상원 의원과 조지아 주지사 선거 역시 중요하게 고려된다.**
G-BT *Arizona Senator and Governor of Georgia elections are also considered important.*
HT 또한 애리조나 주 상원의원 선거 및 조지아 주지사 선거도 재검표에 들어간다.
KT 애리조나 주 상원 의원과 조지아 주지사 선거도 재검 표로 진행됩니다.

Sentence 6-39 was the shortest sentence of the seven cases. Despite the short length and the simple structure, the way GNMT processed the sentence was fluent but inadequate, which showed the infamous trait of many NMT engines. Firstly, the main verb was se encaminan (to head to …), whose GT was “to be considered important”. HT and KT, on the other hand, translated it correctly. Secondly, in the source sentence, there were two types of elections: for Senators and for governor. In GT, it read as “Senator and governor election”, which was not incorrect but more clarification would be of help, such as repeating “election” twice. Thirdly, GT did not consider recuentos (re-count) at all.

Sentence 6-40

ST Esas elecciones fueron polémicas desde el primer momento porque el candidato republicano, Brian Kemp, como secretario de Estado de Georgia, es el responsable de gestionar comicios y los demócratas le acusaron de tomar medidas para restringir el voto de la población negra, que iba a ser clave para Abrams, que sería la primera gobernadora afroamericana de EE UU.

GT **그 선거는 공화당 출신 후보 인 브라이언 켐프 (Brian Kemp)가 조지아 국무 장관으로 선거 운영에 책임이 있고 민주당이 흑인 인구의 투표를 제한하는 조치를 취했다고 비난했기 때문에 처음부터 논란이 많았다. 미국 최초의 아프리카 계 미국인 총재가 될 아브람에게 열쇠.**

B-GT	*The election was controversial from the outset because Republican candidate Brian Kemp accused Georgia of being the secretary of state and responsible for running the elections and the Democrats taking steps to limit the voting of the black population. The key to Abram to be America's first African American president.*
HT	공화당 후보인 브라이언 켐프 조지아 주 국무장관이 선거 관리를 담당하고 있어 이번 선거는 시작부터 논란이 일었다. 민주당은 그가 미국의 첫 흑인 주지사를 노리고 있는 에이브럼스 후보에게 매우 중요한 흑인 표를 제한하기 위해 술수를 쓴다고 비난했다.
KT	그 선거는 조지아 주 국무 장관인 브라이언 켐프 공화당 후보가 선거 관리 책임자이며 민주당이 미국 최초의 아프리카계 미국인 주지사가 될 아브람스의 열쇠가 될 흑인 인구의 투표를 제한하기 위한 조치를 취했다고 비난하면서 처음부터 논란의 여지가 있었다.

Sentence 6-40 was the longest sentence of GT that was regarded as the worst translation. Just by comparing the completeness of the sentence, GT would be ranked as the lowest as it was the only incomplete one, not to mention that the word order was incomprehensible (partly due to the incomplete sentence structure). Besides such point, it seemed like GT dealt well with the source sentence until it reached "[…], que iba a ser clave para Abrams, […]". Except that it did not connect the above-mentioned clause correctly to the main clause, GT seemed quite good.

6.3. Post-Editing Time & Effort

EV	Throughputs		TIme		P ratio
	T	PE	*T*	PE	
EV1	545	740	6.35	3.99	1.35
EV2	679	1,042	5.10	2.83	1.53
EV3	410	461	8.44	6.41	1.12
EV4	1,515	1,923	2.29	1.54	1.26
EV5	897	1,355	3.86	2.18	1.51
EV6	465	665	7.46	4.45	1.43
Sum	-	-	33.5	21	-
Avg	**751**	**1,031**	**5.58**	**3.6**	**1.37**

Table 6.11: Result of the post-editing evaluation contrasting words per hour (WPH) and time (unit: hour) of translation and post-editing.

The post-editing was employed in our experiment to measure the usefulness of Google Translate in post-editing by means of post-editing time (Chapter 6.3.1) and post-editing efforts (Chapter 6.3.2) composed of temporal and technical efforts. Moreover, the HTER score was calculated based on post-edited results in Chapter 6.3.3. Note that the analysis was on a quantitative basis. A qualitative analysis of post-editing data was performed in detail through error

analysis in Chapter 7.

6.3.1. Post-Editing Time

Time was an objective measurement of the operational evaluation. This section reported on the time taken for post-editing and translating from scratch by six novice post-editors. The two tasks were requested for half of the segments each (127 sentences for post-editing and 126 segments for translation from scratch). The results of the experiment were given in Table 6.11. The total time spent for post-editing (named as PE in this section) by the six post-editors was 77,089 seconds (21 hours) and the time taken for translation from scratch (named as *T* in this section) was 120,622 seconds (about 33.5 hours). About 12.5 hours were reduced by performing post-editing. The post-editors introduced 751 words per hour (WPH) in *T* while they added 220 words more per hour in PE.

To analyze the data by post-editor, the throughputs and time were given in the same table. The average time taken for *T* was 5.58 hours and 3.6 hours for PE, with a gap of 1.98 hours per person. Time for PE exhibited a drastic difference per editor, ranging from 1.54 (EV4) - 6.41 hours (EV3). In the same context, WPH ranged from 461 (EV3) to 1,923 words (EV4). It was, in fact, a surprising finding that the productivity of PE of the post-editors who had a similar background varied drastically, with one person post-editing 3.61 times faster than the other. The similar tendency was also observed in translation: the minimum WPH were 410 while the maximum WPH were 1,515.

The result showed that the post-editors worked on average 2 hours faster when post-editing the dataset than translating it, but such a figure could not explain how much faster the post-editing was in an hour than translating. Thus, the productivity of post-editing was computed based on *P* ratio (post-editing ratio), which was a key indicator in the MT evaluation established by words per hour of translation in relation to words per hour of post-editing. In Table 6.11, the *P* ratio was on average 1.37 in a range of 1.12 (EV3) to 1.53 (EV2). It indicated that post-editing was 37% faster than translating from scratch, ranging from a minimum of 12% to a maximum of 53%. In other words, there was a 37%

productivity gain.

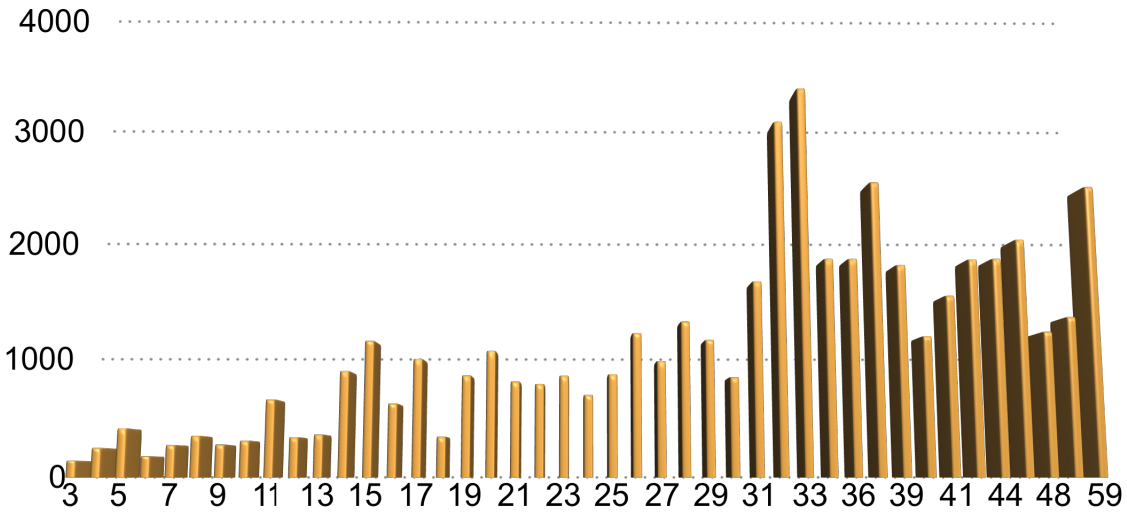


Figure 6.9: Correlation of post-editing time (unit: seconds) and sentence length.

6.3.2. Post-Editing Efforts

The post-editing efforts were analyzed in two branches: temporal efforts and technical efforts. The temporal efforts referred to the amount of post-editing efforts measure by time and throughputs in relation to sentence length. The technical efforts, on the other hand, were measured by an edit distance along with the HTER score. The HTER score was provided separately in Chapter 6.3.3.

6.3.2.1. Temporal Efforts

Previously in the pilot study (in Chapter 4), no significant relation was observed between time/throughputs of post-editing and sentence length, but it showed a possibility that shorter sentences could require less temporal efforts in comparison to longer sentences in both post-editing and translation. As the focus of our study was the post-editing effort, data for post-editing was provided in Figure 6.9. Similar to the result of the pilot study, it did not show a direct correlation of post-editing efforts and time. However, it seemed that the shorter sentences required less post-editing efforts than the longer sentences. The average of the first 10 sentence lengths (of 3 to 12) required 269.70 seconds

per word while the last 10 sentence lengths (of 37 to 59) required 1675.80 seconds, which was 5.2 times more than the former case.

In terms of a correlation of post-editing throughputs and sentence length, Figure 6.10 displayed that some level of post-editing efforts were required almost consistently in all sentence length. In addition, the number of words introduced by the evaluators per hour was not affected by the sentence length, which showed that efforts were irrelevant to the sentence length. It was surprising that the proportion of post-editing inputs could be higher in short sentences than in longer ones.

To sum up, from the two comparison studies, it was found that the temporal post-editing efforts could be affected by the sentence length, but the number of post-editing throughputs was irrelevant to it. The result could be interpreted that the MT outputs required a similar proportion of edits throughout the dataset regardless of the sentence length and that the edits in shorter sentences required less temporal efforts from the post-editors. However, a more thorough investigation would be necessary to specify the temporal efforts as the time measurement in our experiment could include the time spent on other activities such as looking up a dictionary.

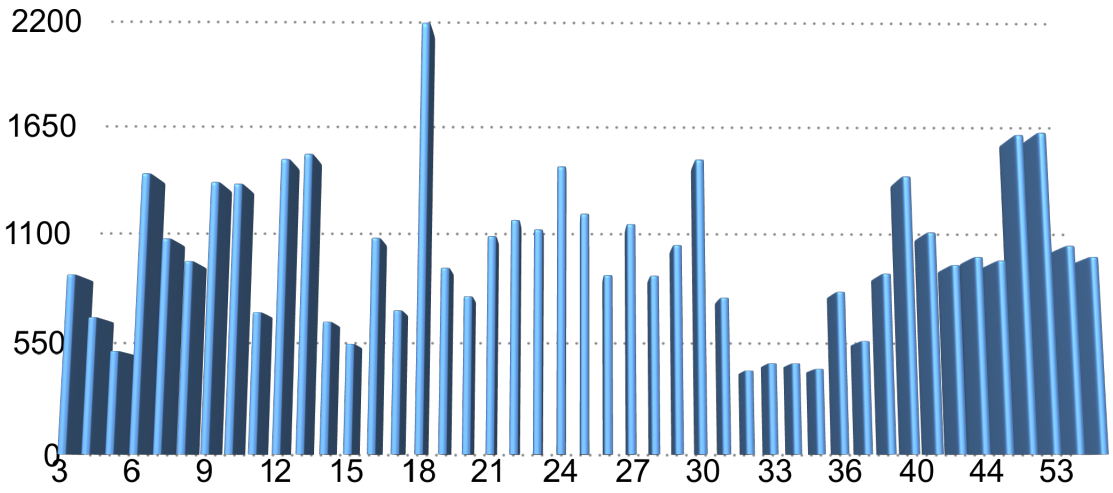


Figure 6.10: Correlation of post-editing throughputs (WPH) and sentence length.

6.3.2.2. Technical Efforts

The technical post-editing efforts in our experiment were measured by an edit distance. The edit distance was expressed on a scale of 0 to 10, where 0 implied that almost no edits were required in a sentence, while 10 meant that "there was no similarity between the post-edited output and a reference translation or the translation was made from scratch" (TAUS, 2010). The calculation was based on Levenshtein's algorithm and normalized by sentence length. Figure 6.11 showed the proportion of edit distance scores. It was found that lower edit distance scores took the largest part of the calculation with 26%. In other words, about 26% of the sentences hardly required any post-editing efforts (*distance* = 0) because they were errorless or the post-editors thought that the sentences were errorless. Moreover, no sentences were proportioned in distance 9 and 10, which could be interpreted, first and foremost, that none of the post-editors translated from scratch for the sentences that they were supposed to post-edit. Secondly, it could mean that there was no sentence that was entirely incorrect and, thus, required efforts as much as or more than translating from scratch.

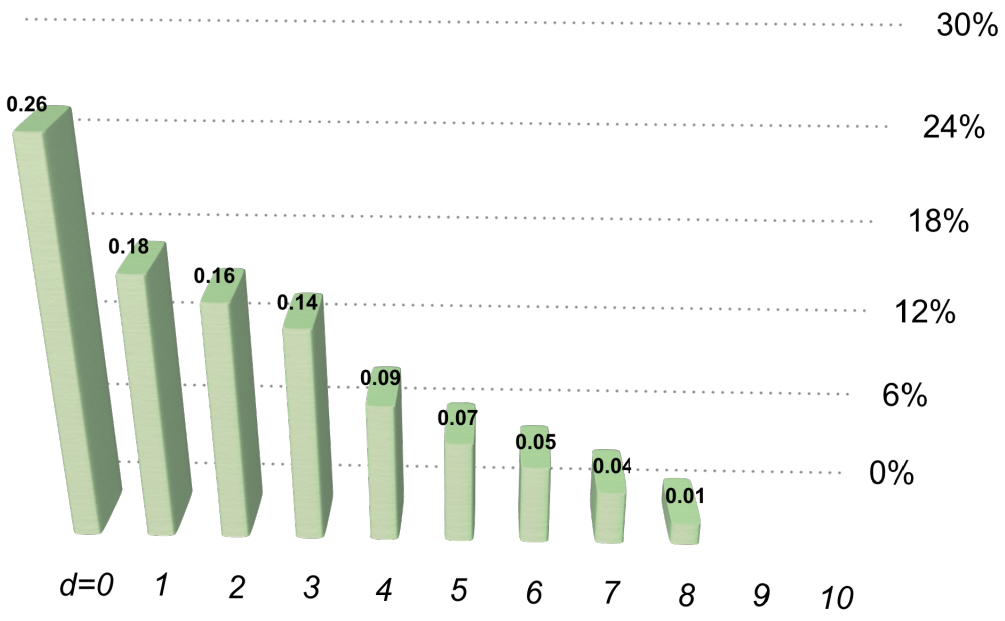


Figure 6.11: Technical efforts measure by edit distance.

6.3.3. HTER Score

HTER was a semi-automatic metric that calculated the number of edits in the post-edited translation normalized by the number of words in a reference translation. In a range of 0 to 1, HTER = 0 meant a perfect match that did not require any edits while HTER = 1 meant no similarity between the two sentences. The HTER score was calculated throughout the six target reference translations and given below in Table 6.12. The HTER score of the half of the dataset[57] in our experiment was 40.3 in a range of 35.7 to 42.9. Such a figure denoted that 40.3% of the dataset required edits. Such figure also showed a level of technical efforts.

	EV1	EV2	EV3	EV4	EV5	EV6	**Avg**
HTER	42.352	42.790	43.522	35.727	42.972	34.628	**40.332**

Table 6.12. HTER scores.

6.4. Chapter Summary

Chapter 6 analyzed the data obtained from the experiment. Firstly, in fluency and adequacy scoring, GNMT exhibited 78% of fluency and 77.75% of adequacy. It could be interpreted that the level of Korean was 78% reliable and the contents of the target text were 77.8% reliable. As such, on a hypothesis that the importance of fluency and adequacy was identical, the reliability of GNMT in the given setup was 77.9%. In terms of a correlation of the two methods to a sentence length, no clear-cut tendency was observed.

In segment ranking, the human reference translation was selected as the best translation with 2.67 (of 3) ranking score, which was equivalent to 89% of quality. GNMT held the lowest ranking of all, with 28.17%. According to a ranking score organized by machine type, 49.61% of GT was positioned as the 2nd. According to a ranking score by ranking choice, 68.78% of Rank 1 was taken by human translation while 53.88% of Rank 3 were taken by GNMT. One surprising finding was that MT was more favored by the evaluators than human translation. The qualitative analysis revealed that the reference translation could

[57] The score was based on half of the dataset because post-editing was performed on 127 sentences while translation from scratch was done on 126 sentences.

be erroneous or the evaluators could misjudge the dataset. More important speculation was the evaluators' preference of a literal translation over a liberal translation.

In post-editing, post-editing time and effort were measured. The post-editors spent 3.6 hours and 5.58 hours in total for post-editing and translation from scratch, respectively. The P ratio indicated that there was a 37% productivity gain of post-editing over translation from scratch. The data showed that post-editing time and throughputs (WPH) varied considerably throughout the post-editors. In terms of post-editing efforts, a temporal effort and technical effort were measured. There was no clear-cut correlations of post-editing time and throughputs to sentence length, but some evidence was found in terms of less temporal efforts in shorter sentences. In the technical efforts measured by edit distance, most of the segments required less technical efforts. In detail, according to the HTER score, it was found that 40% of the dataset required technical post-editing efforts.

7. Error Analysis

The evaluation analysis in Chapter 6 reported that in the given setup Google Translate achieved 78% of fluency and 77.75% of adequacy in fluency and adequacy scoring evaluation, 41.8% of human parity in segment ranking evaluation and a 37% productivity gain in post-editing over translation from scratch in a post-editing evaluation. Additionally, the system obtained an HTER score of 0.403. Having investigated the performance of the NMT engine in the *es-ko* pair in newswire domain in a variety of ways, this chapter quests for consequences of such results by performing error analysis that will be a key part of this thesis. The error analysis is carried out with two sorts of the dataset: i) raw system translations of Google Translate of the whole dataset (253 sentences) and ii) their post-edited version on half of the dataset (129 sentences) by six post-editors. This two-step error analysis is expected to yield an effect of having seven annotators.

Starting from an overview of the error analysis in Chapter 7.1, Chapter 7.2 makes a detailed report on the error analysis on raw outputs of Google Translate, namely GNMT, in a quantitative and qualitative way. Chapter 7.3 describes the results of the error analysis on their post-edited version, underpinning the findings in Chapter 7.2. Chapter 7.4 closes the chapter by a brief summary.

7.1. Overview

This section delineates the design of the binary error analysis. Chapter 7.1.1 clarifies the precise objectives of the error analysis. To this end, Chapter 7.1.2 defines each work step by step and illustrates the organization and contribution of the work. Chapter 7.1.3 establishes a scope of the work in Chapter 7 by pointing out limitations in the two-fold analysis. Having confirmed the framework of the error analysis, Chapter 7.1.4 specifies a methodology of the error analysis in this thesis, covering from an adjusted form of the error taxonomy, to an error calculation method and an interface of the TAUS workbench.

7.1.1. Objective

The primary goal of the error analysis in this thesis is to investigate what type of errors the GNMT engine tends to produce in the *es-ko* pair. Although the reasons for the errors of NMT are known to be untraceable and non-revisable (Klubička et al., 2017), it is of utmost importance to know the strengths and weaknesses of NMT. It is, thus, also our aim to pattern those errors to see the forest for the trees. Secondly, we strive to investigate the relative importance of error types that affect human comprehension of *machine language* by carrying out error analysis on its post-edited version. The result of such analysis is expected not only to underpin the result of the previous error analysis (on the system translation) but also to identify what type of errors are considered to be more/less important to human understanding. Such a comparative study will be of help to confirm human understandability of machine language in the given language pair and domain.

7.1.2. Organization

	Analysis I	**Analysis II**
Dataset	System translation of GNMT	Post-edited system translation of GNMT
Size	253 sentences	129 sentences per editor
Form	Quantitative / Qualitative	Quantitative
Chapter	7.2	7.3

Table 7.1: Organization of the binary error analysis.

To accomplish the aforementioned goals, two types of analysis are proposed as briefed in Table 7.1. The two-fold analysis is based on the same corpus, but Analysis I utilizes raw system translations of GNMT while Analysis II uses post-edited system translations of GNMT. Moreover, the size of the corpus is half in Analysis II as the post-editing has been performed on half of the dataset. The total size of error analysis in Analysis II, however, numbers 834 sentences (129 sentences by six editors).

The error analysis is carried out by the author manually. Firstly, an error typology is studied to find the most suitable categorization for this study including the *es-ko* language pair. And then, the system translation of GNMT is analyzed based on the adopted error taxonomy. Each category of detected errors is qualitatively analyzed and patterned by their traits. In the second error analysis, the post-edited system translations are contrasted to the raw system translation to count the number of errors that the post-editors perceived and to backtrace the type of errors that the post-editors revised. The error taxonomy is identical in the two tasks so that findings in Analysis I and Analysis II can compensate each other, as well as are comparable. By extension, the comparative study of the two-fold analysis will give insights about the most/least significant errors from the perspective of novice post-editors when transforming a machine language into a human language. In this way, such binary error analysis contributes to investigating the most/least frequent error types of NMT in the *es-ko* pair and to monitor the behavior of the Google engine in the course of translation.

7.1.3. Limitation

This error analysis is limited to the following aspects:

- In this error analysis, we take a stance of a machine when approaching the dataset. Such a perspective is crucial to understand the result, considering the organic architecture of NMT. Therefore, linguistic discussions are left out as much as possible, and a phenomenological approach is more concerned.
- This error analysis is based on a manual detection, in such a way that some inconsistency in the result is inevitable, for instance, in the calculation of errors. We assume that a small part of miscalculations, however, cannot distort the overall tendency of the result. Therefore, more stresses are placed on the proportion than the individual number of errors.
- The error calculation follows the HTER system that takes all errors with equal importance.
- The analysis is limited to the distant language combination of Spanish and

Korean. More generalization that can influence other language combinations of either distant from or close to Korean remains to be one of the promising lines of research.

- The analysis is limited to the newswire domain. While acknowledging the possibility of getting a dissimilar result from other domains, we also do not completely rule out a linguistic influence on some type of errors stemming from either Spanish or Korean, irrespective of domain type.
- This error analysis does not attempt to clear up the causes of the errors produced by the GNMT system. We are well aware of the impossibility of tracking down the consequences of the errors in NMT. This study only serves as an observer that finds a general tendency of the errors with some possible hypothesis that can explain the phenomenon. However, the suggested explanations are nothing but speculation.

7.1.4. Methodology

Similar to the MT evaluation, the error classification can be performed by two entities: human or machine. They both count the number of items according to a set of error classes that have been previously established. The key advantage of the manual way is flexibility in terms of error taxonomy. That is, the detection is flexibly adjusted to how the error classes are organized. The rigidness of applicability, on the other hand, is the great challenge of the automatic way (Popović, 2018b). The scope of error detection of the automatic methods such as WER, PER or Hjerson (Popović, 2011) is severely limited to grammatical errors such as inflections or reordering and some of the translational edits such as omissions and additions. In addition to such a challenge, the automatic error classification methods are solely dependent on the reference translation as the automatic evaluation methods do. The consequences of the indirect manner, thus, are identical.

In the meantime, the human way also faces challenges, such as inconsistent and subjective judgments and inefficiency and costliness of the job (Popović, 2018b). The task is especially labor-intensive because the error taxonomy should take into consideration the linguistic features of the language

pairs. Despite such challenges, this thesis adopts the human way, as the scope of the error classification is expected to be broader than what the automatic methods can deal with. To this end, the inherent disadvantages of manual detection are alleviated as much as possible by:

- Systematically digitizing the number of errors on the TAUS workbench
- Benchmarking the counting system of HTER
- Excluding subjective discussions over the errors

The inevitable consequences of employing the human method are not completely out of the picture. We acknowledge a few inconsistent judgments, as discussed in advance in Chapter 7.1.3. The goal of the current thesis, however, is not to obtain a numeric data of error distribution, but to observe the tendency and utilize the data to come up with useful insights. In that sense, while employing the manual method, this thesis strives to enhance the consistency of the result.

7.1.4.1. Error Classification

The error classification and classes are diverse per research setup.[58] The crucial component of the error taxonomy is to select one that best suits the purpose of the study (Popović, 2018b). To that end, we adopt the error classification that TAUS provides in their workbench that is designed based on the Multidimensional Quality Metric (MQM)[59]. The benefits of employing such error classification are that i) it is based on the mainstream error classification of MQM in MT, ii) it is easily adjusted to any language combinations and iii) it facilitates a convenient and, at the same time, systemized error calculation. The original form of error taxonomy of TAUS DQF, which is provided in their website[60], is adapted to the *es-ko* pair and to the purpose of this thesis. As in Table 7.2, there are four high-level error classes, each of which is sub-divided into a series of low-level classes.

---

[58] The overview of the error classification in MT is minutely researched in Popović (2018b), which those who are interested in this regard are referred to.

[59] http://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html.

[60] https://www.taus.net/evaluate/qt21-project#harmonized-error-typology

Accuracy	Fluency	Word Order	Style
Addition	Punctuation		Collocations
Omission	Spelling		Unidiomatic
Mistranslation	Spacing		Culture-specific
Untranslated	Grammar		Other

Table 7.2: Error classification adapted to the *es-ko* pair.

Accuracy It refers to a series of errors caused by not reflecting the source text in a proper manner (TAUS, 2010). The errors on this level are perceived from a translational aspect. There are four sub-categories: addition, omission, mistranslation and untranslated. The following definition is extracted from TAUS (2010) DQF:

- **Addition**: The target text includes text not present in the source.
- **Omission**: Content is missing from the translation that is present in the source.
- **Mistranslation**: The target content does not accurately represent the source content in a way that such an item does not interfere with other categories such as grammar or style.
- **Untranslated**: Content that should have been translated has been left untranslated.

Fluency It refers to an issue related solely to the target text in a linguistic and formal perspective (TAUS, 2010). It includes four sub-level of punctuation, spelling, spacing, and grammar.

- **Punctuation**: incorrect punctuation for the target language
- **Spelling**: incorrect spelling for the target language
- **Spacing**: incorrect spacing for the target language
- **Grammar**: Issues related to the grammar (but not syntax) of the text, other than spelling and orthography.

Word Order It refers to a syntactic error. Considering the linguistic dissimilarity of the *es-ko* pair, the word order is calculated independently from

the *grammar* category. It includes errors on both lexical and phrasal levels.

Style It refers to stylistic errors that do not correspond to the previous three error classes. Some of the sub-categories are, but not limited to, collocations, unidiomatic expressions, and cultural-specific references. The definition of each sub-categories is provided by TAUS (2010) as:

- **Collocations**: The content is a fixed expression, but translated in a literal way.
- **Unidiomatic expressions**: The content is grammatical, but not idiomatic.
- **Culture-specific reference**: Content inappropriately uses a culture-specific reference that will not be understandable to the intended audience.

7.1.4.2. Error Detection

The error calculation in this thesis adopts the HTER system that counts four types of edits: insertion, deletion, substitution, and shift. Each category is equally regarded as one edit. The special case is the shift that captures reordering errors based on a phrasal movement. That is, one edit is equal to one movement of a word or phrase to a proper position. Such a system is applied to the current error analysis so that all categories have equal weight. An instance is given below in Sentence 7-1 to show how the shift is considered. The source text (ST: Spanish) and system translation (GT: Korean) are shown along with their back translations (BT) in English for readers' information. Then, desirable word order is suggested. To correct the word order of GT to the appropriate form, the second shaded phrase (유럽 단일 통화 인 유럽이) should be moved to the front of the first shaded phrase (현재 연합에 의해 설계된). In this case, no matter how many words it skips, the edit count of shift for Sentence 7-1 is calculated as one edit.

Sentence 7-1

ST 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:

BT of ST	*However, when the 25th anniversary of the entry into force of the Maastricht Treaty designed by the current Union and its single currency, the euro, has been completed, the EU faces one of the greatest challenges of its history in the upcoming European elections. :*
GT	그러나 현재 연합에 의해 설계된 마스 트리 히트 조약 발효 25 주년과 유럽 단일 통화 인 유럽이 방금 만났을 때 EU는 다가오는 유럽 선거에서 역사상 가장 큰 도전 중 하나에 직면 해있다. :
BT of GT	*However, the EU is facing one of the biggest challenges in history in the upcoming European elections, when the 25th anniversary of the Maastricht Treaty, now designed by the Union, and the European single currency, Europe, have just met. :*
Model word order	그러나 유럽 단일 통화 인 유럽이 현재 연합에 의해 설계된 마스 트리 히트 조약 발효 25 주년과 방금 만났을 때 EU는 다가오는 유럽 선거에서 역사상 가장 큰 도전 중 하나에 직면 해있다. :

This example also shows that the judgment of error analysis is not based on the comparison of the system translation and reference translation, but on the annotator's intuitive judgment that aims at the ideal scenario that GT can get. Unlike HTER, the role of the reference translation is delimited to an auxiliary material.

Following such an error detection system, the number of edits on each category are calculated as in Sentence 7-2. On the left side of GT, the number of errors is calculated. The next line of PT (post-edited system translation) shows the number of errors according to the post-editors' judgment. The types and number are dissimilar in this case. It demonstrates that there are various ways to edit the sentences and that not all errors should be dealt with to revise the system translation to a proper form.

Sentence 7-2

ST	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:								
GT	그러나 현재 연합에 의해 설계된 마스 트리 히트 조약 발효 25 주년과 유럽 단일 통화 인 유럽이 방금 만났을 때 EU는 다가오는 유럽 선거에서 역사상 가장 큰 도전 중 하나에 직면 해있다. :	add	omi	mis	unt	pun	spc	w.o.	grm	sty
				6		1	6	2	1	2

PT	그러나 현재 연합에 의해 설계된 마스 트리 히트 조약 발효 25 주년과 유럽 단일 통화인 유로화와 EU는 다가오는 유럽 선거에서 역사상 가장 큰 도전 중 하나에 직면 해있습니다.	3		1		1			

7.1.4.3. Interface of TAUS DQF

The calculated errors are stored in TAUS DQF. As in Figure 7.1, the high-level error classes adhere to the default set that is composed of verity, design, locale convention, style, terminology, fluency errors, and accuracy. While the categories of fluency, accuracy, and style are introduced in the corresponding section, word order is calculated under the local convention section, and the rest classes are left unused. The sub-categories of each error are detailed in the comment section.

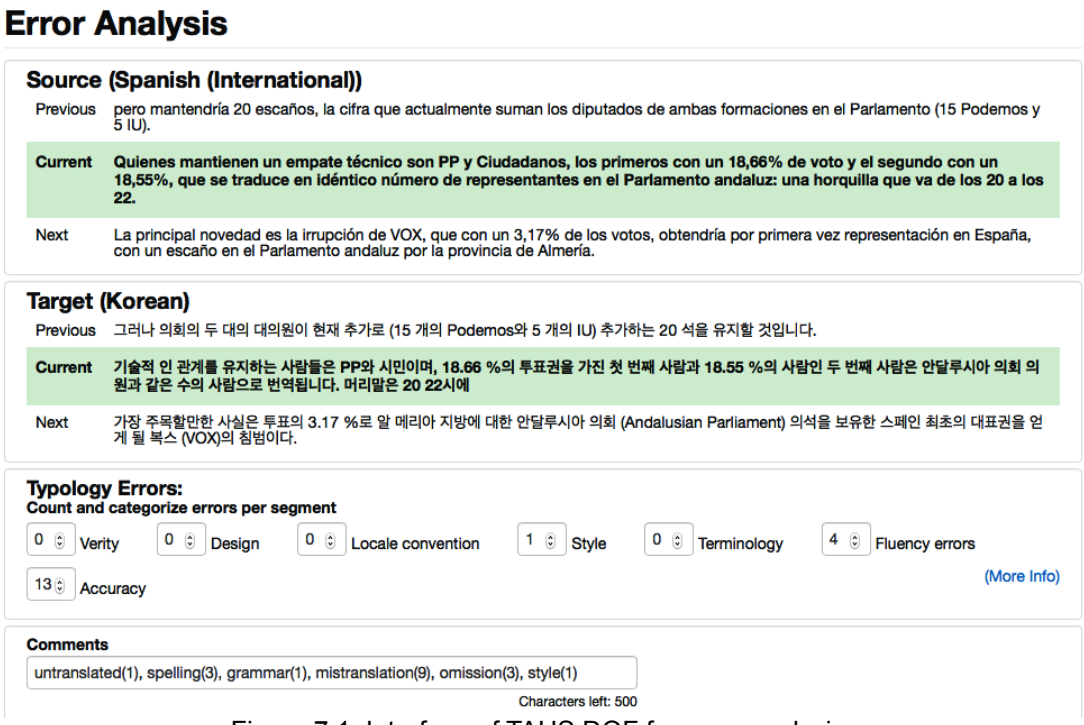


Figure 7.1: Interface of TAUS DQF for error analysis.

7.2. Error Analysis I: On Raw System Translations

The system translation produced by Google Translate is manually analyzed according to the adjusted error classification in a quantitative and qualitative way. The result of the error analysis is organized as a quantitative approach in

Chapter 7.2.1 and a qualitative approach in Chapter 7.2.2.

7.2.1. Quantitative Analysis

	w.o	mis	spc	omi	unt	grm	add	pun	sty	spl	Total
n	115	**783**	465	314	116	73	48	34	48	1	**1997**
n'	3.54	**33.52**	17.85	10.31	4.43	3.28	2.19	1.61	2.06	0.05	**78.84**

Table 7.3: Statistics of errors by the error classification. (w.o=word order; mis=mistranslation; spc=spacing; omi=omission; unt=untranslated; add=addition; pun=punctuation; sty=style; grm=grammar; spl;spelling) The number of errors is provided in *n*, followed by their normalized version in *n'*.

Following the granularity of the error classification described in Chapter 7.1.4.1, the dataset of 6,426 words (253 sentences) was analyzed manually. Considering the punctuation, spacing and spelling errors, the size of the corpus and the number of errors are incomparable. Table 7.3 shows that 1,875 errors in total are detected from the dataset. The largest number of errors is resulted from mistranslations (39.21%), followed by spacing errors (23.28%) and

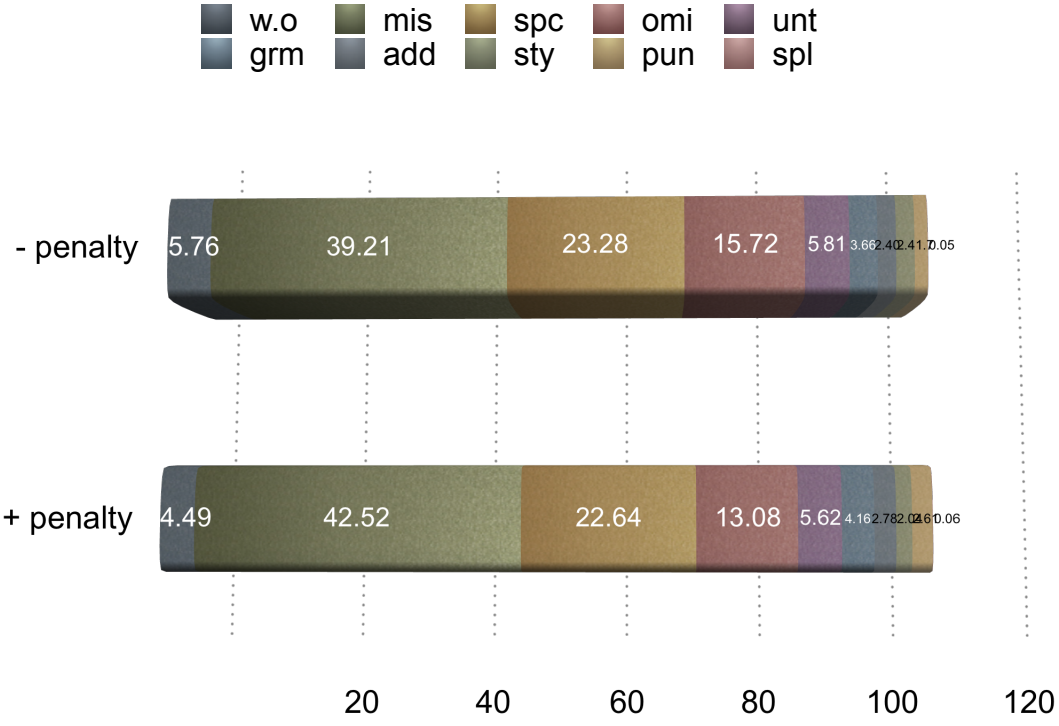


Figure 7.2: Distribution of error classification of with or without penalty of sentence length.

omissions (15.72%). Subsequently, each type of error is normalized by the number of the source sentence. Such normalization serves as a weight to the result on a hypothesis that errors in shorter sentences have a higher penalty than those in longer sentences. The distribution of the errors is compared in Figure 7.2. The overall distribution is similar. The most noticeable change is that the proportion of mistranslation is increased from 39.21% to 42.52% while the proportion of word order is decreased from 5.76% to 4.49%.

Addition There were 48 errors of addition in 29 sentences. The number of additions per sentence ranged from 1 to 4. About 34% of those sentences had one error of addition, constituting the most frequent number in a sentence.

Omission There were a total of 314 errors of omission found in 99 sentences. Such errors were tagged by their part of speeches (POS)[61]. There were 10 types of POS tags that were statistically important: noun (nn), preposition (prep), verb (vb), relative (rel), adverb (adv), adjective (adj), determiner (det), article (art), numeral (nm) and quantifier (ql). The distribution of those ten POS tags was shown in Figure 7.3. The most frequent type of omissions was nouns with 23.9%, followed by prepositions (18.6%).

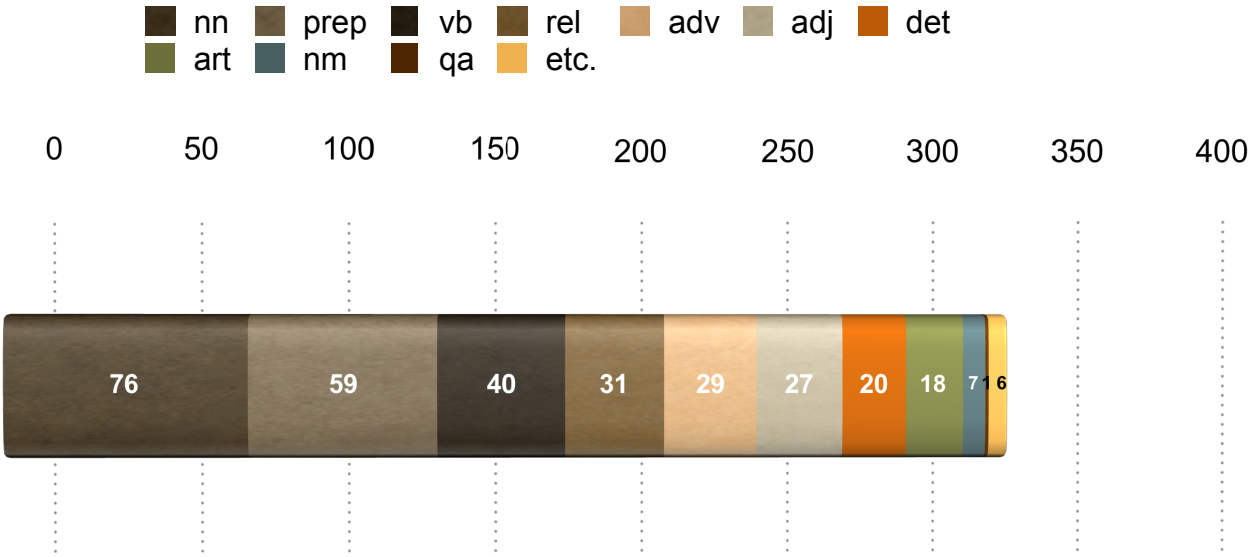


Figure 7.3: Distribution of POS tag sets in omission.

[61] The Korean POS tag sets are referred to as: http://incredible.ai/nlp/2016/12/28/NLP/.

Mistranslation Mistranslation was meaningful to the current error analysis in that it occupied about 39% to 42.5% of the error types. As the definition of mistranslation in this thesis was all inaccurate representation of the source text that did not belong to grammar nor style, the errors in this category represented incorrect word choice.

Untranslated There were 116 untranslated words found in 89 sentences, occupying 5.81% of the total errors. The distribution of the number of errors per sentence is quite even throughout the dataset between 1 and 2 edits with 17.5% and 16.4% respectively. Interestingly, it was found out that some of the untranslated words had actually been translated occasionally in the dataset. Some of the most relevant vocabularies were given in Table 7.4. The data showed the inconsistent performance of the translation on the identical proper nouns. For instance, the name *Trump* was 75% translated, but 15% untranslated. Some of the items such as *PP* and *PSOE* had more untranslated cases than translated cases while items such as Trump and Donetsk had more translated cases than untranslated ones. An in-depth investigation as to what kind of strategies the system applied to each case was further detailed in a qualitative study in Chapter 7.2.2.3.

	UNT	T	%		UNT	T	%
Trump	3	17	15%	*Donald*	3	1	75%
Ciudadanos	2	8	20%	*Adelante*	3	1	75%
Andalucía	3	6	33.3%	*VOX/Vox*	3	1	75%
PP	7	1	87.5%	*Instituto*	1	2	33.3%
Brexit	5	2	71.43%	*Real*	1	1	50%
PSOE	6	1	85.7%	*Abrams*	1	1	50%
Díaz	4	2	66.7%	*ABC*	1	1	50%
Podemos	5	1	83.3%	*Gore*	1	1	50%
Donetsk	1	4	20%	*Broward*	1	1	50%

Table 7.4: List of untranslated (UNT) and translated (T) vocabularies. The percentage of untranslated cases are displayed in gray columns.

Punctuation It constituted the smallest part of the total errors, excluding one spelling error. 34 errors were detected in 24 sentences. Such errors did not occur more than twice in one sentence. The main reason for incorrect punctuation was an omission, with 28.1%.

Spacing The second most frequent error class turned out to be spacing, with 23.28%. The spacing errors could be understood as a consequence of a different morphological segmentation between the two languages. The key point of the spacing error was their level of influence on human understanding, which would be studied in a qualitative part in Chapter 7.2.2.6.

Grammar The grammatical error took 3.66% of the total error classes, detected 73 times in 60 sentences. Most of the grammatical errors were observed from incorrect use of part of speeches and prefixes.

Style The stylistic errors held 2.4% of the error classes with 48 items.

7.2.2. Qualitative Analysis

Addition	Omission	Untranslated	Punctuation	Grammar	Style
Duplicated addition	Complete elimination	Raw source word	Should have been omitted	POS error	Awkward style
Irrelevant addition	Partial elimination	Phonetic transcription	Adaptation	Verb tense error	Collocation error
Overlapped addition	Substitution		Omission	Prefix error	Technical error
Indication of subject					
Addition of Chinese characters					
Supplementation of ellipsis					

Table 7.5: Patterns of errors of each error class.

In this section, the error analysis was qualitatively approached. For each error class, some common patterns were observed and organized with some of the most relevant instances. Some categories were not further studied: i) mistranslation, in order to avoid subjective discussions over word choice, ii) spelling, due to its statistical insignificance. Table 7.5 showed noticeable patterns of each error taxonomy that would be studied subsequently.

7.2.2.1. Addition

The addition referred to an item that appeared in the target text but did not exist in the source text. The computation started from looking into the target sentence and capturing some parts that did not have a correspondence in the source sentence. However, all added words did not represent an error. In that sense, the automatic detection of such error class could be challenging. Having in mind that all the 48 items were not an error, a qualitative description was given in this section by the six most noticeable patterns of the errors.

Duplicated Words A large part of added words in this dataset fell into this type of pattern where identical words were reproduced for some unknown reason. The chances were that the fault was a consequence of a confusion of the context due to an interruption by hyphenated fragments.

Sentence 7-3

ST Una vez en la Casa Blanca, quienes confiaban -sobre todo, sus aliados republicanos- en que adoptaría un tono «presidencialista» se han topado con que Trump sigue en campaña de sí mismo.

GT 백악관에서 한 번 공화당 동맹국 인 공화당 동맹국들에 대해 "대통령 주의적 입장"을 채택 할 것을 신뢰 한 사람들은 트럼프가 여전히 스스로 선거 운동을하고 있음을 발견했다.

BT *Once in the White House, those who trusted in adopting a "presidential" position for Republican allies, Republican allies, found that Trump was still campaigning for himself.*

In Sentence 7-3, *Republican allies* (공화당 동맹국) were repeated twice, but no evidence was found as to which part it originated from. The only clue for the incorrect translation of that phrase was due to the insertion of the phrase "…-

sobre todo, sos aliados republicanos-…", and thus, the omission of sobre todo (especially).

Sentence 7-4

ST El portavoz del Parlamento Europeo, Jaume Duch, también trata de lidiar con el auge de los partidos populistas y euroescépticos:

GT 유럽 의회 (Jaume Duch)의 유럽 대륙 의회의 대변인은 또한 대중 주의자와 유로 지성인들의 부상에 대처하기 위해 노력한다.

BT *The spokesperson of the European Parliament of the European Parliament (Jaume Duch) also strives to cope with the rise of popularists and Euro-intellectuals.*

Similarly, Sentence 7-4 showed that 유럽 의회 (the European Parliament) was duplicated while the proper noun Jaume Duch was omitted. In addition, the second item had 대륙 (continent) in between, which was not shown in the English back translation of the target text (BT).

Sentence 7-5

ST La candidata socialista obtendría el 37,41% de los votos y entre 45 y 47 escaños.

GT 사회당 후보는 37.41 %의 표와 45 ~ 47 표의 표를 얻는다.

BT *Socialist candidates get 37.41% votes and 45-47 votes.*

Some of the examples exhibited some shreds of evidence that could be understood. In Sentence 7-5, 표 (vote*)* was used three times: first, as a translation of votos; second, as a translation of escaños; and third, presumably as a part of the translation of the verb obtendría (to acquire). The first case belonged to the correct translation while the second case was an error of mistranslation. What mattered in this section was the third case. It was speculated that the system translated the verb into 'get a vote' instead of just 'to get'. The repeated part was invisible in BT.

Sentence 7-6

ST En el Senado, el vuelco demócrata es mucho más difícil, porque buena parte de los escaños en juego están en territorio favorable para los republicanos.

GT 상원 의석에서 민주당 전복은 훨씬 어려워진다. 공화당 의원들 대부분이 공화당의 유리한 영역에 있기 때문이다.

BT *In the Senate seat, the Democratic uprising becomes much more difficult. Most Republicans are in the favorable area of the Republican Party.*

The addition of Sentence 7-6 occurred with the word republicanos (Republicans). The translation clearly showed that the vocabulary was repeated, but when compared to ST, it suggested that los escaños en juego could not be handled properly in the system for some reason and was substituted to *Republicans*. Such a pattern was also witnessed in the following sentence.

Sentence 7-7

ST El Parlamento Europeo está discutiendo el nuevo marco financiero plurianual para los años 2021-2027, donde está sobre la mesa la opción de otorgar más capacidad económica a la Autoridad para controlar el uso de los Fondos de la UE.

GT 유럽 의회는 2021 년에서 2027 년까지의 새로운 다년간 금융 체제에 대해 논의 중이며 EU 당국이 EU 자금 사용을 통제 할 수있는 경제적 능력을 더 많이 부여 할 수있는 방안이 논의되고있다.

BT *The European Parliament is discussing a new multi-year financial system from 2021 to 2027 and discusses ways in which the EU authorities can give more economic power to control EU funds.*

In Sentence 7-7, the proper noun la Autoridad (the authority) was over-translated with an existing vocabulary in the sentence, *EU authorities*. The reasons were unknown but the pattern seemed alike.

Sentence 7-8

ST El partido de Albert Rivera obtendría representación en todas las provincias, con un avance muy significativo en Cádiz, donde pasaría de 1 a 3 o 4.

GT 앨버트 리베라 (Albert Rivera)의 당은 카디스에서 매우 중요한 진보와 함께 모든 지역에서 대표권을 얻었으며 카디스는 1에서 3 또는 4로 넘어갔습니다.

BT *The party of Albert Rivera, with significant progress in Cadiz, received representation in all regions and Kadis went from one to three or four.*

In some cases like Sentence 7-8, the system repeated the same word because it was regarded as the subject for the subordinate clause. The complex sentence with a subordinate clause of donde (where) was well translated into two clauses in GT. The error occurred when the subject of the subordinate clause was not the preceding word Cádiz, but representación (representation) in Cádiz.

Sentence 7-9

ST Los apoyos de Trump -con una opinión muy favorable de sus bases- han puesto y quitado candidatos republicanos, que no tienen más remedio que proclamar su adhesión al jefe.

GT 트럼프의 지원은 공화당 후보들을 두었다. 공화당 후보들은 입후보자의지지를 선언하는 것 외에는 선택의 여지가 없다.

BT *Trump's support left Republican candidates. Republican candidates have no choice but to declare their candidacy.*

The similar case was Sentence 7-9 where candidatos republicanos (Republican candidates) was repeated as the subject of the subordinate clause que. In this time, the translation conveyed the correct meaning of ST because fortunately the subject of the subordinate clause was the *Republican candidates* despite the role of que was not correctly understood. The problem rather occurred in the main clause where the inserted phrase deterred the structure of the sentence.

Sentence 7-10

ST pero mantendría 20 escaños, la cifra que actualmente suman los diputados de ambas formaciones en el Parlamento (15 Podemos y 5 IU).

GT 그러나 의회의 두 대의 대의원이 현재 추가로 (15 개의 Podemos와 5 개의 IU) 추가하는 20 석을 유지할 것입니다.

BT *But two delegates to the Congress will now hold 20 seats to add (15 Podemos and 5 IUs) in addition.*

While the previous examples were mostly nouns, the error in Sentence 7-10 was a consequence of the verb sumar (to add). While GT translated it to 추가하는 (to add), the similar word was additionally inserted as an adverb 추가로 (in addition) so that the sentence formed "[…] to add in addition".

Sentence 7-11

ST Según la Comisión Nacional Electoral, 78 mil contadores de votos, 12 mil observadores, 105 mil administradores y 2.500 máquinas para el recuento automatizado de papeletas han sido movilizados para el escrutinio a nivel nacional.

GT 전국 선거위원회 (National 선거위원회)에 따르면 78000 명의 투표자, 12000 명의 옵서버, 105000 명의 관리자 및 2,500 대의 자동 투표기가 전국 조사를 위해 동원되었습니다.

BT According to the National Election Commission, 78,000 voters, 12,000 observers, 105,000 managers, and 2,500 automatic voting machines were mobilized for national surveys.

Some addition was inside the brackets as in Sentence 7-11. While the intention to give an additional explanation of a proper noun was positive, the phrase had to be regarded as an error because it was translated even inside the brackets in Korean, as well as partly in English.

Irrelevant Words This type of pattern referred to a series of added words that had nothing to do with either the source or the target sentence. The reason for such errors was not explicable, but some interesting tendency was that many of the added words were contextually deducible in the target sentence, in such a way that in many cases the translation itself sounded quite natural.

Sentence 7-12

ST También se encaminan a recuentos las elecciones a senador por Arizona y a gobernador de Georgia.

GT 애리조나 상원 의원과 조지아 주지사 선거 역시 중요하게 고려된다.

BT *Arizona Senator and Governor of Georgia elections are also considered important.*

In Sentence 7-12, the verb encaminar (to direct) was mistranslated to *considered important*. The first part (considered) was a mistranslation and the second part (important) was an addition. While the two words came from nowhere GT itself sounded natural.

Sentence 7-13

ST Este es el resultado que arrojó la encuesta realizada por la agencia de estudios de opinión pública Realmeter a 1.005 surcoreanos mayores de edad entre el 13 y el 14 de agosto.

GT 이는 여론 조사 기관인 Realmeter가 8 월 13 일에서 14 일 사이에 거주하는 1,005 명의 한국인을 대상으로 실시한 설문 조사의 결과입니다.

BT *This is the result of a poll conducted by a survey of 1,005 Korean residents, Realmeter, between August 13th and 14th.*

Similarly, in Sentence 7-13, GT read as "…Koreans who stayed during August 13 to 14…" without translating mayores de edad (adults). There was no clue as to how such a decision was made, but the expression was highly contextual and natural in the target sentence. BT translated the segment to *Korean residents*.

Sentence 7-14

ST La presidenta de la Junta de Andalucía, Susana Díaz, ganaría las próximas elecciones autonómicas con más de 10 puntos de ventaja, pero lejos de la mayoría absoluta (55 escaños de un total de 109), según la encuesta hecha pública este miércoles por el Centro de Investigaciones Sociológicas (CIS).

GT 유다 데 안달루시아 (Susana Díaz) 의장은 다음 주 지방 선거에서 10 점 이상의 우위를 점할 것으로 예상했지만, 이번 수요일에 센터에 의해 공개 된 설문 조사에 따르면, 절대 다수 (109 개 가운데 55 석) 사회학 연구 (CIS).

BT *Susana Díaz, chairman of the Supreme Council for Democracy, predicted that she would have an advantage of more than ten in the next provincial election, but according to a survey released by the Center this Wednesday, the absolute majority (55 out of 109)).*

In Sentence 7-14, the error occurred when translating las próximas elecciones (the next election). GT translated it to *next week's election* while BT, interestingly enough, omitted the problematic part 'week'. Although the meaning expressed in the source text was entirely changed, the expression itself (next week) was too natural to be even detected as an error. Additionally, it was surprising that the verb ganaría (would win) that was underlined in Sentence 7-14 was translated to *expected to win* instead of *would win,* which sounded more natural than its literal equivalence.

Sentence 7-15

ST Nadie le puede negar que entonces funcionó.

GT 아무도 그 경로가 작동한다고 부인할 수 없습니다.

BT *No one can deny that the path works.*

In Sentence 7-15, the verb 'work' was translated into 그 경로가 작동한다고 (the path works). That is, the subject of the verb was generated in an incorrect way, but it sounded natural.

Sentence 7-16

ST Hermano del también dimisionario Boris y político de gran prestigio y popularidad, Jo Johnson, que votó contra el Brexit, era el encargado de pactar el estratégico acuerdo sobre Transportes, un sector clave para las futuras relaciones del Reino Unido con la UE.

GT 부시 대통령과 블레치 총리에게 투표 한 조 존슨 (Jo Johnson) 대통령은 보리스 (Boris)를 사임하고 영국과 EU의 미래 관계를위한 핵심 분야 인 교통에 관한 전략적 협상을 담당했다.

BT *President Jo Johnson, who voted for President Bush and Prime Minister Blech, resigned from Boris and was in charge of strategic negotiations on transportation, a key area for future UK and EU relations.*

In the case of Sentence 7-16, Brexit was mistranslated to 블레치 총리에게 (Prime Minister Blech), which was followed by addition of 부시 대통령과 (with President Bush) that, without knowing the source text, sounded just natural.

Overlapped Words The overlapped addition referred to a double translation of a source word that caused a redundancy. It was not a mainstream tendency, and a few errors belonged to such a pattern, but this phenomenon of causing a redundant translation seemed intriguing.

Sentence 7-17

ST Ante la incapacidad de Londres de gestionar este proceso, el intento de organizar otro referéndum es solo la vía de eludir la elección entre lo malo y lo peor.

GT 런던이이 과정을 관리 할 능력이 없다는 점을 감안할 때 다른 국민 투표를 조직하려는 시도는 악과 최악의 선택을 피하는 유일한 방법 일뿐입니다.

BT *Given that London is not capable of managing this process, attempts to organize a different referendum are only the way to avoid evil and the worst choices.*

In Sentence 7-17, the word solo (just) was translated twice to a different POS in Korean: firstly, into an adjective 유일한 (only) and secondly, into a dependent noun -뿐_(just). While the lexical selection was correct, GT sounded redundant but no ungrammatical as it could be understood as an emphasis.

Sentence 7-18

ST En la vecina también autoproclamada República Popular de Lugansk (LNR) se celebran igualmente elecciones y el que parte con ventaja es el que desempeña interinamente el cargo de máximo dirigente, Leonid Pasechnik.

GT 이웃 국가들에서도 스스로 자칭 된 LUMR (Lugansk Republic of Lugansk) 선거도 개최되며 유리한 방향으로 이끌어가는 당은 일시적으로 최고 지도자 인 Leonid Pasechnik의 입장을 취하고있다.

BT *The neighboring countries also hosted the LUMR (Lugansk Republic of Lugansk) election, which is self-proclaimed, and the party that leads in favorable direction is temporarily elected by Leonid Pasec*

In Sentence 7-18, the prefix ‘auto-’ is translated to ‘스스로 (themselves)’ and an adjective ‘self-claimed’. In both of the cases, the translation is not fluent but

understandable. One interesting thing is that the back translation into English does not have the same problem.

Indication of Subject This type of addition referred to a special indication of a subject which was usually hidden and should only be inferred from the inflection of a verb. Therefore, errors were usually caused when the generated subject was incorrect. It also meant that if a subject was generated correctly, it should not be considered as an error. In fact, such a case could be also understood as a mistranslation considering the Spanish grammar, but it was regarded as an addition in this thesis.

Sentence 7-19

ST Pero en Moscú piensan de otra manera.

GT 그러나 모스크바에서 그들은 다르게 생각합니다.

BT *But in Moscow they think differently.*

In Sentence 7-19, the subject of the verb piensan (they think) represented people in general, but GT designated it as *they,* and so, the sentence seemed to be intended for unspecified individuals.

Sentence 7-20

ST sigue siendo una democracia».

GT 그것은 여전히 민주주의입니다. "

BT *It is still democracy. "*

In Sentence 7-20, the translation of the subject inferred from the verb sigue (*keep* in 3rd person singular) matched the translation (그것은 - *it*) correctly. The sentence would sound more natural without the word, but it could not be considered as an error. Such a case should be excluded in automatic detection.

Sentence 7-21

ST Ha destapado fraudes y ahora ha habido un caso en el Tribunal en el que se ha pedido devolver el dinero», dijo Laurin Berresheim, quien prepara el trabajo de Arndt Khon (eurodiputado de los socialdemócratas alemanes) en el Comité de Control Presupuestario.

GT 그것은 사기를 밝혀 냈고 지금은 재판소에 돈을 돌려 줄 것을 요청한 사례가있다 "라고 예산 통제위원회의 Arndt Khon (독일 사회 민주주의 자들의 MEP)의 작업을 준비하는 Laurin Berresheim은 말했다.

BT *It has revealed fraud, and now we have asked to return the money to the court, "said Laurin Berresheim, who is preparing for the work of Arndt Khon (MEP of the German Social Democrats) in the Budget Control Committee.*

In Sentence 7-21, the subject of the verb ha (*have* in 3rd person singular) could be understood from the previous sentence, which was *the European Parliament*. GT generated the subject as 그것은 (it), which was not appropriate in this sentence.

Chinese Characters This type of addition referred to an added Chinese character inside the brackets. It was a common writing style in Korean, especially in a newspaper, when the word was a proper noun that needed to be clarified or a word that had multiple meanings and could confuse readers, otherwise. The addition of such kind was not erroneous.

Sentence 7-22

ST No obstante, se analiza que el interés de la población en la política, que empezó a aumentar a partir de las protestas contra el gobierno de la ex presidenta Park Geun Hye, su destitución, y las elecciones presidenciales prematuras organizadas en mayo de 2017;

GT 그러나 2017 년 5 월 조직 된 박근혜 (朴 槿 惠) 전 (前) 대통령 정부의 해임과 해임과 대통령 선거에 대한 항의 시위에서부터 시작된 정치인들의 관심이 증가한 것으로 분석된다.

BT *However, it is analyzed that the interest of politicians started from the protests against the dismissal and dismissal of former President Park Geun-hye, organized in May 2017, and the presidential election.*

In Sentence 7-22, the name of the Korean ex-president Park Geun Hye (박근혜) was translated and detailed with Chinese characters of her name to specify the person. Moreover, there was another word that was specified with a Chinese character: 전 (ex-). The Chinese character of the prefix was provided to clearly suggest the meaning because the word 전, for example, had 34 meanings[62].

[62] https://stdict.korean.go.kr/search/searchResult.do

Sentence 7-23

ST El estado de Vermont fue el primero en abrir las puertas de sus colegios electorales, a las 5 de la madrugada, hora local.

GT 버몬트 주 (州)는 현지 시간으로 오전 5시에 처음으로 투표소 문을 열었습니다.

BT *Vermont has opened its polling place for the first time at 5 am local time.*

Similarly, Sentence 7-23 also had a Chinese character of 주 (state) for the identical reason. It was assumed that the training dataset contained texts of such kind. In that sense, this pattern of translation allowed us to peek through what type of the training data the engine had been trained with.

Supplementation of Ellipsis This type of addition referred to an inevitable generation of ellipsis in a source text. Therefore, if the generated word was correct, it should not be considered as an error. The addition was inevitable because without it the segment could not convey the original meaning.

Sentence 7-24

ST Quienes mantienen un empate técnico son PP y Ciudadanos, los primeros con un 18,66% de voto y el segundo con un 18,55%, que se traduce en idéntico número de representantes en el Parlamento andaluz: una horquilla que va de los 20 a los 22.

GT 기술적 인 관계를 유지하는 사람들은 PP와 시민이며, 18.66 %의 투표권을 가진 첫 번째 사람과 18.55 %의 사람인 두 번째 사람은 안달루시아 의회 의원과 같은 수의 사람으로 번역됩니다. 머리말은 20 22시에

BT *The people who maintain the technical relationship are the PP and the citizen, the first person with 18.66% of the vote and the second person with 18.55% will be translated into the same number of people as the Andalusian parliamentarians. The preamble is at 20*

In Sentence 7-24, los primeros (the first) and el segundo (the second) could be understood as ‘the former’ and ‘the latter’. The complete form of the two words was *the first party, which was* PP, and *the second party, which was Ciudadanos*. The reproduction in GT, however, was *the first person* and *the second person*, which led to a mistranslation as well.

Sentence 7-25

ST pero mantendría 20 escaños, la cifra que actualmente suman los diputados de ambas formaciones en el Parlamento (15 Podemos y 5 IU).

GT 그러나 의회의 두 대의 대의원이 현재 추가로 (15 개의 Podemos와 5 개의 IU) 추가하는 20 석을 유지할 것입니다.

BT *But two delegates to the Congress will now hold 20 seats to add (15 Podemos and 5 IUs) in addition.*

In Sentence 7-25, the gray phrase was a reduced form of 15 escaños de Podemos y 5 escaños de IU (15 seats of Podemos and 5 seats of IU). While *seat* should be generated, GT selected a dependent noun 개 which meant a *number*, and the addition, in this case, was erroneous.

Sentence 7-26

ST "Es una desgracia", dijo Trump, quien en las presidenciales de 2016 atribuyó a irregularidades falsas su derrota, en el número total de votos, frente a la demócrata Hillary Clinton.

GT "그것은 불명예 스럽다." 2016 년 대통령 선거에서 민주당 힐러리 클린턴에 반대하여 총 투표 수에서 거짓 부정에 대한 자신의 패배를 돌린 트럼프는 말했다.

BT *"It is disgraceful," Trump said in his 2016 presidential election, against his Democratic challenger Hillary Clinton, who turned his defeat against false injustice in total votes.*

In Sentence 7-26, the segment las presidenciales (the presidential) was a reduced form of las elecciones presidenciales (the presidential election). Surprisingly, GT generated the right word for the segment as BT showed. Such type of addition should not be regarded as an error.

7.2.2.2. Omission

This type of error referred to a content that was present in the source text but was absent from the target text. There were 314 omitted words in the dataset. Such calculation was based on the number of source words. In order to pattern them, all items were manually tagged by their part-of-speech (POS) according to an original 87-tag Brown corpus tag set (Francis and Kucera, 1982). The POS tagging was especially challenging for a sequential omission such as Sentence 7-27 given below. Although the POS did not have much sense in this case, the phrase was tagged as an independent item. As such, the phrase had 7 omissions in total that were composed of 2 verbs, 3 conjunctions, 1 article, and 1 noun.

Sentence 7-27

ST Si bien la competencia es reñida, los expertos prevén que el Partido Demócrata renovará la mayoría en la cámara baja por primera vez en ocho años -suponiendo un gran obstáculo para Trump, que busca la reelección en 2020-, **mientras que estiman que el Senado mantendrá** el predominio republicano.

GT 경쟁이 치열한 상황에서 전문가들은 8 년 만에 처음으로 민주당이 하원에서 다수를 갱신 할 것으로 예상한다. 트럼프는 2020 년 재선을 모색하는데 큰 걸림돌이된다. 공화당 우세

The use of POS tagging facilitated a systematic categorization of the types of omission. Some of the most frequent POSs were article, determiner, noun, adjective, adverb, verb, preposition, relative, numeral and quantifier. The other statistically insignificant POSs were included in the etc. section. Table 7.6 was reproduced from Figure 7.3. The omission was mostly occurred with nouns, with 24.2%, followed by the preposition with 18.79% and verb with 12.74%. Among the aforementioned types of tags, noun, determiner, verb, and article were minutely analyzed.

	nn	prep	vb	rel	adv	adj	det	art	nm	qa	etc.
n	76	59	40	31	29	27	20	18	7	1	6
%	24.20	18.79	12.74	9.87	9.24	8.60	6.37	5.73	2.23	0.32	1.91

Table 7.6: Distribution of POS tag sets in omission.

Noun The largest proportion of the omission was dedicated from the noun (24.2%). Six types of noun tags were observed: namely, singular common noun (_NN), plural common noun (_NNS), singular proper noun (_NP)[63], possessive personal pronoun (_PP$), objective personal pronoun (_PPO) and nominative wh- pronoun (_WPS). The distribution of the tags was displayed in Table 7.7. Most of the omitted nouns belonged to a common noun (_NN and

[63] It is to note that the common noun in a capital form to designate a certain institution or organization is included in the proper noun.

_NNS) with 71.3%. It could be interpreted as the omission of such type was purely an error of lexical choice. As the discussion over the lexical choice could be subjective, the main focus of the discussion in this section was on the proper noun (_NP) and possessive personal pronoun (_PP$).

_NN	_NNS	_NP	etc.
62.3	9	12.9	15.8

Table 7.7: Distribution of noun tags. (unit: %)

The omitted pronoun noun held 12.9% of the total omitted nouns. There were six pronouns (_NP) that were omitted from the translation. Most of them were names of a certain group.

Sentence 7-28

ST En 2016, Morten Messerschmidt, eurodiputado danés del ahora desaparecido Movimiento por una Europa de Libertades y Democracia (MELD) que utilizó fondos europeos para apoyar campañas nacionales, se vio obligado a devolver este dinero.

GT 2016 년에, Morten Messerschmidt, 자유주의와 민주주의 (MELD)의 국가 운동을 지원하기 위하여 유럽 기금을 사용한 지금 무효 한 운동의 덴마크 MEP는,이 돈을 돌려 보내야했다.

BT *In 2016, Morten Messerschmidt, the Danish MEP of the now invalid movement, used the European funds to support the national movement of liberalism and democracy (MELD), had to return this money.*

In Sentence 7-28, the word Europa (Europe) was omitted. While the original name of the party was Movimiento por una Europa de Libertades y Democracia (Europe of Freedom and Direct Democracy), GT read as 자유주의와 민주주의의 국가 운동 (movement of liberalism and democracy).

Sentence 7-29

ST «En este momento tenemos tres actores diferentes que verifican el uso de los Fondos de la UE: Dirección General de Finanzas (DGFIN), el Comité Independiente de Personas Eminente y también la Autoridad para los Partidos Políticos Europeos y las fundaciones políticas europeas (EUPPF) (...)

GT "현재 EU 재무 기금 (DGFIN), 저명한 인물위원회 (Independent Committee of Eminent Persons), 그리고 유럽 정당 및 유럽 정치 재단 (EUPPF) 당국의 EU 기금 사용을 검증하는 3 가지 행위자가있다. (...)

BT *"There are now three actors to verify the use of EU funds by the DGFIN, the Independent Committee of Eminent Persons and the European Political and European Political Foundation (EUPPF)*

In Sentence 7-29, three pronouns were omitted: Dirección (directorate), General (general) and Independiente (independent). GT translated the first group as 재무 기금 (financial fund), and the second group as 저명한 인물위원회 (Independent Committee of Eminent Persons) with its English translation inside the brackets. The first case was either an omission and mistranslation.

Sentence 7-30

ST la investigación del supuesto complot de la campaña de Trump con Rusia, cuyos resultados no tardarán en llegar.

GT 주장 음모 캠페인 트럼프와의 조사를 그의 결과는 오래 가지 않을 것이다.

BT *The results of the investigation with Trump and the assassination plot campaign will not last long.*

The omission of proper nouns, previously observed in the examples of Sentence 7-28 and 7-29, could be understood as a problem of out-of-vocabulary (OOV) words. Sentence 7-30, however, omitted *Russia* in the translation, which was hard to justify.

Sentence 7-31

ST El porcentaje es el más bajo en lo que va de su administración.

GT 그 비율은 행정부에서 지금까지 최저치이다.

BT *The ratio is the lowest in the administration so far.*

In relation to the possessive pronouns (_PP$), the problem observed in the dataset was equally due to vagueness of the 3rd person singular pronoun su (i.e. his, her, its). In Sentence 7-31, the omitted segment su administración (his administration) represented either *his administration* or, more directly, *Moon's administration*. However, the translation of the possessive pronoun was omitted in GT even though BT translated it as *the administration*.

Sentence 7-32

ST La tasa de apoyo al presidente Moon Jae In y a su partido, The Minjoo, cayó al nivel más bajo desde mayo de 2017, mes en que se celebraron elecciones presidenciales prematuras tras la destitución de la ex presidenta Park Geun Hye.

GT 대통령 문재인 일행의 Minjoo에 대한 지원의 비율 월 2017 년 이후 가장 낮은 수준으로 떨어졌습니다, 전 대통령 박근혜의 축출 이후 한 달 조기 대통령 선거가 일어났다.

BT *The percentage of support for Minjoo, the president 's party, has fallen to its lowest since May 2017, a month ahead of the presidential election of former President Park Geun - hye.*

In Sentence 7-32, the same pronoun su was omitted. A desirable translation of the segment su partido would be 그의 정당 (his party), which represented the ruling party, The Minjoo. In GT, however, the segment a su partido was combined with presidente Moon Jae In (President Moon Jae In) to produce *support for the president's party*.

Sentence 7-33

ST «No existe un solo partido euroescéptico, existen veintitantos, cada uno con su sensibilidad.

GT "유로 이탈리즘 당이 하나도 없으며, 감성이있는 스무 명의 사람들이 있습니다.

BT *"There is not one Euro break, and there are 20 people with emotions.*

The situation was slightly different in Sentence 7-33, where the pronoun su did not refer to a specific person but a person in general. As such, the translation of the segment con su sensibilidad would be 자신만의 감각 (with one's own sense) if the nuance was considered. In anyhow, all these errors of omission occurred with the pronouns seemed detrimental to the output that could distort the meaning of the translation.

Determiner Three types of tag sets for the determiner were frequently detected: a singular determiner (_DT) (i.e. this, that), a singular or plural determiner (_DTI) (i.e. some, any) and a plural determiner (_DTS) (i.e. these, those, them). This type of omission should be categorized as an incorrect translation.

Sentence 7-34

ST [SUB-HEADLINE] Los estadounidenses eligen este martes a congresistas y autoridades estatales y locales.

GT 미국인들은 화요일에 국회의원과 주 및 지방 당국을 선택합니다.

BT Americans choose lawmakers, state and local authorities on Tuesday.

In Sentence 7-34, the singular determiner este (this) was omitted. A desirable form of GT was 이번 화요일에 (this Tuesday). The translation was understandable without the determiner, but strictly speaking, it was erroneous and did not convey all the meaning expressed in the source text.

Sentence 7-35

ST Las corrientes más izquierdistas del partido demócrata optarán por ese camino, lo que convertiría la segunda mitad del mandato de Trump en un circo político que estimularía al electorado republicano en defensa de su presidente.

GT 민주당의 가장 좌파 전류는 대통령의 방어 공화당 유권자를 자극 할 정치적 서커스에서 트럼프의 위임 하반기을 만들 것이라고 경로를 선택합니다.

BT *The Democratic Party's most leftist election would make the second half of the trump's mandate in a political circus that would stimulate the President's defense Republican voters.*

In a similar way, the determiner ese (that) was omitted in Sentence 7-35. As such, the source text read as 경로를 선택합니다 (select path), which sounded awkward. Interestingly, BT did not produce the translation of this segment.

Sentence 7-36

ST Las autoridades han dicho no tener pruebas de fraude alguno, aunque ese condado ha estado envuelto en polémicas en el pasado.

GT 당국은 과거에 논란에 휘말린 상태이지만 사기에 대한 증거가 없다고 말했습니다.

BT *Officials say they have been involved in controversy in the past, but there is no evidence of fraud.*

In Sentence 7-36, two determiners were omitted: _DTI and _DT. Firstly, the determiner alguno (any) was omitted. The translation might sound correct without it, but as such determiner emphasize the meaning of the modifying noun pruebas (evidence), a desirable form would be 어떠한 증거도 (any evidence). Secondly, ese (that) was omitted. This was the same case as Sentence 7-35. In Sentence 7-36, however, the modifying noun condado (county) was also omitted.

Article In the case of the article, the fact that Korean did not use the article made its omission a correct translation. Most of the articles in the corpus, therefore, was not calculated as an omission. The type of articles that this section strived to discuss was those with explicit meaning that contributed to a nuance of the sentence.

Sentence 7-37

ST [SUB-HEADLINE] Pero los comicios son ante todo un juicio al «trumpismo« que determinará el futuro del presidente

GT 그러나 선거는 무엇보다도 대통령의 미래를 결정할 "트럼프즘"에 대한 재판이다.

BT *But the election is, above all, a trial of Trumpfism, which will determine the president 's future.*

The omitted article los (the) in Sentence 7-37 was a definite article. As it designated a specific election (los comicios), the midterm election, in this context, the translation should represent it explicitly to the translation as 이/그 선거 (this/that election). The nuance of GT was not expressed in BT as English also had an article.

Sentence 7-38

ST En concreto, la acusación contra el político se centra en no respetar el periodo oficial de campaña al presentar sus promesas electorales y visiones antes de que el mismo empezara, y por divulgar falsa información al comentar sin pruebas objetivas que en la afluencia masiva de capital chino en el desarrollo urbano de Jeju estuvieron implicados, tanto el político que fue su rival en las elecciones, como un ex gobernador de Jeju.

GT 특히 정치인에 대한 비난은 공식 선거 운동 기간을 존중하지 않는 데 초점을두고있다. 선거 전의 약속과 비전을 시작하기 전에 제시하고, 중국 자본의 대규모 유입 선거에서 경쟁자였던 정치인이자 전 총재직이었던 제주도의 도시 개발에 참여했다.

BT *In particular, criticism of politicians is focused on not respecting the official campaign period. Before the election promises and visions were introduced, he participated in the city development of Jeju Island, a politician and former governor who was a competitor in the massive influx of Chinese capital.*

In Sentence 7-38, the role of the article in el político (the politician) was to delimit the word to a certain politician. However, the omission of this article in GT resulted in an entirely different sentence, referring to a politician in general, as seen in BT: criticism of politicians. The translation needed to specify the word with some determiner such as 그 정치인 (that politician) or it was a good way in Korean to repeat the proper noun.

Sentence 7-39

ST En tanto, en las legislativas parciales, que por primera vez en la historia tuvieron lugar simultáneamente a las regionales, la participación marcó el 60,7%, rebasando en gran medida el 53,9% de las últimas celebradas el 12 de abril del año pasado.

GT 한편, 역사상 처음으로 지역과 동시에 일어난 부분 입법부에서는 참여율이 60.7 %로 4 월 12 일의 53.9 %를 크게 웃돌았다 과거의

BT *On the other hand, for the first time in history, the participation rate in the partial legislative branch, which occurred simultaneously with the region, exceeded the 53.9%*

In Sentence 7-39, the omission of three meaningful articles led to poor adequacy of GT. It was a different case from Sentence 7-37 and 7-38 in that the role of the articles in Sentence 7-39 was to notify that there was an omitted noun in ST. In the first segment las legislativas parciales (the partial legislative), the complete form was *the partial legislative elections.* Likewise, the second segment las regionales (the regional) was a reduced form of *the regional election*. The third segment las últimas celebradas (the last held) was a reduced form of *the last held election*. It seemed that GT did not recognize the ellipsis of such noun and translated those phrases as a noun without the translation of the articles. A desirable translation would be to reproduce the noun and translate the article.

Sentence 7-40

ST Hasta ahora, los de Rivera se han mantenido firmes en su intención de no volver a apoyar bajo ningún concepto a la dirigente socialista.

GT 지금까지 리베라는 어떤 식 으로든 사회주의 지도자를지지하지 않을 것이라는 확고한 견해를 유지해 왔습니다.

BT *So far, Rivera has maintained a firm belief that it will not support socialist leaders in any way.*

Similar but not similarly, it was clear that the segment in Sentence 7-40 los de Riviera (those of Riviera) possessed an omitted noun after the article los. Considering the context, it was partidos (parties —it was written in a plural form presumably because the name of his party *Ciudadanos* was plural). GT, however, omitted this segment entirely and translated it as Rivera. A desirable form in Korean would be 리베라의 당 (the party of Riviera) or 시우다다노스 (Ciudadanos).

Sentence 7-41

ST Eurodiputados de los verdes y socialdemócratas dicen exactamente «pagué 200 euros por un traje», pero no están obligados a hacerlo.

GT 녹색당과 사회 민주당 원들의 MEP들은 정확하게 "나는 200 유로를 지불했다"고 말하지만 그렇게 할 의무는 없다.

BT *The MEPs of the Greens and Social Democrats say precisely "I paid 200 euros," but I have no obligation to do so.*

In terms of an indefinite article, Sentence 7-41 clearly showed that it should be explicitly translated. The segment of Sentence 7-41, un traje (a suit), should be translated as 정장 한 벌, but it was omitted as a whole from GT.

Verb Considering the important role a verb had in a sentence, the omission of the verb was lethal to the translation. Four types of verb-related POS tags were commonly detected in the dataset: a base form (_VB), past tense (_VBD), past participle (_VBN) and present participle gerund (_VBG). Most of the errors were caused in _VB with 70%. The reason for the omission was inexplicable, but an interesting question arose as to how the system managed to complete the sentence without the verb if it did. The focus of this

section was, therefore, to study the strategies of GNMT of compensating the verb omission. Three strategies were detected: a complete elimination, a partial elimination, and a substitution. Each solution was discussed subsequently.

Complete Elimination It was the most convenient way to compromise the omitted verb while maintaining the fluency of the translation. 15 out of 35 sentences with the verb omission fit in this category with 42.8%. Some of the most relevant examples were given.

Sentence 7-42

ST El Eurobarómetro dice que los temas más importantes son el empleo, el crecimiento económico, la gestión de la migración, el cambio climático, medio ambiente y protección social», afirma a ABC.

GT 유로 바로 메타는 가장 중요한 문제는 고용, 경제 성장, 이주 관리, 기후 변화, 환경 및 사회 보호라고 말했습니다.

BT *Eurobalmeta said the most important issues are employment, economic growth, migration management, climate change, environmental and social protection.*

In Sentence 7-42, the inserted clause afirma a ABC (states to ABC) was eliminated as a whole. GT made a sacrifice of adequacy for the sake of fluency.

Sentence 7-43

ST Si bien la competencia es reñida, los expertos prevén que el Partido Demócrata renovará la mayoría en la cámara baja por primera vez en ocho años -suponiendo un gran obstáculo para Trump, que busca la reelección en 2020-, mientras que estiman que el Senado mantendrá el predominio republicano.

GT 경쟁이 치열한 상황에서 전문가들은 8 년 만에 처음으로 민주당이 하원에서 다수를 갱신 할 것으로 예상한다. 트럼프는 2020 년 재선을 모색하는데 큰 걸림돌이된다. 공화당 우세

BT *In a competitive environment, experts expect the Democrats to renew their majority in the House for the first time in eight years. Trump is a major obstacle to re-election in 2020. Republican dominance*

The long phrase was entirely omitted as well in Sentence 7-43. While the clause mientras que (while) was omitted, not to mention that the verbs estiman (to estimate) and mantendrá (to maintain) were deleted, words (el predominio

republicano) were partially left in GT, which severely damaged the fluency of the sentence.

Partial Elimination Unlike the complete deletion of the segment that included a verb, the partial elimination referred to a deletion of the verbal part only. It resulted in a sentence without a verb, which severely damaged the fluency of a sentence.

Sentence 7-44

ST Quienes mantienen un empate técnico son PP y Ciudadanos, los primeros con un 18,66% de voto y el segundo con un 18,55%, que se traduce en idéntico número de representantes en el Parlamento andaluz: una horquilla que va de los 20 a los 22.

GT 기술적 인 관계를 유지하는 사람들은 PP와 시민이며, 18.66 %의 투표권을 가진 첫 번째 사람과 18.55 %의 사람인 두 번째 사람은 안달루시아 의회 의원과 같은 수의 사람으로 번역됩니다. 머리말은 20 22시에

BT *The people who maintain the technical relationship are the PP and the citizen, the first person with 18.66% of the vote and the second person with 18.55% will be translated into the same number of people as the Andalusian parliamentarians. The heading is*

In Sentence 7-44, the verb va (to go) was not visible from GT, and the sentence was left partially translated. Such destructured translation obtained the lowest fluency and adequacy scores in the fluency and adequacy scoring by the evaluators. The average fluency score was 2.17 (of 4), which was equivalent to 54.25% fluency. The average adequacy score was 1.67 (of 4), which was equal to 41.75%. Such low scores also represented other errors in the sentence, but the unfinished translation due to the omission of the verb was assumed to have influenced the low fluency score in particular.

Sentence 7-45

ST Pese a que el PSOE volvería a ganar las elecciones, la valoración de los votantes respecto de la gestión de los socialistas en esta última legislatura es mayoritariamente regular (39,8%), mala (27,9%) o muy mala (15,2%).

GT PSOE가 다시 선거에서 승리 할지라도,이 마지막 기간에 사회 주의자들의 관리에 관한 유권자 평가는 대부분 규칙적 (39.8 %), 나쁜 (27.9 %) 또는 매우 나쁘다 (15, 2 %).

BT *Even if the PSOE is to win again in the elections, voter ratings on the management of socialists in this last period are mostly regular (39.8%), bad (27.9%) or very bad (15, 2%).*

In Sentence 7-45, the omission of es (to be) also caused a disfluent translation of GT. There should be 이다 (is) at the end of the sentence. Although the rest of the segments were well translated, the omission of the verb affected largely the fluency of the sentence, resulting in a relatively low fluency score of 2.83 (of 4), which was 70.75%, in the MT evaluation.

Substitution It was the most interesting way to deal with the verb omission, in that there was evidence that the system *tried* to make up for such omission. The strategy was that a surrounding verb, noun or adjective substituted the original verb.

Sentence 7-46

ST Su lenguaje de confrontación, el insulto habitual, las referencias de tono racista y sexista y las filtraciones sobre el caos en la Casa Blanca han copado los medios, que aman odiar al presidente.

GT 그의 대립 언어, 통상적 인 모욕, 인종 차별주의 및 성 차별주의의 언급, 백악관의 혼돈에 대한 누출이 언론을 점령 해 대통령을 싫어한다.

BT *His oppositional language, common denigration, racism and sexism, and leaks to the chaos of the White House hate the president by occupying the media.*

In Sentence 7-46, there were two verbs: aman (to love) and odiar (to hate). The original meaning of the segment aman odiar was *love to hate*, GT omitted the verb aman. The translation, therefore, became an entirely different sentence from “they love …” to “they hate …”.

Sentence 7-47

ST Las encuestas dicen que los demócratas recuperarán la Cámara de Representantes, y la duda es cuál será la ventaja final.

GT 여론 조사는 민주당 원들이 하원 의원을 되찾을 것이고, 의문점은 최종 우위가 될 것이라고 말했습니다.

BT *The poll said the Democrats will reclaim the House of Representatives, and the question will be the ultimate advantage.*

In Sentence 7-47, the verb es (to be) was omitted from the translation, and the

following verb será (to be) substituted it. The original meaning of the segment was *[…] the question is what will be the ultimate advantage*, but as BT partly showed, GT read as *[…] the question will be the ultimate advantage*, omitting the interrogative pronoun cuál as well.

Sentence 7-48

ST A este último, ya disuelto, se le impuso la devolución 1,1 millones de euros al Parlamento Europeo por uso indebido de dinero europeo.

GT 이미 해체 된 후자는 유럽의 돈을 오용하여 유럽 의회에 110 만 유로를 상환했다.

BT *The latter, which had already been disbanded, misused European money and repaid € 1.1 million to the European Parliament.*

While the previous examples showed the substitution of a verb from another verb, Sentence 7-48 was an example of a substitution of a verb from a nearby noun. The omitted verb, in this case, was se impuso (to impose). While ST read as *a money return was imposed*, GT read as *money was returned*.

Sentence 7-49

ST El descenso se debe, según Realmeter, a la reforma del sistema de pensión nacional y a la confianza de la población en el actual Gobierno, que decrece tras ser declarado inocente el ex gobernador de Chungcheong del Sur, Ahn Hee Jung, en su juicio por abuso sexual.

GT Realmeter에 따르면, Realmeter에 따르면, 국민 연금 제도의 개혁과 현 정부의 국민에 대한 신뢰 (무죄로 선언 된 후 감소 함), 남 충의 전 총리 안희정 (An Hee Jung)은 학대에 대한 재판에서 성적인

BT *According to Realmeter, according to Realmeter, the reform of the national pension system and the trust of the present government (decreasing after being declared innocent) and the former prime minister of South Korea An Hee Jung,*

Similarly, in Sentence 7-49, the verb se debe a (to be due to) was substituted by the preceding noun descenso (decline). As such, GT read as *decreased after […]* while the original meaning was *the decline was due to […]*. The translation, however, was somehow understandable and did not harm much the original meaning of the text.

Sentence 7-50

ST se tradujo en altas tasas de participación electoral para las regionales, superando incluso el 20% en la votación anticipada.

GT 지방 선거 투표율이 높았고 조기 투표에서 20 %를 능가했다.

BT *Local elections had a high turnout and surpassed 20% in early voting.*

In Sentence 7-50, GT omitted the verb se tradujo (to translate) and used the adverb altas (high) instead as the main verb so that it read as *[…] rate was high*. ST, on the other hand, read as *it resulted in high rates […]*. Interestingly, BT generated a sentence similar to ST by using the verb "have": *had a high turnout*.

Sentence 7-51

ST En unas elecciones de 2016, eliminó un archivo de papeletas físicas pero mantuvo uno digitalizado.

GT 2016 년 선거에서 그는 실제 투표 용지 파일을 삭제했지만 하나는 디지털화 했습니다.

BT *In the 2016 election, he deleted the actual ballot file, but one digitized it.*

Similar to Sentence 7-50, the nearby adjective became the main verb instead of the original verb in Sentence 7-51. While mantuvo (to maintain) was deleted, digitalizado (digitized) acted as the verb so that GT read as *[…] but digitized one*. BT, in this case, mistranslated GT: *one digitized it*. The original meaning of this segment was *[…] but maintained one digitized*.

7.2.2.3. Mistranslation

Mistranslation could embrace, in theory, a large part of errors due to its definition. The definition of mistranslation in this thesis was limited to an incorrect translation that did not interfere with other categories. Such definition delimited the scope of the mistranslation to incorrect word choice. There were some exceptions such as four English translations that translated a Spanish word into English or incorrect use of the word. As they were statistically insignificant, they were understood in the boundary of wrong word choice.

There were 783 mistranslated items in the dataset. As an instance, Sentence 7-52 was given below. The mistranslation was detected in una amplia mayoría (the large majority). While the word "majority" could be used not only in

daily life but also in the election, the Korean word in GT 대다수 was not appropriate in the context of an election. A better choice would be 과반수. As this kind of discussion could lean toward an individual preference, mistranslation was not discussed in this thesis.

Sentence 7-52

ST	[HEADLINE] El CIS otorga a Susana Díaz una amplia mayoría en las elecciones andaluzas
GT	CIS는 Susana Díaz에게 안달루시아 선거에서 대다수를 준다.
BT	*CIS gives Susana Díaz a majority in the Andalusian elections.*

7.2.2.4. Untranslated

This type of error referred to a content that was left unchanged from the source to the target sentence. Out of 116 untranslated words in the dataset, 74 words (64.79%) could not be regarded as an error because they provided additional information of the word with its original spelling.

Sentence 7-53

ST	El pasado agosto, fue asesinado en un atentado todavía sin esclarecer el que había sido jefe de la autoproclamada República Popular de Donetsk (DNR) durante casi cuatro años, Alexánder Zajárchenko.
GT	지난 8 월, 알렉산더 자 자르 첸코 (Alexander Zajárchenko)는 자신이 주장한 도네 치크 인민 공화국 (Donetsk People 's Republic, DNR)의 머리로 4 년 가까이 머물렀다는 점을 아직도 분명히 모르는 공격에서 암살 당했다.
BT	*In August, Alexander Zajárchenko was assassinated in an attack that he still did not know clearly that he stayed for four years as head of the Donetsk People 's Republic (DNR).*

In Sentence 7-53, for instance, four words were untranslated. Firstly, the name of a person Alexánder Zajárchenko was left inside the brackets, which made the translation not an error but rather an elaborate work. In the second segment, the pronoun República Popular de Donetsk was partially untranslated inside the brackets with its acronym DNR. It would have been acceptable if the segment was entirely untranslated as the previous case, but the fact that a part of the

segments was translated in English when the translation was from Spanish to Korean made it erroneous. The acronym, on the other hand, was regarded as acceptable. As such, the English except for Donetsk, the rest three untranslated items should be considered as a correct translation.

Most of the untranslated words were proper nouns. This section focused on how the engine treated the proper noun, which could give insights to the treatment of OOV of the NMT engine in the *es-ko* pair. There were two most common strategies that this engine adopted: phonetic transcription and raw source word. Each strategy was discussed with examples, subsequently.

Transcription It was found that 44.8% of the untranslated segments were transcribed in Korean. Some of the most relevant examples were given below.

Sentence 7-54

ST Esas elecciones fueron polémicas desde el primer momento porque el candidato republicano, Brian Kemp, como secretario de Estado de Georgia, es el responsable de gestionar comicios y los demócratas le acusaron de tomar medidas para restringir el voto de la población negra, que iba a ser clave para Abrams, que sería la primera gobernadora afroamericana de EE UU.

GT 그 선거는 공화당 출신 후보 인 브라이언 켐프 (Brian Kemp)가 조지아 국무 장관으로 선거 운영에 책임이 있고 민주당이 흑인 인구의 투표를 제한하는 조치를 취했다고 비난했기 때문에 처음부터 논란이 많았다. 미국 최초의 아프리카 계 미국인 총재가 될 아브람에게 열쇠.

BT *The election was controversial from the outset because Republican candidate Brian Kemp accused Georgia of being the secretary of state and responsible for running the elections and the Democrats taking steps to limit the voting of the black population. The key to Abram, the first African-American president in America.*

In Sentence 7-54, the name Brian Kemp was left untranslated inside the brackets. This was a rather complete translation that translated (or could be transcribed) the proper noun in Korean and additionally showed its original spelling inside the brackets.

Sentence 7-55

ST Mientras en Georgia, la demócrata Stacey Abrams ha prometido tomar acciones legales para garantizar que todos los votos se cuenten correctamente.

Sentence 7-55

ST Mientras en Georgia, la demócrata Stacey Abrams ha prometido tomar acciones legales para garantizar que todos los votos se cuenten correctamente.

GT 조지아 주에서 민주당의 스테이시 에이 브람스 (Stacey Abrams)는 모든 득표가 올바르게 집계 될 수 있도록 법적 조치를 취할 것을 약속했습니다.

BT *Democrat Stacey Abrams in Georgia promised to take legal action to ensure that all votes are counted correctly.*

In Sentence 7-55, it seemed clear that the name Stacey Abrams was transcribed, seeing the wrong spacing of 에이 브람스 and that the same name was written in Sentence 7-54 as 아브람. This, again, was a desirable way of translation.

Sentence 7-56

ST El líder regional de Ciudadanos, Juan Marín, y el coordinador general de IU en Andalucía y número dos de Adelante Andalucía, Antonio Maíllo, obtienen una calificación de 3,5, seguidos de la cabeza de lista por la confluencia de izquierdas, Teresa Rodríguez, con un 3,4, y el candidato popular, Juan Manuel Moreno, 3,1.

GT 후안 마린과 안달루시아 IU 총지배인 인 아델란테 안달루시아 (Andelucía), 안토니오 메이 올로 (Antonio Maíllo)의 지역 지도자는 3.5 점을 얻었고 왼쪽의 합류점 인 테레사 로드리게스 (Teresa Rodríguez) 3.4, 인기 후보 인 후안 마누엘 모레노 (3.1).

BT *Andalucía and Antonio Maíllo, regional leaders of Juan Marin and Andalusia IU general managers, got 3.5 points, Teresa Rodríguez 3.4 on the left, Juan Manuel Moreno (3.1) .*

In Sentence 7-56, there were many proper nouns in one sentence. Firstly, the noticeable point was that not all names were provided with their original spelling inside the brackets. While the shadowed names had an untranslated translation in GT, the underlined names were transcribed to Korean. The interesting point was found with Adelante Andalucía. While it was the name of a Spanish party, only a part of the word was provided with its original spelling, but in an incorrect way: Andalucía > Andelucía. Such a case was considered a mistranslation.

Raw Source Word The transcription was the best strategy of dealing with a person's name, as previously seen, and the untranslated parts provided excellent information. On the other hand, there were cases where the proper noun was left as it was without their transcription in the target language. Some

of the cases were acceptable as a translation while others harm the quality of a translation.

Sentence 7-57

ST [SUB-HEADLINE] La encuesta sitúa a la coalición de Podemos e IU en segundo lugar y anticipa un empate técnico entre PP y Ciudadanos

GT 이 설문 조사는 Podemos와 IU의 연합을 두 번째 장소에두고 PP와 Citizens 사이의 기술적 인 관계를 예상합니다

BT *This survey puts the coalition of Podemos and IU in the second place and expects a technical relationship between PP and Citizens*

In Sentence 7-57, there were four names of Spanish political parties. While three of them (Podemos, IU, PP) were left untranslated, Ciudadanos was translated in English as *Citizens*. Firstly, we thought that Podemos should be transcribed to Korean as 포데모스. Secondly, IU should be translated to 좌파연대 (United Left) accompanied by the acronym. Thirdly, PP could be translated to 인민당 (People's Party), but considering their popularity, the acronym alone could convey the original meaning. Lastly, Ciudadnos should be transcribed into 시우다다노스, which was why it was regarded as a mistranslation. As seen from this example, the proper nouns of the same level had their own best option of translation. Some of them could be left as a raw source word while others should be transcribed.

Sentence 7-58

ST Podemos obtuvo 15 diputados, mientras Ciudadanos e IU obtuvieron nueve y cinco respectivamente.

GT Ciudadanos와 IU가 각각 9 점과 5 점을 얻은 동안 우리는 15 명의 대리인을 얻을 수 있습니다.

BT *We can get 15 agents while Ciudadanos and IU get 9 and 5 respectively.*

The same proper noun Ciudadanos was left as it was in Sentence 7-58. It was a better option than translating it to an English word but still was an error. The more erroneous case was Podemos which was regarded as a modal verb poder (can) in 1st person plural. It was an obvious mistranslation.

7.2.2.5. Punctuation

The use of punctuation differed from language to language. The punctuation error constituted a trivial part of the error classes in this study with 34 incidents. They were erroneous because i) it should have been omitted, ii) it should have been adapted to a Korean way or iii) it was omitted from the translation. Note that the error class of punctuation was prioritized over other classes such as mistranslation and omission so that all punctuational errors could be dealt with in this section.

Should have been omitted Two types of punctuations were detected in the current corpus that could not be translated into punctuation in Korean: colon (:) and semicolon (;). As they did not have an equivalence, they had to be expressed in words in Korean. That is, if they remained in the translation, they were erroneous.

Sentence 7-59

ST De forma divertida, el partido más euroescéptico extremista húngaro es antieslovaco y el eslovaco, antihúngaro;

GT 재미있는 방법으로, 대부분의 Eurosceptic 극단적 인 헝가리 당은 반대로 슬로바키아어이고 반대로 헝가리 인 슬로바키아어이다;

BT *Interestingly, most Eurosceptic extreme Hungarian parties are Slovak, which is anti-Slovak and anti-Hungarian;*

In Sentence 7-59, a semicolon was situated at the end of the sentence. In practice, more words existed after the semicolon, but it was segmented as such in order to avoid too long sentences. GT also had the punctuation, which made the sentence erroneous. The simple deletion was a desirable way to revise the error.

Sentence 7-60

ST 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:

GT 그러나 현재 연합에 의해 설계된 마스 트리 히트 조약 발효 25 주년과 유럽 단일 통화인 유럽이 방금 만났을 때 EU는 다가오는 유럽 선거에서 역사상 가장 큰 도전 중 하나에 직면 해있다. :

BT *However, the EU is facing one of the biggest challenges in history in the upcoming European elections, when the 25th anniversary of the Maastricht Treaty, now designed by the Union, and the European single currency, Europe, have just met. :*

Similarly, in Sentence 7-60, a colon was left at the end of the sentence. The difference was that GT finished the sentence with a period and added the colon next to it. The same went with BT. It was erroneous, but it could be speculated that the engine regarded it importantly in translation.

Sentence 7-61

ST Para Berresheim, los eurodiputados tienen todas las mismas reglas sobre cómo gastar su dinero;

GT Berresheim에게있어 MEP는 돈을 쓰는 법에 대해 모두 동일한 규칙을 가지고 있습니다. "

BT *For Berresheim, MEP has all the same rules about how to spend money. "*

In Sentence 7-61, the semicolon was left as the case of Sentence 7-59. The difference was that it was translated into a quotation mark.

Adaptation The strategy of adaptation referred to punctuation that should be adapted to a Korean style. Two types of such kind were detected in the corpus: dash (-) and guillemets («, »). The dashes were usually translated to brackets, and the guillemets were converted into quotation marks (“, ”). They were found erroneous either because they were adapted to wrong punctuation or because they were not converted.

Sentence 7-62

ST La batalla se juega en el terreno del segundo puesto, donde las tres formaciones principales, PP, Adelante Andalucía —la confluencia entre Podemos e IU— y Ciudadanos llegan muy igualadas, según el sondeo.

GT 이 전투는 2 위의 지형에서 진행되는데 PP, Adelante Andalucía (Podemos와 IU- 시민들의 합류는 매우 평평하다.

BT *This battle takes place in the second topography, PP, Adelante Andalucía (Podemos and IU-citizens are very flat at the confluence.*

In Sentence 7-63, one dash was translated to a bracket, but the other was left unchanged.

Sentence 7-63

ST Sobre todo tras un cierre de campaña virulento, agitado por la campaña del miedo desatada por Trump a raíz de la caravana de inmigrantes y por los episodios de violencia —los paquetes bomba a sus enemigos politicos y, sobre todo, la matanza en la sinagoga de Pittsburgh— que parte de la opinión pública relaciona con el mensaje agresivo del presidente.

GT 특히 악의적 인 선거 운동 종료 후, 이민자들의 캐러밴과 폭력 사건의 결과로 트럼프가 쏟아내는 두려움에 대한 캠페인에 동요 됨. - 정치적 원수 폭파, 특히 무엇보다도 시나고그의 학살 피츠버그 - 여론의 일부는 대통령의 공격적인 메시지와 관련이 있습니다.

BT *Especially after the end of the malicious election campaign, the campaign against the fears of Trump as a result of immigrants' caravans and violent incidents was shaken. - Political bombing, especially above all the slaughter of the Sinaigg Pittsburgh - Some of the public opinion is related to the president's aggressive message.*

The same dashes were left unchanged in Sentence 7-63. The reason was unknown, but the reaction of the engine towards the punctuation was quite intriguing.

Sentence 7-64

ST «Pretended que estoy en la votación», pidió a sus seguidores.

GT «내가 투표 한 척해라»그는 추종자들에게 물었다.

BT *«Pretend I voted» he asked his followers*

In the case of the guillemets, they remained as they were in GT of Sentence 7-64. A desirable way was to change them into quotation marks.

Sentence 7-65

ST El «presidente» en funciones de Donetsk y favorito para obtener la mayoría de los votos, Denís Pushilin, explicó a comienzo de mes que la república «necesita celebrar estas elecciones» para dotarse de líder y asamblea local.

GT 도네츠크의 대통령과 대다수의 투표를 얻으려는 데니스 푸 스틸 린 (Denis Pushilin)은 이달 초 공화당이 "선거를 개최해야한다"고 설명하면서 지도자와 지방 의회가되었다.

BT *Dennis Pushilin, the president of Donetsk and the vast majority of voters, became leaders and provincial councils earlier this month, explaining that the Republicans "must hold elections".*

In Sentence 7-65, there were two pairs of guillemets. Firstly, they were used as an emphasis of the word presidente (president), which was omitted in GT. They should have been translated to quotation marks. Secondly, they were used as a direct quotation, which was correctly translated into quotation marks.

Omission A large part of the errors in the punctuation error class stemmed from being omitted. The guillemets were occasionally omitted in the translation. Sometimes, some sentences did not have a period at the end of the sentence. Meanwhile, the omission of the comma was not considered as erroneous.

Sentence 7-66

ST «Las enmiendas propuestas se centran en proporcionar más transparencia, mejorar la legitimidad democrática y fortalecer la aplicación de la ley», según el informe sobre estatuto y el financiación de partidos políticos y las fundaciones europeas redactado por Wieland y Bresso.

GT "제안 된 개정안은 Wieland와 Bresso에 의해 작성된 정당과 유럽 재단의 법령과 기금에 관한 보고서에 따르면,보다 투명성을 제 공하고, 민주적 정당성을 향상시키고, 법률 적용을 강화하는 데 중점을 둡니다.

BT *"The proposed amendment focuses on providing greater transparency, enhancing democratic legitimacy, and strengthening legislation," according to a report by parties and European Foundation statutes and funds written by Wieland and Bresso.*

In Sentence 7-66, one guillemet was not reproduced in GT. The interesting point was that BT somehow reproduced the nonexistent guillemet.

Sentence 7-67

ST El portavoz del Kremlin, Dmitri Peskov, dijo el martes que las elecciones organizadas por los separatistas «no vulneran los acuerdos de paz».

GT 드미트리 페스 코프 (Dmitri Peskov) 러시아 대변인은 화요일 분리 주의자들에 의해 조직 된 총선은 평화 협정을 위반하지 않는다고 밝혔다.

BT *A Russian spokesman, Dmitri Peskov, said Tuesday that a general election organized by separatists does not violate the peace treaty.*

In Sentence 7-67, the guillemets entirely disappeared from GT. The engine transformed the direct quotation to an indirect quotation, which was not erroneous from the perspective of fluency.

Sentence 7-68

ST	En tanto, en las legislativas parciales, que por primera vez en la historia tuvieron lugar simultáneamente a las regionales, la participación marcó el 60,7%, rebasando en gran medida el 53,9% de las últimas celebradas el 12 de abril del año pasado.
GT	한편, 역사상 처음으로 지역과 동시에 일어난 부분 입법부에서는 참여율이 60.7 % 로 4 월 12 일의 53.9 %를 크게 웃돌았다 과거의
BT	On the other hand, for the first time in history, the participation rate in the partial legislative branch, which occurred simultaneously with the region, exceeded the 53.9%

In Sentence 7-68, the omitted punctuation was a period. The sentence was unfinished due to the absence of the period, but also the translation itself was incomplete.

7.2.2.6. Spacing

The spacing in Korean constitutes an important part of semantic disambiguation. There is a famous Korean sentence that shows the role of spacing in that regard. Two examples are given below. They are formally all identical. The sole difference is that the example (1) has a spacing after the particle 가[ga] so that the sentence looks like 아버지가 방에. The example (2), on the other hand, has a spacing before the particle so that the sentence looks like 아버지 가방에. As their GTs show, one spacing difference caused a totally different sentence. In the example (1), the particle *ga* serves as a subject particle and makes *father* the subject of the sentence. However, in the example (2), ga is attached to the other word and becomes a new word —a bag (가방).

(1)		(2)	
ST	아버지가 방에 들어가신다.	ST	아버지 가방에 들어가신다.
GT	Father enters the room.	GT	He goes into his father 's bag.

Albeit such importance of the spacing, most of the incorrectly-spaced words rarely impede comprehension of a sentence to native Koreans. There were 465 spacing errors in the dataset, recording the third most frequent error taxonomy. Some of them were over-spaced and others were under-spaced. Such errors, we believed, were a consequence of a different morphological segmentation of Spanish and Korean. One noticeable pattern of the spacing errors was observed from a numeric expression. Many numeric expressions in the corpus erroneously had spacing between the number and the following word or mark. As Sentence 7-69 showed, 56,4% in ST was translated into 56.4 % in GT. In anyhow, as far as it did not disturb human comprehension, it was considered as a minor error.

Sentence 7-69

ST Sin embargo, para el 56,4% de los electores de la comunidad, lo más importante a la hora de elegir la papeleta serán los temas propios de Andalucía.

GT 그러나 지역 사회 유권자의 56.4 %가 안달루시아의 대상이 될 것입니다.

BT *But 56.4% of the community voters will be targeted for Andalusia.*

7.2.2.7. Grammar

The grammar in the error classification of MQM included morphology, POS, agreement, tense/mood/aspect, word order, and missing, extra, incorrect function words (Lommel et al., 2014). It was thought, nevertheless, that such fine-grained classification would challenge the human annotator in maintaining consistency in the analysis. As an adjusted way, a general error taxonomy suggested by Popović (2018b) was applied for this thesis. They drew a distinction of grammar into three sub-levels: inflection, derivation and composition. The inflection was not of our interest as Korean was an agglutinative language. The composition would be discussed in word order in Chapter 7.2.2.8. Thus, the core study of grammar error detection was on the derivation; especially POS, verb tense and partially on the use of a prefix. Each type was discussed subsequently.

POS Error Prior to the analysis, it was crucial to define what was regarded as a POS error. It was easy to think that it referred to the different POS of the source and target text, but such type of detection could result in a subjective result due to a linguistic divergence between the two languages. The POS error that this section was mainly concerned was a content whose POS should have been changed for the naturalness of the target language but was not. It was to note that such a definition could overlap at some point with the stylistic error in Chapter 7.2.2.9. It was ignored from the calculation of the Style category to have a better picture of how the POS error was produced.

Sentence 7-70

ST Estas elecciones han sido también criticadas e incluso calificadas de «ficticias e ilegítimas» por la Unión Europa, EEUU y las autoridades de Kiev.

GT 이 선거는 유럽 연합, 미국 및 우크라이나 당국에 의해 "허구와 사생아"로 비판 받고 심지어 비난을 받아왔다.

BT *This election has been criticized and even condemned as "fiction and illegitimate children" by the European Union, US and Ukrainian authorities.*

In Sentence 7-70, two adjectives ficticias, ilegítimas (fictitious, illegitimate) were translated into nouns in an incorrect way. Such mistake was presumably because the engine could not recognize the relations between the adjectives and their modifying noun elecciones (election).

Sentence 7-71

ST Lo firmado en Minsk, el 12 de febrero de 2015, contempla la devolución a Ucrania del control de la frontera con Rusia, en los tramos que corresponden a Donetsk y Lugansk, y la celebración en ambos territorios de elecciones realmente libres y democráticas con arreglo a la legislación ucraniana.

GT 2015 년 2 월 12 일 민스크에서 서명 한이 협정은 도네츠크와 루 간츠크에 해당하는 부분에서 러시아와의 국경 통제에 대한 우크라이나로의 회귀와 그에 따른 자유와 민주주의 선거의 두 영역 모두에서의 축하를 고찰한다. 우크라이나 입법.

BT *This agreement, signed by Minsk on February 12, 2015, examines the return of Ukraine to border control with Russia in the areas of Donetsk and Lugansk, and congratulations on both the liberal and democratic elections accordingly . Ukraine Legislation.*

Similarly in Sentence 7-71, while ST read as elecciones libres y democráticas (free and democratic election), GT put the two adjectives into nouns: 자유와 민주주의

(freedom and democracy). In the meantime, BT reproduced them to adjectives somehow.

As seen from the previous examples, the wrong recognition of POS was quite rare in this dataset. A more frequent error was that the engine translated a source word to a target word of the identical POS, but it sounded unnatural to the target language. The majority of these kinds of errors was found with a so-called *nominalization*. While the nominalization was common in Spanish, Korean preferred to use verbs.

Sentence 7-72

ST Su lenguaje de confrontación, el insulto habitual, las referencias de tono racista y sexista y las filtraciones sobre el caos en la Casa Blanca han copado los medios, que aman odiar al presidente.

GT 그의 대립 언어, 통상적 인 모욕, 인종 차별주의 및 성 차별주의의 언급, 백악관의 혼돈에 대한 누출이 언론을 점령 해 대통령을 싫어한다.

BT *His oppositional language, common denigration, racism and sexism, and leaks to the chaos of the White House hate the president by occupying the media.*

RT 적대적인 언어 사용, 의례적인 인신공격성 발언, 인종 차별 및 성 차별적 발언, 백악관의 내부 혼란 유출 등 그의 행동은 대통령을 끌어내리기 좋아하는 언론을 뒤덮었다.

In Sentence 7-72, ST was introduced along with GT, the back translation of GT in BT and a reference translation in RT. The subject of the sentence was composed of four nouns (lenguaje, insulto, referencias, filtraciones). They were translated in GT with their POS maintained so that the adequacy of the translation was high, but the fluency of the sentence was damaged. As RT showed, each noun should be explained with verbs or adjectives. For instance, su lenguaje de confrontación (his language of confrontation) was translated to 그의 대립 언어 (his language of confrontation), but a more natural translation would be 그의 적대적인 언어 사용 (his use of hostile language).

Sentence 7-73

ST Sobre todo tras un cierre de campaña virulento, agitado por la campaña del miedo desatada por Trump a raíz de la caravana de inmigrantes y por los episodios de violencia -los paquetes bomba a sus enemigos politicos y, sobre todo, la matanza en la sinagoga de Pittsburgh- que parte de la opinión pública relaciona con el mensaje agresivo del presidente.

GT 특히 악의적 인 선거 운동 종료 후, 이민자들의 캐러밴과 폭력 사건의 결과로 트럼프가 쏟아내는 두려움에 대한 캠페인에 동요 됨. - 정치적 원수 폭파, 특히 무엇보다도 시나고그의 학살 피츠버그 - 여론의 일부는 대통령의 공격적인 메시지와 관련이 있습니다.

BT *Especially after the end of the malicious election campaign, the campaign against the fears of Trump as a result of immigrants' caravans and violent incidents was shaken. - Political bombing, especially above all the slaughter of the Sinaigg Pittsburgh - Some of the public opinion is related to the president's aggressive message.*

RT 특히 캐러밴 이민 행렬로 트럼프가 퍼트린 두려움과 폭력을 행사한 일화(선거 경쟁자에 폭탄 우편물을 투척했다거나 특히 피츠버그대학살) 등으로 촉발된 공격적인 선거 운동을 뒤에는 말이다. 일부 여론은 대통령의 공격적인 메시지와 연관되어 있다.

In Sentence 7-73, two incidents were inserted inside the dashes: las paquetes bomba (bomb packages) and la matanza en la sinagoga de Pittsburgh (slaughter in a synagog of Pittsburgh). In GT, the translation of the two nouns paquetes and matanza in a literal way impeded the understanding of the sentence. Even though they were correct in theory, a revision was required to naturalize them in the target language. A desirable way would be to add an additional verb to the contents such as 폭탄 우편물 (bomb packages) > 폭탄 우편물을 투척 (to send bomb packages).

Verb Tense Error 27 errors were found in 26 sentences, which was a statistically small amount. The focus of this section was on monitoring a type of verb tense that was found erroneous in the source language and investigating how they were translated to the target language. Five types of verb tense (morphologically speaking) in the source language and three in the target language were detected. The interesting point was that 59% of the errors occurred with future tense in Spanish. Moreover, the engine translated 70% of the errors to present tense.

Sentence 7-74

ST Las corrientes más izquierdistas del partido demócrata optarán por ese camino, lo que convertiría la segunda mitad del mandato de Trump en un circo político que estimularía al electorado republicano en defensa de su presidente.

GT 민주당의 가장 좌파 전류는 대통령의 방어 공화당 유권자를 자극 할 정치적 서커스에서 트럼프의 위임 하반기을 만들 것이라고 경로를 선택합니다.

BT *The Democratic Party's most leftist election would make the second half of the trump's mandate in a political circus that would stimulate the President's defense Republican voters.*

In Sentence 7-74, the future-tense verb optarán (will choose) was translated into a present-tense verb 선택합니다 (choose). Out of 15 future-tense verbs of this category in the dataset, the majority of them were translated into present tense while two were translated to past tense.

Sentence 7-75

ST El partido de Albert Rivera obtendría representación en todas las provincias, con un avance muy significativo en Cádiz, donde pasaría de 1 a 3 o 4.

GT 앨버트 리베라 (Albert Rivera)의 당은 카디스에서 매우 중요한 진보와 함께 모든 지역에서 대표권을 얻었으며 카디스는 1에서 3 또는 4로 넘어갔습니다.

BT *The party of Albert Rivera, with significant progress in Cadiz, received representation in all regions and Kadis went from one to three or four.*

In Sentence 7-75, the two future-tense verbs obtendría (will obtain) and pasaría (will pass) were translated into past-tense verbs: 얻었으며 (obtained) and 넘어갔습니다 (passed).

Sentence 7-76

ST no va a ser tan sencillo ponerlos en común», agrega.

GT 공통점을 두는 것은 그렇게 쉬운 일이 아닙니다. "라고 그는 덧붙입니다.

BT *It is not so easy to have something in common. "He adds.*

An interesting pattern was observed with a phrasal verb such as *be going to* whose morphological form was present but was semantically futuristic. There were two erroneous segments as a consequence of such reason, and Sentence 7-76 was one of them. While ST used va a ser (is going to be), GT translated it to

present tense as 아닙니다 (is not).

Sentence 7-77

ST Aunque esos partidos van a ser más influyentes, el responsable del apartado español del mismo título, Ignacio Molina (Real Instituto Elcano), cree que será más difícil articular una posición común.

GT 비록 그 당사자들이 더 영향력이 있지만, 같은 제목의 스페인어 부문 책임자 인 Ignacio Molina (Real Instituto Elcano)는 공동의 입장을 분명히하기가 더 어려울 것이라고 믿는다.

BT *Although the parties are more influential, Ignacio Molina (Real Instituto Elcano), head of the Spanish section of the same title, believes it will be harder to clarify the position.*

Similarly, Sentence 7-77 also showed that the verb phrase van a ser (are going to be) was translated into simple present tense.

In the case of past tense, almost all of the items were translated into present tense except one. In Sentence 7-78 given below, there were two verbs in the sentence, which were past-tense: acabó (ended up) and hacía (took over). The translation of the latter verb in GT was in past tense, but it was an error. The reason was that the tense agreement did not exist in Korean. Thus, a desirable translation would be to change it to future tense: 맡을지 (will take over).

Sentence 7-78

ST Florida acabó determinado quién se hacía con la presidencia.

GT 플로리다는 누가 대통령직을 맡았는지 결정했다.

BT *Florida has decided who was in the presidency.*

Besides such case, other past-tense errors stemmed from not translating them as past tense.

Sentence 7-79

ST como candidato a la presidencia llamó «mentiroso patológico» y «cobarde llorón» a Trump en 2016 y el mes pasado se vio obligado a invitarlo a un mitin para conservar su puesto como senador en Texas.

GT 대통령 후보로 그는 2016 년에 트럼프에게 "병적 인 거짓말 쟁이"와 "겁쟁이 족제비"라고 부르며 지난달 텍사스에서 상원 의원 자격을 유지하기 위해 그를 집회에 초대해야했다.

BT *As a candidate for president, he called Trump a "morbid liar" and a "coward weasel" in 2016 and had to invite him to the congress last month to remain senator in Texas.*

In Sentence 7-79, the past-tense verb llamó (called) was translated in GT to 부르며 (call) in present tense. BT somehow reproduced the past tense.

Some errors were produced due to perfect tense. That was because the perfect tense did not exist in Korean grammar. Most of such errors occurred because the meaning of the perfect tense was not expressed in a correct manner.

Sentence 7-80

ST A lo largo de la última década, las instituciones de la UE han reforzado sus anticuerpos para enfrentarse a prácticas abusivas con dinero europeo.

GT 지난 10 년 동안 유럽 연합 기관들은 유럽의 돈으로 학대적인 관행을 다루기 위해 항체를 강화했습니다.

BT *Over the last decade, European Union agencies have strengthened their antibodies to deal with abusive practices with European money.*

In Sentence 7-80, there was perfect-tense verb han reforzado (have reinforced). Considering the context that the effort was put for 10 years, the translation should be more than putting it as past tense. A desirable translation of the verb would be 강화해 왔다 that would stress on the continuity of the action. BT somehow translated the past-tense Korean verb to perfect tense.

Sentence 7-81

ST Dos años después de que los ingleses decidieran recuperar su soberanía —¿la habían perdido?—, disponen de solo cuatro meses y medio para elegir entre el vasallaje o el caos.

GT 영국군이 주권을 되찾기로 결정한 지 2 년 후 - 그들은 그것을 잃었을 까? - 그들은 단지 4 개월 반 밖에 걸리지 않았기 때문에 계략이나 혼란 중에서 선택했다.

RT *Two years after the British decided to regain sovereignty - did they lose it? Because they only took four and a half months, they chose between riddles and chaos.*

In Sentence 7-81, the verb was in past perfect tense —habían perdido (had lost) —while the translation in GT was in past tense: 잃었을까. As the role of the past perfect tense was to express the time gap that the verb happened prior to the other verb, the translation should represent the nuance more clearly: 잃었었을까.

Prefix Error There were certain prefixes in this dataset that provoked a grammatical problem. Three prefixes were detected that the system failed to translate in the right way: *euro-, anti-* and *auto-*. Firstly, in relation to the prefix euro-, the majority of the errors was due to the fact that the engine translated it to an English word. 10 out of 14 such items were translated into English. Such behavior additionally manifested that the engine made use of English corpus in the *es-ko* translation.

Sentence 7-82

ST «Mientras que los subsidios europeos debían reforzar el sistema político de la UE y abordar el déficit democrático, los euroescépticos han utilizado principalmente sus recursos para apoyar a sus partidos nacionales», escribe Wouter Wolfs, investigador del Instituto de Gobernanza Pública de KU Leuven, para el Foro contra la corrupción y la integridad de la OCDE, donde resalta la evolución de las formaciones euroescépticas desde hace veinte años, pasando del total boicot a una participación muy intensa en las instituciones y elecciones europeas.

GT 쿠루 벤 (KU Leuven) 공공 거버넌스 연구소의 연구원 인 우스터 울프 스 (Wouter Wolfs)는 "유럽의 보조금은 EU 정치 체제를 강화하고 민주적 인 적자를 다루어야하지만 Eurosceptics는 자국의 정부를 지원하기 위해 주로 자원을 사용했다" OECD의 부패와 청렴성 반대 포럼 - 전체 보이콧에서부터 유럽 제도와 선거에 대한 매우 치열한 참여에 이르기까지 20 년 동안 유로 지성 형성의 진화를 강조합니다.

BT *Wouter Wolfs, a researcher at the KU Leuven Institute for Public Governance, said: "While European subsidies must strengthen the EU political system and deal with democratic deficits, Eurosceptics is mainly using resources to support its governments. "The OECD's Anti-Corruption and Integrity Forum emphasizes the evolution of euro-state formation for 20 years, from the entire boycott to the very intense participation in European institutions and elections.*

In Sentence 7-82, the noun euroescéptico (Eurosceptic) was translated into English. The correct translation of the word would be 유럽연합 회의론자, 유로회의론자 or EU 회의론자. The same vocabulary was translated into Eurosceptic most of the times, but it was also translated in 유로 지성인 (EU intellectual) or 유로 이탈리즘 (EU Italism).

Sentence 7-83

ST Ha destapado fraudes y ahora ha habido un caso en el Tribunal en el que se ha pedido devolver el dinero», dijo Laurin Berresheim, quien prepara el trabajo de Arndt Khon (eurodiputado de los socialdemócratas alemanes) en el Comité de Control Presupuestario.

GT 그것은 사기를 밝혀 냈고 지금은 재판소에 돈을 돌려 줄 것을 요청한 사례가있다 "라고 예산 통제위원회의 Arndt Khon (독일 사회 민주주의 자들의 MEP)의 작업을 준비하는 Laurin Berresheim은 말했다.

BT *It has revealed fraud, and now we have asked to return the money to the court, "said Laurin Berresheim, who is preparing for the work of Arndt Khon (MEP of the German Social Democrats) in the Budget Control Committee.*

In Sentence 7-83, the word eurodiputado (Eurodeputy, MEP) was translated into English, as well. Its Korean translation was 유럽의회의원.

Sentence 7-84

ST Las llamas se observan también en Getlink, gestora del Eurotúnel, por el que anualmente cruzan sin controles fronterizos 1,6 millones de camiones, 2,5 millones de coches y 21 millones de personas.

GT Eurotunnel 매니저 인 Getlink에서도 매년 160 만 대의 트럭, 250 만 대의 자동차 및 2 천 1 백만 명의 사람들이 국경 통제없이 교차하고 있습니다.

RT *Eurotunnel manager Getlink also has 1.6 million trucks, 2.5 million cars and 21 million people crossing the border each year.*

In Sentence 7-84, the word Eurotúnel (Channel Tunnel) translated in English, but in a wrong way: Eurotunnel. A correct English translation was Channel Tunnel, and a correct Korean translation was 채널 터널.

In relation to *anti-*, most of the cases treated the prefix as an independent noun or sometimes adverb. Both of the solutions were incorrect. A desirable translation would be to translate it also as a prefix 반-.

Sentence 7-85

ST De forma divertida, el partido más euroescéptico extremista húngaro es antieslovaco y el eslovaco, antihúngaro;

GT 재미있는 방법으로, 대부분의 Eurosceptic 극단적 인 헝가리 당은 반대로 슬로바키아어이고 반대로 헝가리 인 슬로바키아어이다;

BT *Interestingly, most Eurosceptic extreme Hungarian parties are Slovak, which is anti-Slovak and anti-Hungarian;*

In Sentence 7-85, there were two words with the prefix anti-: antieslovaco (anti-

Slovak) and antihúngaro (anti-Hungarian). Both of them were translated in GT as an adverb 반대로 (in reverse). A desirable equivalence would be 반슬로바키아 and 반헝가리.

Sentence 7-86

ST Los demócratas, por su lado, compiten por demostrar quién es más «anti Trump».

GT 반면에 민주당은 누가 더 "트럼프 반대"인지를 증명하기 위해 경쟁합니다.

BT *Democrats, on the other hand, compete to prove who is more "anti-trump".*

In Sentence 7-86, the same prefix anti- was written as an independent noun, and it was translated in GT in a noun: 반대 (opposition). A correct translation should be 반트럼프 in the same way as Sentence 7-85. BT somehow generated the word in the correct way.

Sentence 7-87

ST A cambio, Kiev debe conceder a los dos enclaves un sistema de autogobierno.

GT 그 대가로 키에프는 두 영토를 자치 체제로 만들어야합니다.

BT *In return, Kiev must make both territories autonomous.*

In the case of 'auto-', the options were diverse. In Sentence 7-87, the word autogobierno (self-government) was translated into a noun 자치 (autonomy), omitting the *government* part. A suitable translation could be 자치 정부 (autonomous government).

Sentence 7-88

ST En la vecina también autoproclamada República Popular de Lugansk (LNR) se celebran igualmente elecciones y el que parte con ventaja es el que desempeña interinamente el cargo de máximo dirigente, Leonid Pasechnik.

GT 이웃 국가들에서도 스스로 자칭 된 LUMR (Lugansk Republic of Lugansk) 선거도 개최되며 유리한 방향으로 이끌어가는 당은 일시적으로 최고 지도자 인 Leonid Pasechnik의 입장을 취하고있다.

BT *In neighboring countries, the LUMR (Lugansk Republic of Lugansk) election, which is self-proclaimed, is held, and the party that leads in favorable direction temporarily takes the position of top leader Leonid Pasechnik.*

In Sentence 7-88, the same prefix was attached to proclamada (proclaimed) to form an adjective. Its translation in GT was redundant: 스스로 자칭된 (self-styled for itself)'. It would be better if the adverb was omitted so that it left 자칭 (so-styled).

Sentence 7-89

ST	El pasado agosto, fue asesinado en un atentado todavía sin esclarecer el que había sido jefe de la autoproclamada República Popular de Donetsk (DNR) durante casi cuatro años, Alexánder Zajárchenko.
GT	지난 8 월, 알렉산더 자 자르 첸코 (Alexander Zajárchenko)는 자신이 주장한 도네 치크 인민 공화국 (Donetsk People 's Republic, DNR)의 머리로 4 년 가까이 머물렀다는 점을 아직도 분명히 모르는 공격에서 암살 당했다.
RT	*In August, Alexander Zajárchenko was assassinated in an attack that he still did not know clearly that he stayed for four years as head of the Donetsk People 's Republic (DNR).*

The same vocabulary autoproclamada (self-proclaimed) was translated in Sentence 7-89 as 자신이 주장한 (proclaimed by himself). The fact that its translation was entirely different in the two similar contexts was quite surprising.

7.2.2.8. Word Order

The biggest challenge in MT, in general, for a distant pair such as Spanish and Korean has been the long-ranged word order difference. For instance, Daems et al. (2014) claimed that their error analysis for an English-to-Dutch translation revealed that word order was the biggest reason for the errors. Surprisingly enough, the quantitative result of the error analysis demonstrated that the syntactic error constituted only 5.76% without penalty and 4.49% with a penalty of the total sum. This chapter was dedicated to explaining again the calculation of a shift (movement) and to show some of the most relevant examples.

As previously stated, the calculation was based on the HTER metrics. A block of words that moved to a right place was considered as one shift or one edit. While the metrics normalized the score by the number of words in the reference sentence, the normalized score was not employed in this study because the main concern was to monitor the number of errors, not their

relative importance. Thus, it was safe to say that the HTER metrics were partially applied.

One possible challenge was that such calculation could vary per person as the calculation was carried out manually. We strived to calculate the minimum number of shifts. A maximum of five shifts were observed in one sentence. As an instance, in Sentence 7-90 given below, the source sentence had a relative clause led by donde (where). GT separated the sentence into two without an explicit ending of the previous sentence. While there could be many ways to deal with the sentence, such as connecting the two sentences into one, we acknowledged the decision that the engine made and tried to preserve it as much as we could. Thus, considering that the two sentences could be connected with some additional words shown in the targeted reference (TR), no shift was calculated for this segment.

Sentence 7-90

ST «Mientras que los subsidios europeos debían reforzar el sistema político de la UE y abordar el déficit democrático, los euroescépticos han utilizado principalmente sus recursos para apoyar a sus partidos nacionales», escribe Wouter Wolfs, investigador del Instituto de Gobernanza Pública de KU Leuven, para el Foro contra la corrupción y la integridad de la OCDE, donde resalta la evolución de las formaciones euroescépticas desde hace veinte años, pasando del total boicot a una participación muy intensa en las instituciones y elecciones europeas.

GT 쿠루 벤 (KU Leuven) 공공 거버넌스 연구소의 연구원 인 우스터 울프 스 (Wouter Wolfs)는 "유럽의 보조금은 EU 정치 체제를 강화하고 민주적 인 적자를 다루어야하지만 Eurosceptics는 자국의 정부를 지원하기 위해 주로 자원을 사용했다” OECD의 부패와 청렴성 반대 포럼 - 전체 보이콧에서부터 유럽 제도와 선거에 대한 매우 치열한 참여에 이르기까지 20 년 동안 유로 지성 형성의 진화를 강조합니다.

BT *Wouter Wolfs, a researcher at the KU Leuven Institute for Public Governance, said: "While European subsidies must strengthen the EU political system and deal with democratic deficits, Eurosceptics is mainly using resources to support its governments. "The OECD's Anti-Corruption and Integrity Forum emphasizes the evolution of euro-state formation for 20 years, from the entire boycott to the very intense participation in European institutions and elections.*

TR 쿠루 벤 (KU Leuven) 공공 거버넌스 연구소의 연구원 인 우스터 울프 스 (Wouter Wolfs)는 "유럽의 보조금은 EU 정치 체제를 강화하고 민주적 인 적자를 다루어야하지만 Eurosceptics는 자국의 정부를 지원하기 위해 주로 자원을 사용했다” [라고] OECD의 부패 반대와 청렴성 포럼[에서 밝혔다. 이 포럼은] 전체 보이콧에서부터 유럽 제도와 선거에 대한 매우 치열한 참여에 이르기까지 20 년 동안 유로 지성 형성의 진화를 강조합니다.

Following such rule, up to five shifts were observed in one sentence. Out of 61 sentences with word order problems, 47.5% had one shift edit and 27.8% had two shift edits, which showed that 75.3% of the total syntactic errors were minor with 1 - 2 edits irrespective of the size of the block. The most shifts that could happen in a sentence in this dataset was five, and there were three sentences of such kind. Some of the most relevant examples were introduced subsequently. Note that to enhance the readers' comprehension, a targeted reference (TR) was provided. It would help to explain how the shift could be made. Moreover, the GT was corrected if necessary in TR to justify the shifts. The shift occurred either on a word level or on a sequential level.

Sentence 7-91

ST Nelson, su rival demócrata, le acusa de utilizar su potestad como gobernador para sugerir que puede pedir una investigación policial.

GT 그의 민주당 경쟁자 인 넬슨 (Nelson)은 주지사로 그의 권력을 사용하여 그가 경찰 수사를 요청할 수 있다고 제안했다고 비난했다.

BT *His Democratic rival, Nelson, accused the governor of using his power to suggest that he could ask for a police investigation.*

TR 그의 경쟁자 인 민주당[의] 넬슨 (Nelson)은 주지사로 그의 권력을 사용하여 그가 경찰 수사를 요청할 수 있다고 제안했다고 비난했다.

Sentence 7-91 showed a simple shift on a word level. While the structure of GT was partially correct, it sounded vague if Nelson was his rival inside the Democratic Party or if he was a Republican while Nelson was a Democrat. Part of such fault was due to the structure of the source text. Considering the context, a correct way was to exchange 민주당 (Democratic party) and 경쟁자 (rival).

Sentence 7-92

ST Hermano del también dimisionario Boris y político de gran prestigio y popularidad, Jo Johnson, que votó contra el Brexit, era el encargado de pactar el estratégico acuerdo sobre Transportes, un sector clave para las futuras relaciones del Reino Unido con la UE.

GT 부시 대통령과 블레치 총리에게 투표 한 조 존슨 (Jo Johnson) 대통령은 보리스 (Boris)를 사임하고 영국과 EU의 미래 관계를위한 핵심 분야 인 교통에 관한 전략적 협상을 담당했다.

BT *President Jo Johnson, who voted for President Bush and Prime Minister Blech, resigned from Boris and was in charge of strategic negotiations on transportation, a key area for future UK and EU relations.*

TR 부시 대통령과 블레치 총리에게 투표 한 조 존슨 (Jo Johnson) 대통령은 사임[한] 보리스 (Boris)[의 형제로] 영국과 EU의 미래 관계를위한 핵심 분야 인 교통에 관한 전략적 협상을 담당했다.

The shift in Sentence 7-92 was as simple as that in Sentence 7-91. The difference was that in Sentence 7-92, many source words were omitted. Such a condition made the shift rather simple. Supposing that the omitted words would be added as in TR, the words 보리스 (Boris) and 사임하고 (resigned) were switched.

Sentence 7-93

ST La cita con las urnas es un juicio a su presidencia -formidable para unos, infame para otros, polarizante sin duda-, un referéndum al «trumpismo».

GT 투표소와의 약속은 그의 대통령 대행의 재판이다. 다른 사람들에게는 악명이 높으며 의심의 여지없이 "트럼프즘"에 대한 국민 투표가 실시된다.

BT *The appointment with the polling place is the trial of his acting president. It is notorious for others, and there is no doubt that a referendum on "trumpfism" takes place.*

TR 투표소와의 약속은 다른 사람들에게는 악명이 높으며 의심의 여지없이 그의 대통령 대행의 재판이다. "트럼프즘"에 대한 국민 투표가 실시된다.

The one shift could happen with a block as long as that of Sentence 7-93. Even though there was one shift, a block was composed of 4 and 6 words.

Sentence 7-94

ST Pese a las críticas y advertencias de la OSCE y la mayor parte de los países de Occidente, las autoridades rebeldes de las repúblicas separatistas del este de Ucrania, Donetsk y Lugansk, se disponen hoy a celebrar «comicios» para elegir a sus respectivos líderes y a los diputados de los llamados «Consejos Populares».

GT OSCE와 대부분의 서방 국가에 대한 비판과 경고에도 불구하고 우크라이나 동부 분리주의 반군 당국 인 도네츠크와 루간 스크는 오늘 각자의 지도자를 선출하기 위해 "선거"를 개최 할 준비를하고있다. 소위 "인민위원회"의 대리인.

BT *Despite criticism and warnings against the OSCE and most western nations, Donetsk and Lugansk, eastern Ukrainian separatist rebels, are preparing to hold "elections" to elect their leaders today. Agent of the so-called "People's Committee".*

TR OSCE와 대부분의 서방 국가에 대한 비판과 경고에도 불구하고 우크라이나 동부 분리주의 반군 당국 인 도네츠크와 루간 스크는 오늘 소위 "인민위원회"의 각자의 지도자[와] 대리인를 선출하기 위해 "선거"를 개최 할 준비를하고있다.

Sentence 7-94 was an example of two shifts. The two words —오늘 (today) and 소위 “인민위원회”의 (of the so-called “People’s Committee”)— were positioned in an incorrect way. Each word moved to the right position and that produced two shifts.

Sentence 7-95

ST En la vecina también autoproclamada República Popular de Lugansk (LNR) se celebran igualmente elecciones y el que parte con ventaja es el que desempeña interinamente el cargo de máximo dirigente, Leonid Pasechnik.

GT 이웃 국가들에서도 스스로 자칭 된 LUMR (Lugansk Republic of Lugansk) 선거도 개최되며 유리한 방향으로 이끌어가는 당은 일시적으로 최고 지도자 인 Leonid Pasechnik의 입장을 취하고있다.

BT *In neighboring countries, the LUMR (Lugansk Republic of Lugansk) election, which is self-proclaimed, is held, and the party that leads in favorable direction temporarily takes the position of top leader Leonid Pasechnik.*

TR 스스로 자칭 된 LUMR (Lugansk Republic of Lugansk) 이웃 국가들에서도 선거도 개최되며 유리한 방향으로 이끌어가는 당은 일시적으로 최고 지도자 인 입장을 취하고 있[는] Leonid Pasechnik의

Similarly, Sentence 7-95 also showed two shifts in one sentence. The two blocks were shifted while the structure of GT was maintained as much as possible. Most of the two shifts occurred with a long block of words.

Sentence 7-96

ST 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.

GT 처음으로 전통적인 중도 좌파와 중도 좌파의 유럽 연합 블록은 50 % 이하로 떨어질 것으로 예상되며 여론 조사원들은 유럽과 유럽의 가치관을 어지럽히도록 노력하는 국민 강대국당이 3 분의 1의 좌석을 차지할 것으로 추정하고있다. 역설적이게도 그들은 유니언 기금을 이용하여 재정을 강화했습니다.

BT *For the first time, the traditional middle-left and center-left European Union bloc is expected to fall below 50%, and polls estimate that the National Power Party, which is trying to clutter European and European values, will occupy one-third of the seats . Paradoxically, they used the union funds to finance.*

TR 처음으로 중도 좌파와 중도 좌파의 전통적인 유럽 연합 블록은 50 % 이하로 떨어질 것으로 예상되며 여론 조사원들은 국민 강대국당이 3 분의 1의 좌석을 차지할 것으로 추정하고있다. 그들은 유럽과 유럽의 가치관을 어지럽히도록 노력하[고] 역설적이게도 유니언 기금을 이용하여 재정을 강화했습니다.

There were also cases such as Sentence 7-96 where one word formed a block in two shifts. The word 전통적인 (traditional) needed to be situated before the European Union bloc in GT. The second block was much longer than the previous one, which moved to the next sentence in GT. The existence of long blocks demonstrated that the NMT engine did not produce a mere word salad, but phrases with meaning despite some incorrect positions.

Sentence 7-97

ST La presidenta de la Junta de Andalucía, Susana Díaz, ganaría las próximas elecciones autonómicas con más de 10 puntos de ventaja, pero lejos de la mayoría absoluta (55 escaños de un total de 109), según la encuesta hecha pública este miércoles por el Centro de Investigaciones Sociológicas (CIS).

GT 유다 데 안달루시아 (Susana Díaz) 의장은 다음 주 지방 선거에서 10 점 이상의 우위를 점할 것으로 예상했지만, 이번 수요일에 센터에 의해 공개 된 설문 조사에 따르면, 절대 다수 (109 개 가운데 55 석) 사회학 연구 (CIS).

BT *Susana Díaz, chairman of the Supreme Council for Democracy, predicted that she would have an advantage of more than ten in the next provincial election, but according to a survey released by the Center this Wednesday, the absolute majority (55 out of 109)).*

TR 이번 수요일에 센터에 의해 공개 된 사회학 연구 (CIS)[의] 설문 조사에 따르면, 안달루시아 유다 데 (Susana Díaz) 의장은 다음 주 지방 선거에서 10 점 이상의 우위를 점할 것으로 예상했지만, 절대 다수 (109 개 가운데 55 석)

In Sentence 7-97, three shifts were performed in a sentence. The main reordering was because of the longest block —*according to a survey […].* The most natural position of the block was either in front of the sentence or at the end of the sentence, but the current position did not impede human comprehension as the block was separated by commas. The calculation of the shift, however, stuck to the rule and aimed at the most natural option.

Sentence 7-98

ST Sobre todo tras un cierre de campaña virulento, agitado por la campaña del miedo desatada por Trump a raíz de la caravana de inmigrantes y por los episodios de violencia -los paquetes bomba a sus enemigos politicos y, sobre todo, la matanza en la sinagoga de Pittsburgh- que parte de la opinión pública relaciona con el mensaje agresivo del presidente.

GT 특히 악의적 인 선거 운동 종료 후, 이민자들의 캐러밴과 폭력 사건의 결과로 트럼프가 쏟아내는 두려움에 대한 캠페인에 동요 됨. - 정치적 원수 폭파, 특히 무엇보다도 시나고그의 학살 피츠버그 - 여론의 일부는 대통령의 공격적인 메시지와 관련이 있습니다.

BT *Especially after the end of the malicious election campaign, the campaign against the fears of Trump as a result of immigrants' caravans and violent incidents was shaken. - Political bombing, especially above all the slaughter of the Sinaigg Pittsburgh - Some of the public opinion is related to the president's aggressive message.*

TR 캐러밴 이민자들과 여론의 일부는 대통령의 공격적인 메시지와 [연관을 짓는] 폭력 사건의 결과로 (정치적 원수 폭파, 특히 무엇보다도 피츠버그 시나고그의 학살 [과 같은]) 트럼프가 쏟아내는 두려움에 대한 캠페인에 동요 됨. 특히 악의적 인 선거 운동 종료 후,

Sentence 7-98 was one of the most destructred sentences in the dataset with having five shifts. Some of them were short blocks while others were long blocks. This sentence obtained 2.17 fluency score (54.25% fluency) and 2.83 adequacy score (70.75% adequacy score) by the evaluators. The rest two sentences that had five shifts in one sentence equally obtained 1.67 fluency score (41.75% fluency) and 2.17 adequacy score (54.25% adequacy).

7.2.2.9. Style

The definition of the stylistic error was a content that was grammatical, but erroneous from the viewpoint of the design. In this study, the scope that this category covered was broader than usual by including not only awkward cases but also wrong collocations, terminology, locale convention and verity, most of which were an independent error class in TAUS DQF. The main reason to combine such categories was due to their minor contribution to the result. Three most statistically important patterns were discussed, subsequently: awkward style, collocations error, and technical error.

Awkward Style The awkward expressions referred to a literal translation

that sounded unnatural to the target language. In Sentence 7-99, the translation of Es una desgracia (It is disgraceful) was grammatical and understandable, but did not sound natural in Korean. Many alternate options existed, such as '불명예스러운 일이다' or '명예롭지 못하다' that sounded more natural to native Korean speakers.

Sentence 7-99

ST "Es una desgracia", dijo Trump, quien en las presidenciales de 2016 atribuyó a irregularidades falsas su derrota, en el número total de votos, frente a la demócrata Hillary Clinton.

GT "그것은 불명예 스럽다." 2016 년 대통령 선거에서 민주당 힐러리 클린턴에 반대하여 총 투표 수에서 거짓 부정에 대한 자신의 패배를 돌린 트럼프는 말했다.

BT *"It is disgraceful," Trump said in his 2016 presidential election, against his Democratic challenger Hillary Clinton, who turned his defeat against false injustice in total votes.*

Among various awkward expressions, some interesting differences were observed. Firstly, an inanimate subject was accompanied by an active verb, which was unnatural in Korean.

Sentence 7-100

ST Las encuestas dicen que los demócratas recuperarán la Cámara de Representantes, y la duda es cuál será la ventaja final.

GT 여론 조사는 민주당 원들이 하원 의원을 되찾을 것이고, 의문점은 최종 우위가 될 것이라고 말했습니다.

BT *Polls say Democrats will regain congressional members, and questioning will be the ultimate advantage.*

For instance, in Sentence 7-100, a survey could not *tell* in the Korean language. Other inactive verbs should be used or the subject should be changed to a different POS such as *according to a survey*.

Sentence 7-101

ST [SUB-HEADLINE] La encuesta sitúa a la coalición de Podemos e IU en segundo lugar y anticipa un empate técnico entre PP y Ciudadanos

GT 이 설문 조사는 Podemos와 IU의 연합을 두 번째 장소에두고 PP와 Citizens 사이의 기술적 인 관계를 예상합니다

BT *This survey puts the coalition of Podemos and IU in the second place and expects a technical relationship between PP and Citizens*

Similarly, in Sentence 7-101, a survey could not put (sitúa) or expect (anticipa) in the Korean language because it was an inanimate noun. A more suitable expression would be to change the inanimate subject to a preposition: 설문 조사에 따르면 (according to a survey).

Sentence 7-102

ST Finalmente, el secretario del Estado (autoridad electoral) de Florida hizo oficial a mediodía del sábado que la diferencia de votos obliga a un recuento automático.

GT 마지막으로, 플로리다 주 (장관) 선거 당국은 토요일에 정오에 투표 결과의 차이가 자동 재계 표를 요구했다고 공식 발표했다.

BT *Finally, the Florida state (state) electoral authorities officially announced at noon Saturday that the difference in voting results required an automatic recount.*

In Sentence 7-102, a problem occurred when GT said *the difference […] required a recount*. A better way to express ST would be to convert the subject into a preposition as the previous examples: *because of the difference, […] was required*.

Secondly, the agreement of the subject and verb was erroneous especially when the prepositional phrase of 'according to' existed in a sentence.

Sentence 7-103

ST La presidenta de la Junta de Andalucía, Susana Díaz, ganaría las próximas elecciones autonómicas con más de 10 puntos de ventaja, pero lejos de la mayoría absoluta (55 escaños de un total de 109), según la encuesta hecha pública este miércoles por el Centro de Investigaciones Sociológicas (CIS).

GT 유다 데 안달루시아 (Susana Díaz) 의장은 다음 주 지방 선거에서 10 점 이상의 우위를 점할 것으로 예상했지만, 이번 수요일에 센터에 의해 공개 된 설문 조사에 따르면, 절대 다수 (109 개 가운데 55 석) 사회학 연구 (CIS).

BT *Susana Díaz, chairman of the Supreme Council for Democracy, predicted that she would have an advantage of more than ten in the next provincial election, but according to a survey released by the Center this Wednesday, the absolute majority (55 out of 109)).*

In Sentence 7-103, GT read as *[…] expects to win, according to a survey.* Firstly, the subject of the verb *expect* was omitted. As discussed previously that

the inanimate subject could not *expect*, a desirable alternative would be to change the verb into a passive form: 예상되다 (is expected). As such, when *according to* was used, it was easy to be vague about the subject and the verb.

Sentence 7-104

ST Según su opinión, «los resultados de esta convocatoria jamás serán reconocidos por la Comunidad Internacional (...) y provocarán la adopción de nuevas sanciones contra Rusia».

GT 그의 견해로는 "이 호의 결과는 국제 사회 (...)에 의해 인정되지 않으며 러시아에 대한 새로운 제재의 채택으로 이어질 것"이라고 말했다.

BT *In his view, "the outcome of this issue is not recognized by the international community (...) and will lead to the adoption of new sanctions against Russia," he said.*

Similarly in Sentence 7-104, while there was a prepositional phrase *according to his opinion*, the main clause did not have a subject but a verb 말했다 (told). A correct way, in this case, would be to erase the verb *told.* Then, the sentence would look like: 그의 견해로는 […] 이어질 것이다.

Thirdly, a literal translation could also give an unnatural feeling when the literally translated word did not fit in to the context because it was a dictionary definition. Three verbs of such kind were detected in the dataset: justificar (to justify), judicializar (to judicialize) and digitalizar (to digitize). A desirable way of dealing with these words would be to explain the verb, but it was hard to conclude that it should be. While suggesting a better option, we did not state categorically that those literal translations were erroneous.

Sentence 7-105

ST En términos de ingresos fijos, cada eurodiputado recibe un salario de alrededor de 6.600 euros netos al mes, 4.400 para gastos de oficina sin justificar y alrededor de 24.000 para asesores.

GT 고정 수입 측면에서, 각 MEP는 월 6,600 유로의 순이익을, 정당화되지 않은 사무 경비는 4,400, 고문은 약 24,000의 급여를받습니다.

BT *In terms of fixed income, each MEP receives a net profit of 6,600 euros per month, unwarranted office expenses of 4,400 and torture of approximately 24,000 salaries.*

In Sentence 7-105, The word justificar was translated in GT as 정당화되다 (to be justified), which existed in a Korean dictionary but made the context sound

unnatural. A better expression that took the context into account would be 증빙이 필요 없는 (that does not require documentation).

Sentence 7-106

ST Paralelamente, la elección ha sido judicializada a través de demandas de los dos aspirantes a senador.

GT 병행하여, 선거는 두 명의 야심 찬 상원 의원의 요구를 통해 사법 화되었습니다.

BT *In parallel, the election was made justified by the demands of two ambitious Senators.*

In Sentence 7-106, the word judicializar was translated in GT as 사법화되다 (to be judicialized). The definition of *judicialize* was "to convert or integrate into a judicial system; to bring under the remit of the law."[64] Such word was rarely used even in English. The Korean translation still conveyed the meaning of ST, but it could be improved by adapting the word to the context (and to the newswire domain) such as 법적 효력을 갖다 (to have a legal effect).

Sentence 7-107

ST En unas elecciones de 2016, eliminó un archivo de papeletas físicas pero mantuvo uno digitalizado.

GT 2016 년 선거에서 그는 실제 투표 용지 파일을 삭제했지만 하나는 디지털화했습니다.

BT *In the 2016 election, he deleted the actual ballot file, but one digitized it.*

In Sentence 7-107, the word digitalizar was translated to 디지털화하다 (to digitize), which meant "to convert (pictures or sound) into a digital form that can be processed by a computer."[65] A better expression that could enhance the fluency of GT would be to explain the contents: 컴퓨터 상에 (파일로) 남겨두다 (to leave it as a file on a computer).

Collocations Error The Merriam-Webster dictionary defined a collocation as "a noticeable arrangement or conjoining of linguistic elements

[64] https://en.wiktionary.org/wiki/judicialize

[65] https://www.merriam-webster.com/dictionary/digitize

(such as words)".[66] The collocations were generally a fixed expression that could not be translated otherwise. Errors occurred when the collocational expressions were not recognized as a whole and translated independently. While introducing the corresponding examples, this section also included a few terminologies that appeared in the dataset. The terminology in this context represented a political term that could not be literally translated. However, considering the domain (newswire), only a few examples were observed.

Sentence 7-108

ST Su renuncia echa gasolina al incendio declarado en algunas aerolíneas que, alarmadas por el desastre, tratan como apestados a sus accionistas británicos y les amenazan con eliminar sus derechos de voto para demostrar a Bruselas una mayoría de propiedad comunitaria y operar así sin problemas en toda la UE.

GT 그들의 사임은 재난으로 경계를당한 일부 항공사들에 의해 선언 된 화재에 휘발유를 붓고 그들의 영국 주주들을 괴롭 히고 브뤼셀에 대다수의 공동체 재산을 보여주고 투표권을 없애겠다고 위협했다. EU

BT *Their resignation threatened to pour gasoline into fires declared by some of the airlines caught up in disaster, plague their British shareholders, show the majority of community property in Brussels and eliminate voting rights. EU*

In Sentence 7-108, the phrase Echa gasolina al incendio was an idiomatic expression that, when translated in a literal way, could not transfer the original meaning. While GT output a literal translation that it *poured gasoline into fire* (화재에 휘발유를 붓다) as seen in BT, a desirable translation would be 불 난 데 부채질하다 (Add fuel to the fire/flames).

Sentence 7-109

ST Quienes mantienen un empate técnico son PP y Ciudadanos, los primeros con un 18,66% de voto y el segundo con un 18,55%, que se traduce en idéntico número de representantes en el Parlamento andaluz: una horquilla que va de los 20 a los 22.

GT 기술적 인 관계를 유지하는 사람들은 PP와 시민이며, 18.66 %의 투표권을 가진 첫 번째 사람과 18.55 %의 사람인 두 번째 사람은 안달루시아 의회 의원과 같은 수의 사람으로 번역됩니다. 머리말은 20 22시에

BT *The people who maintain the technical relationship are the PP and the citizen, the first person with 18.66% of the vote and the second person with 18.55% will be translated into the same number of people as the Andalusian parliamentarians. The preamble is at 20*

[66] https://www.merriam-webster.com/dictionary/collocation

In Sentence 7-109, a term in boxing was used in the context of the election. The term empate técnico meant “when a fight has to be stopped because a fighter is unable to continue from an accidental injury (usually cuts) or foul”.[67] The English equivalence was *a technical draw* and the Korean translation was 테크니컬 무승부. However, GT translated it as 기술적인 관계 (technical relationship) as BT showed. Considering the political domain, a desirable translation of the collocation would be 동률 (tie).

Sentence 7-110

ST La última pésima noticia ha sido la dimisión, el viernes pasado, del ministro de Transportes, Jo Johnson, autor de esa diatriba entre vasallaje y caos, es decir, entre el sometimiento a las reglas del mercado único sin voz ni voto de Londres (la oferta de May rechazada por todos) o la desastrosa salida sin acuerdo el 29 de marzo.

GT 최악의 소식은 지난 금요일, 교통 장관, 조 존슨 (Jo Johnson)의 사퇴, 즉 매매와 혼돈 사이의 비난 즉, 목소리 나 런던의 투표가없는 단일 시장의 규칙에 대한 제출 사이다. 3 월 29 일에 합의없이 모든 사람이 거부 할 수 있음) 또는 비참한 출발.

BT *The worst news was last Friday's minister of transportation, Jo Johnson's resignation, the blaming between trade and chaos, the submission of a single market rule without a voice or London vote. Everyone can refuse without consent on March 29) or a miserable departure.*

In Sentence 7-110, a political term was used in the sentence: salida sin acuerdo (exit without agreement). It was a fixed expression that should not be translated to *exit without agreement* but to *no-deal Brexit*. GT translated it as 합의없이 […] 출발 (without consent […] departure). The correct expression was 노딜 브렉시트 (a phonetic transcription of the English term).

Sentence 7-111

ST desde la profunda división del país en dos bloques al bandazo del Brexit duro al blando y la inestabilidad política, pasando por la voladura de las líneas rojas exhibidas por Londres frente a Bruselas o la incertidumbre, letal para la economía.

GT 브뤼셀 앞에서 런던이 전시 한 빨간 선이 폭발하거나 경제에 치명적인 불확실성으로 인해 브리짓이 강경 한 브리짓 (Br Brexit)이 부드럽고 정치적으로 불안정한 상태에 이르기까지 두 블록에 걸친 국가의 깊은 분파에서 일어났습니다.

[67] https://en.wikipedia.org/wiki/Technical_draw

BT *The explosion of London's red line in front of Brussels, or uncertainty over the economy, has resulted in a two-block deep branch of the country, from Brigitte's tough Br Brexit to a soft, politically unstable state.*

In Sentence 7-111, there were three political terms in the sentence, all of which were not considered properly. In the first place, Brexit duro and Brexit blando were equivalent to 하드 브렉시트 (hard Brexit) and 소프트 브렉시트 (soft Brexit). The Korean expressions were a phonetic transcription of the English terms. GT translated them to 강경한 브리짓 (tough Bridget) and 부드럽고 (soft), which led to a big confusion. Secondly, the phrase las líneas rojas was a fixed expression that meant "a figurative point of no return or line in the sand, or a limit past which safety can no longer be guaranteed"[68]. While the Korean equivalence was 레드 라인 (red line), GT translated the meaning of the words (빨간 선) and caused a misunderstanding.

Technical Error While Collocation Error was a style error from a linguistic aspect, Technical Error approached from a viewpoint of a translation technique. There was one noticeable error observed in this regard. That was, in news, a name of the city used to refer to their government, in such a way that Seoul represented the Korean government. Such a technique, on the other hand, was not preferred in Korean. A general way of translating those items was to use the name of the country (Seoul > Korea) or to add *government* after the name of the city (Seoul > Seoul government). In all cases, the system translated them as the name of the city. Such a conversion would be hard for a machine.

Sentence 7-112

ST Pero en Moscú piensan de otra manera.

GT 모스크바에서 그들은 다르게 생각합니다.

BT In Moscow they think differently.

In Sentence 7-112, GT had 모스크바, which should be revised to 러시아 (Russia), 모스크바 정부 (Moscow government) or 러시아 정부 (Russia government).

[68] https://en.wikipedia.org/wiki/Red_line_(phrase)

Sentence 7-113

ST [SUB-HEADLINE] La Unión Europea, Estados Unidos y Kiev califican los comicios de «ficticios e ilegítimos»

GT 유럽 연합 (EU)과 미국, 키예프는 선거를 "허구와 불법"이라고 부른다.

BT *The European Union, the United States and Kiev call the elections "fictitious and illegal.*

Similarly, Sentence 7-113 had Kiev to represent the government of Ukraine. While GT translated it as *Kiev*, a better translation would be to use 키예프 정부 (Kyiv government), 우크라이나 정부 (Ukraine government) or 우크라이나 (Ukraine).

7.3. Error Analysis II: On Post-Edited System Translations

With the result of error analysis on the system translations at hand, this section attempts to carry out error analysis on the post-edited version of the system translations by the six post-editors. The core work of the second error analysis is to deduce what the post-editors perceived as an error from their edited data. The post-edited results from the experiment are extracted and manually analyzed according to the same error classification that was used in Analysis I. All six post-edited texts are applied, but as the experiment is designed to perform post-editing on half of the dataset, 129 sentences of each editor are assessed. Such work underpins the result of the previous error analysis done by one annotator. Furthermore, it gives insights over the relative importance of the error classes to human understanding of the machine language.

The six post-edited versions of the system translations were analyzed by the same annotator to the previous error analysis, and the result was given in Table 7.8. Each version had 129 sentences, and a total of 834 sentences were analyzed. To compare the result, the number of errors calculated in the same sentences in Analysis I was given, as well.

Analysis I	Analysis II						
	EV1	EV2	EV3	EV4	EV5	EV6	***Avg***
636	754	747	922	362	557	427	***628.17***

Table 7.8: Number of errors in the post-edited translations.

Table 7.7 showed that the annotator detected 636 errors from the system translation while the post-editors' result varied substantially, from minimum 362 to maximum 922 and on average 628. Such difference showed that post-editing was different from a mere revision. That is, post-editing was more concerned of conveying the idea of the source text so that more or fewer edits than a revision could be made depending on the situation. In other words, as far as the sentences were of acceptable quality, some remaining errors could be left unhandled as in EV4, EV5, or EV6 or more errors could be witnessed as in EV1, EV2, and EV3. Such divergence could show a minimum level of human understandability of a system translation. It could be also interpreted as the gap between a human language and a machine language.

The error analysis I reported on 636 errors in those 129 sentences. While those errors included all 10 error classes in Table 7.3, the errors that were observed in the error analysis II only covered word order, mistranslation, untranslated, addition, omission, and punctuation. Spacing, grammar, style, and spacing were not found or their size was statistically minor. For example, the post-editors were told to ignore spacing errors in the guideline. In that sense, the errors of Error Analysis I needed to be adjusted to feature the same categories. Figure 7.4 showed the proportions of the error classes in Error Analysis I when the six error classes were considered only. The proportions of a full and half dataset were compared. In the 129 sentences, almost all categories

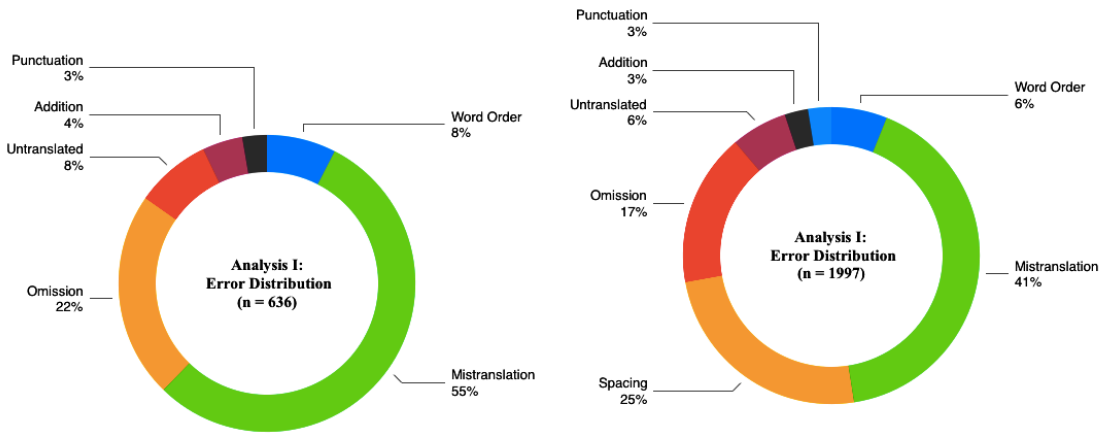


Figure 7.4: Distribution of error classes in Error Analysis I: a comparison of half dataset and full dataset.

were more highly-proportioned than in the full dataset. For instance, mistranslation was 14% higher and word order error was 2% higher in *n*=636.

Having obtained the adjusted distribution of error classes of the first analysis, that of the post-edited versions of the system translations were graphically shown in Figure 7.5 altogether. In Figure 7.5, the first column showed the percentile of Analysis I, which was contrasted to the six post-edited versions. The interesting point was that irrespective of the number of errors they detected, the proportions were almost alike. The largest pie was taken by mistranslation, ranging from 55% (in Analysis I, EV1 and EV4) to 65% (in EV6). The second-largest proportion was occupied by omission with 19 - 28%. In relation to word order, the proportion ranged from 6% (in EV1 and EV6) to maximum 10% (in EV2).

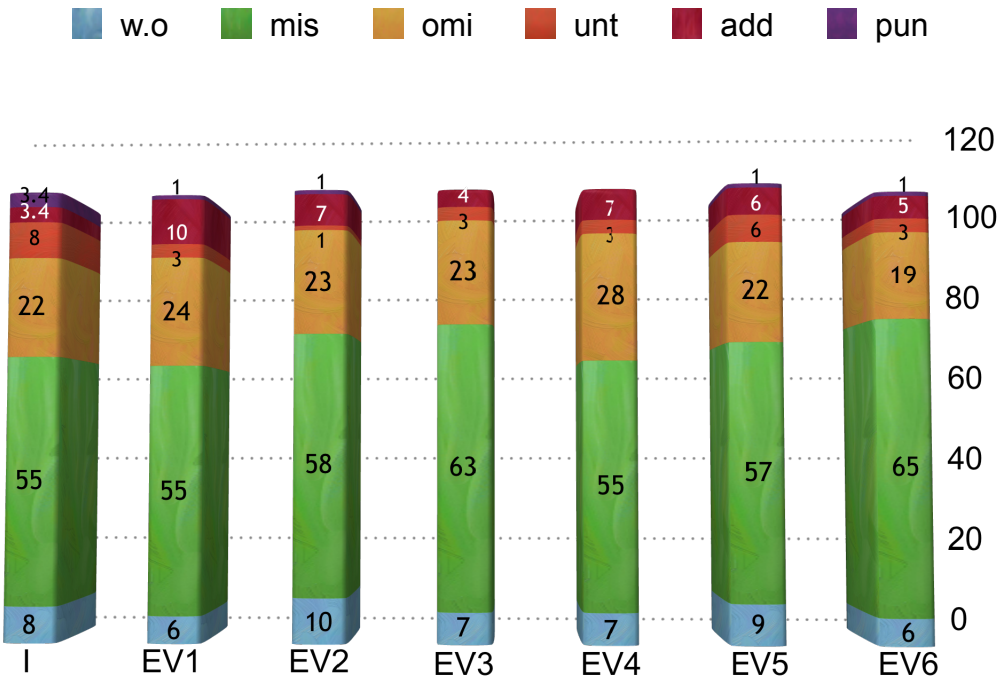


Figure 7.5: Error distribution of Error Analysis I and II.

From such comparative study, the first and foremost finding was that the homogeneous error distribution proved the validity of the result in the error analysis I. Secondly, the post-editors all agreed that there were a few word order errors in the dataset. It, in return, demonstrated that the Google NMT engine was robust in dealing with syntactic issues in the *es-ko* pair. If the ten

error classes were divided into a syntactic error and a lexical error, all but the word order belonged to the lexical error. It meant that 90% - 94% of the errors occurred in GNMT in the *es-ko* pair were a lexical issue. It could be interpreted as the NMT engine broke down the boundaries of close and distant language pairs.

Furthermore, the majority of the lexical errors was turned out to be due to mistranslation, which was interpreted in this thesis as incorrect word choice. Considering that the post-editing productivity in Chapter 6 was not statistically remarkable, it would be an interesting line of research to monitor how the errors from wrong lexical choice influenced the post-editing productivity.

7.4. Chapter Summary

In Chapter 7, we reported on many valuable findings in terms of the errors occurred in the NMT engine (Google Translate) in the *es-ko* pair. A binary error analysis was carried out. The main error analysis on raw system translations produced by GNMT detected 1,977 errors in the dataset. From ten error classes —Word Order, Mistranslation, Omission, Addition, Untranslated, Punctuation, Spacing, Spelling, Grammar, and Style—, the largest proportion was taken by Mistranslation. Meanwhile, the word order occupied 6% of the total errors, which was an unexpected result when one of the major issues in distant pairs was known to be syntax.

The qualitative analysis of the data was described subsequently by error type. The most common patterns in Addition was that the same words were duplicated. In Omission, when the omitted words were manually tagged by Brown tag sets, 10 POS tags were statistically dominant. Among them, the most frequent omission occurred in nouns, with 23.9%. The most common POS of the nouns was a common noun with 71.3%. In Untranslated, words of the source text that also remained in the target text were considered untranslated. Following such rule, 39% of the untranslated items were erroneous while 61% were a correct translation. In Word Order, 75.3% of the sentences in this regard had one or two shifts, which was a minor reordering. The examples of shifts demonstrated that most of the shifts moved as a block. This could mean that

the outputs of the NMT engine were not a *word salad* but a meaningful combination of phrases. One interesting behavior of the NMT engine was that a back translation of a system translation (Korean) to English sometimes corrected the corresponding errors even though they were not traceable from the Korean sentence.

Additionally, the second error analysis was performed on the post-edited system translations of GNMT. There were a total of 834 sentences (129 sentences per post-editor). The error analysis revealed that half of them detected more errors than the error analysis I —up to 1.45 times— while the other half did less —up to 56.9% less than the error analysis I. An interesting finding was that albeit such inconsistent number of detections, the proportion of the error classes were almost alike: the biggest pie was occupied by Mistranslation (on average 58.3%). The Word Order held on average 7.6%. From the results, we claimed that the NMT engine of Google Inc. resolved the syntactic issue of the *es-ko* pair to a great extent and that the key challenge at this point in time was lexical choice.

[blank page]

CONCLUSION

The technological development in the MT field has been continuously advanced until it reached to a point where technologies of AI and MT have met and created the robust NMT system. Since its introduction to free commercial machine translators in around 2016, the interests of stakeholders of the translation field have spiked, and at the same time, the tension between human and machine has heightened (see Chapter 3.5).

Despite the active research and use of MT and subsequently post-editing in the workflow of translation worldwide, these topics are quite brand-new in the Korean research community, with about half of the total number of articles in relation to MT from the period of 2004 to 2018 being published after 2017 (see Chapter 1.2). Additionally, the survey conducted in this thesis has demonstrated that the usage of MT and post-editing in the Korean commercial sector is seemingly limited to date (see Chapter 3.6). Such circumstances have motivated us to evaluate the performance of NMT in the *es-ko* pair. This study can be of help to monitor the growth of NMT at this point in time and to let us cope with the current anxiety about the future of human translators by finding a way to co-exist with the machine.

To this aim, this thesis launched two core investigations. In the first investigation, we performed an NMT evaluation experiment in the Spanish-to-Korean translation in the newswire domain with six native Korean translators. The dataset was composed of 6,424 words of 253 sentences (see Chapter 5) and the NMT engine was represented by Google Translate (GT) of version 2018. The evaluation was largely based on human evaluation methods — fluency and adequacy scoring, segment ranking and post-editing— and semi-automatic evaluation measurement, HTER, considering the unproved reliability of automatic methods in a distant language pair (see Chapter 3).

In the fluency and adequacy evaluation on a 4 point Likert scale, the data showed that the engine achieved 78% of fluency level in a range of 66.5% - 88.3% and 77.8% of adequacy level in a range of 67.8% - 84%. From the analysis, it was found that the GT was more fluent than adequate in the *es-ko* pair and that when the two scores were combined on a hypothesis that the two

methods had equal importance, the overall reliability of the raw MT output was 78%. Such a figure could be also interpreted, in return, that there were about 22% errors in the output. (see Chapter. 6.1)

In the segment ranking evaluation in comparison to an NMT engine of Kakao i and human translation, GT was ranked as the lowest, with 28.2% while human translation held the first rank with 41.8%. A rather more intriguing finding was that when the result was approached from the viewpoint of a machine (Google and Kakao i) versus human, the machine-translated outputs were more favored by the evaluators with a margin of 16.4%. Some pieces of evidence were found in terms of such phenomenon, suggesting a new notion of 'a better translation', that is, to prioritize high *adequacy* of a sentence over *fluency*. In other words, despite the unnaturalness of the sentence, the *machine language* could be understood in a far broader sense than before with human cognition that was developed in such way presumably by a proliferation of raw MT outputs in our daily life. (see Chapter. 6.2)

From the post-editing evaluation, it was found that there was a 37% post-editing productivity gain over translating from scratch in a range of 12% - 53%. The average words per hour (WPH) in translation were 751 words while those of post-editing were 1,031 words. There was no standard value as to how much gain was considered to be acceptable to conclude that post-editing was more effective than translation from scratch, but considering the previous researches (see Chapter 3.5), the result we obtained was not statistically significant. From the HTER metric that calculated how many edits were performed by post-editors in the raw MT output, it was found that about 41% of the dataset was corrected to achieve an affordable quality. One of the possible reasons of the insignificant role of the given engine in post-editing could be either because the performance of NMT was not robust enough to ease the post-editing effort to facilitate a productive work or because editing incorrect word choices required considerable technical and cognitive efforts. In any case, no marked gain was detected in either post-editing productivity or post-editing effort from the given engine. (see Chapter. 6.3)

With the interesting result at hand, the second investigation was dedicated to error analysis of the system translation produced by GT. A general

error taxonomy suggested by MQM and TAUS was adapted to the scenario of the *es-ko* pair so that it featured nine error types: addition, omission, mistranslation, untranslated, punctuation, spacing, grammar, word order, and style. There was a total of 1,997 errors detected in the dataset of 6,424 words. When comparing the proportion of the errors to the dataset, it can be said that 30.77% are erroneous. Such a figure was also in line with the result obtained from the fluency and adequacy evaluation that reported the reliability of this engine to be about 78%. In a distribution of the error taxonomy, the most common errors were mistranslation, accounting for 39.2%.

In the meantime, another remarkable finding was that a word order error type accounted for only 5.76% of all types of errors (when with penalty). Such a figure was an extremely small proportion considering that the biggest challenge of SMT in a distant pair used to be the word order. As the word order was a single parameter that considered errors on a syntactic level, in contrast to a lexical level, in our error taxonomy, it could be concluded that about 6% of the errors produced by NMT in the *es-ko* pair were syntactically challenged while 94% of the errors were for a lexical reason. Such a result shed light on a possibility that the NMT engine could alleviate one of the fundamental issues that distant pairs such as the *es-ko* pair exhibited in SMT. (see Chapter 7.2)

The validity of all such findings was reconfirmed by an additional error analysis of the post-edited dataset (129 sentences —half of the dataset). Having examined the edit type of post-editing according to the identical error taxonomy applied in the previous section, it turned out that while the number of total edits differed considerably per person in a range of 362 ~ 922, the proportion of the error taxonomy was consistent throughout the post-editors, including the result of the previous error analysis. The biggest pie was taken by mistranslation, ranging from 55% to 65%, while the word order ranged from 6% to 10%. Moreover, the most common type of mistranslations that they considered erroneous was 91.47% to 97.83% a word choice. (see Chapter 7.3)

The data acquired from this thesis demonstrates that in a binary level of syntax and lexicon, the majority of the problems for both NMT and post-editing is lexical. This manifests that NMT is capable of processing a distant language pair in the same way as a close language pair, translating the languages

irrespective of their language family. In that context, we expect that the direction of the development of NMT is desirable and that the future of NMT that lies ahead is highly promising because it will keep improving itself by a large volume of big data, and such lexical issues will be dealt with accordingly for both close and distant pairs. Especially when considering that NMT has just been reported to be the new state-of-the-art in this field, it is an important finding for both NMT and the *es-ko* pair.

In the meantime, the results acquired from post-editing have demonstrated that post-editing does not contribute much to complement NMT from the perspective of time and effort, for reasons that are not clarified. However, as no clear-cut feature has been observed in this regard from this thesis, it is thought that post-editing is still valid. At this point in time, we anticipate a crucial role of *pre-editing, which* originally stands for an act of preparing a well-formed source text before using MT to enhance translatability. As a one step forward, there is a chance that some machine-oriented pre-editing rules will serve as a “manual" for MT engines. In fact, some studies have proved its usefulness in enhancing the raw MT quality and the efficiency of post-editing in SMT engines. The study of pre-editing rules tailored for NMT in the *es-ko* pair is left to be a future line of research.

In any sense, the discussion of either post-editing or pre-editing is only viable for those who have professional knowledge in both languages and does not comply with the final goal of NMT being available to the lay public. Regardless of either post-editing or pre-editing, NMT will have to exceed such limit to achieve the goal. The good news, nevertheless, is that this study has witnessed the future potential of NMT by proving that it can translate without regard to language families.

Bibliography

Ackley, H., Hinton, E., Sejnowski, J. “A Learning Algorithm for Boltzmann Machines.” Cognitive Science, 9.1 (1985): 147.

ALPAC. “Languages and Machines: Computers in Translation and Linguistics." A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council, Washington, D.C.: National Academy of Sciences, National Research Council (1966).

Aranberri, N. “What Do Professional Translators Do when Post-Editing for the First Time? First Insight into the Spanish-Basque Language Pair.” HERMES-Journal of Language and Communication in Business, (56) (2017): 89–110.

Aranberri, N., Labaka, G., Ilarraza, A., Sarasola, K. Comparison of Post-Editing Productivity between Professional Translators and Lay Users.” (2014).

Arnold, D., Balkan L., Humphreys L., Meijer S., and Sadler L. “Machine Translation: An Introductory Guide” Manchester and Oxford: NCC Blackwell (1994).

Aziz, W., de Sousa, S. C. M. and Specia, L. "PET: a tool for post-editing and assessing machine translation.” The Eighth International Conference on Language Resources and Evaluation, LREC ’12, Istanbul, Turkey, May 2012 (2012).

Babych, B. “Automated MT Evaluation Metrics and Their Limitations.” Revista Tradumàtica, No. 12, pp464–70, http://revistes.uab.cat/ojs-tradumatica/tradumatica/issue/view/5 (2014).

Bahdanau, D., Cho K., and Bengio Y. “Neural machine translation by jointly learning to align and translate.” CoRR, Accepted for oral presentation at the International Conference on Learning Representa- tions (ICLR) 2015, abs/1409.0473 (2014).

Banerjee, S. and Lavie, A. “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments.” in Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, Michigan, June 2005 (2005).

Bengio Y., Ducharme R., and Vincent P. "A neural probabilistic language model.” Journal of Machine Learning Research, 3 (2003):1137-1155.

Bentivogli, L., Bisazza A., Cettolo M., and Federico M. “Neural versus phrase-based machine translation quality: a case study.” In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016 (2016).

Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. “Neural versus Phrase-Based MT Quality: an In-Depth Analysis on English-German and English-French.” Computer Speech & Language. 49. 10.1016/j.csl.2017.11.004 (2017).

Birch, A., Osborne, M., and Blunsom, P. “Metrics for MT Evaluation: Evaluating Reordering” http://www.itl.nist.gov/iad/mig/tests/metricsmatr/2008/ (2009).

Blagodarna, O. “Enhancement of post-editing performance: introducing machine translation post-editing in translator training.” Ph.D. Thesis, Universitat Autónoma de Barcelona (2018). https://www.tdx.cat/bitstream/handle/10803/666847/olbl1de1.pdf?sequence=1&isAllowed=y.

Bojar, O. “Analyzing Error Types in English-Czech Machine Translation.” The Prague Bulletin of Mathematical Linguistics 95 (2011):63-76. 10.2478/v10108-011-0005-2.

Brown, P., Cocke J., Della Pietra S., Della Pietra V., Jelinek F., Mercer R., and Roossin P. “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 (1988).

Buchanan, B. “A (Very) Brief History of Artificial Intelligence.” AI Magazine 26 (2005): 53-60.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. “(Meta-) evaluation of machine translation.” In Proceedings of the SecondWorkshop on Statistical Machine Translation (WMT07), Prague, Czech Republic (2007).

Carl, M., Dragsted B., Elming J., Hardt D., and Jakobsen A. “The pro- cess of post-editing: a pilot study.” In Proceedings of the 8th international NLPSC workshop. Special theme: Human-machine interaction in translation, Copenhagen Business School, 20-21 August 2011. (Copenhagen Studies in Language 41), Fred- eriksberg: Samfundslitteratur (2011): 131- 142.

Carroll, J. “An Experiment in Evaluating the Quality of Translations." Mechanical Translation and Computational Linguistics 9.3-4 (1966): 55-66.

Castilho, S., Moorkens J., Gaspari F., Sennrich R., Sosoni V., Georgakopoulou Y., Lohar P., Way A., Valerio A., Barone M., Valerio A., and Gialama, M. “A Comparative Quality Evaluation of PBSMT and NMT using Professional Translators.” (2017b).

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., and Way, A. “Is neural machine translation the new state of the art?” The Prague Bulletin of Mathematical Linguistics, 108.1 (2017a):109–120.

Chang, A. “Analysis of the Current Development of Machine Translation and Interpretation in Korea: Focusing on Korean-Chinese Language Pairs.” The Journal of Translation Studies 18(2) (2017): 171-206.

Chen, Y., Liu Y., Cheng Y., and Li V. “A teacher-student framework for zero-resource neural machine translation.” In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) Vancouver, Canada, July 2017. Association for Computational Linguistics (2017): 1925–1935. URL http://aclweb.org/ anthology/P17-1176.

Cheng, Y., Yang Q., Liu Y., Sun M., and Xu W. "Joint training for pivot-based neural machine translation.” In Proceedings of IJCAI. (2017).

Cho, K., Merriënboer B., Bahdanau D., and Bengio Y. “On the Properties of Neural Machine Translation: Encoder-Decoder Approaches." In Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar (2014b): 103–111.

Cho, K., Merriënboer B., Gülc¸ehre C., Bahdanau D., Bougares F., Schwenk H., and Bengio Y. “Learning phrase representations using rnn encoder–decoder for statistical machine translation.” In Proceedings ofEMNLP 2014, Doha, Qatar, October. Association for Computational Linguistics (2014a).

Choi, D. “An Analysis of Errors in Translating English Sentences with Inanimate Subjects into Korean by Machine Translation.” Studies in Linguistics 29 (2013): 279-299.

Choi, H. “A Study on the Quality Assessment of Korean-English Patent Translation: Focusing on Quality Assessment Based on the Characteristics and Functions of Korea Patent Abstract and Korean-English Machine Translation.” Ph.D. Thesis, Ewha Woman’s University (2016).

Choi, H. and Lee, J. “A study on the evaluation of Korean-English patent machine translation - Focusing on KIPRIS K2E-PAT translation.” Interpretation & Translation, 19.1 (2017): 139-178.

Chung H. “Automatic Evaluation of Human Translation.” Journal of Interpretation & Translation Institute, 22.4 (2018): 265-287.

Conneau, A., Lample G., Ranzato M., Denoyer L., and Jégou H. "Word translation without parallel data." arXiv preprint arXiv:1710.04087 (2017).

Daems, J., Macken L., and Vandepitte S. “On the origin of errors: a fine-grained analysis of MT and PE errors and their relationship.” In Proc. of LREC, Reykjavik, Iceland (2014).

Dai, Z., Yang, Z., Yang, Y., Cohen, W. W., Carbonell, J., Le, Q. V., and Salakhutdinov, R. “Transformer-xl: Attentive language models beyond a fixed-length context." arXiv preprint arXiv:1901.02860 (2019).

de Almeida, G. and O’Brien, S. “Analysing Post-Editing Performance: Correlations with Years of Translation Experience.” EAMT May 2010 St Raphael, France [online], available: http://www.mt- archive.info/EAMT-2010-Almeida.pdf (2010).

de Sousa, M., Aziz W., and Specia L. “Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles.” Proceedings of the International Conference Recent Advances in Natural Language Processing 2011 [online] (2011). available: http://clg.wlv.ac.uk/papers/ranlp-2011_sousa.pdf.

Denkowski, M. and Lavie, A. “Choosing the Right Evaluation for Machine Translation: an Examination of Annotator and Automatic Metric Performance on Human Judgment Tasks." In Proc. of AMTA 2010 (2010a).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. “BERT: Pre- training of deep bidirectional transformers for language understanding.” arXiv preprint arXiv:1810.04805 (2018).

Devlin, J., Zbib R., Huang Z., Lamar T., Schwartz R., Makhoul J. “Fast and Robust Neural Network Joint Models for Statistical Machine Translation.” 1 (2014): 1370-1380. 10.3115/v1/P14-1129.

Doddington, G. “Automatic evaluation of machine translation quality using n-gram co-occurrence statistics." In Proceedings of the second international conference on Human Language Technology Research, HLT ’02, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. (2002): 138–145.

Doherty, S. “The impact of translation technologies on the process and product of translation”, International Journal of Communication 10 (2016): 947–969.

Doherty, S. and Brien, S. “Can MT Output be Evaluated Through Eye Tracking?” (2009).

Domingos, P. “The Master Algorithm.” Basic Books; 1 edition (2015).

Dorr, B., Snover M., and Madnani N. “Part 5: Machine Translation Evaluation [Online].” (2009). Available: https://www.cs.cmu.edu/~alavie/papers/GALE-book-Ch5.pdf.

Doyon B., White, J., Kathryn, T. “Task-Based Evaluation for Machine Translation.” (1999).

Duh, K. “Ranking vs. regression in machine translation evaluation.” (2008). 10.3115/1626394.1626425.

Echizen-ya, H., Ehara T., Shimohata S., Fujii A., Utiyama M., Yamamoto M., Utsuro T., Kando N. “Meta-evaluation of automatic evaluation methods for machine translation using patent translation data in ntcir-7.” In Proceedings of the 3rd Workshop on Patent Translation (2009): 9–16.

Escartín, P. and Arcedillo, M. :A fuzzier approach to machine translation evaluation: A pilot study on post-editing productivity and automated metrics in commercial settings.: In Proceedings of the ACL 2015 Fourth Workshop on Hybrid Approaches to Translation (HyTra), Beijing, China. ACL (2015): 40–45.

Farajian, A., Turchi M., Negri M., Bertoldi N., and Federico M. “Neural vs. Phrase-Based Machine Translation in a Multi-Domain Scenario.” (2017): 280-284. 10.18653/v1/E17-2045.

Firat, O., Cho K., and Bengio Y. “Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism." In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, California (2016a): 866–875.

Firat, O., Sankaran B., Al-Onaizan Y., Vural F., and Cho K. “Zero-resource translation with multi-lingual neural machine translation.” In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, November 2016b. Association for Computational Linguistics (2016b): 268–277. URL https://aclweb. org/anthology/D16-1026.

Francis, N. and Kuera, H. “Frequency analysis of English usage." Boston, MA: Houghton Mifflin (1982).

Frey, B. and Michael O. “The Future of Employment: How Susceptible are Jobs to Computerization?” Oxford Martin School, September (2013).

García, I. “A brief history of post-editing and of research on post-editing. New Directions in Translation Studies.” Special Issue of Anglo Saxonica, 3.3 (2012):292–310.

Görög, A. “Quality evaluation today: the dynamic quality framework.” in Proceedings of Translating and the computer 36. The International Association for Advancement in Language Technology, London: United Kingdom (2014): 155-164.

Graham Y., Baldwin T., Dowling M., Eskevich M., Lynn T., and Tounsi L. “Is all that glitters in machine translation quality estimation really gold?” In COLING (2016).

Graves, A. “Generating sequences with recurrent neural networks.” (2013). arXiv:1308.0850 [cs.NE].

Groves, D and Schmidtke, D. ”Identification and analysis of post-editing patterns for MT." (2009).

Han, L., Wong, D., and Lidia, C.“Machine Translation Evaluation: A Survey.” (2016). https://arxiv.org/abs/1605.04515.

Hochreiter, S. and Schmidhuber, J. “Long Short-term Memory” Neural Computation (1997): 1735-1780.

Hong, M. “Pre-editing oder Post-editing? - Über die Einsatzbarkeit eines Deutsch-Koreanisch MÜ Systems mithilfe der kontrollierten Sprache und des Post-editings.” Korean German Journal 21 (2010): 251-273. DOI : 10.24814/kgds.2010..21.251.

House, J. “Translation Quality Assessment: Past and Present.” (2014). 10.1057/9781137025487_13.

Humphreys, L. “User-Oriented MT Evaluation and Text Typology.” (1991). http://www.mt-archive.info/ISSCO-1991-Humphreys.pdf.

Hutchins, W. “Machine Translation: a brief history." Oxford: Pergamon Press (1995): 431-445.

Hutchins, W. “The history of machine translation in a nutshell, Technical Report." University of East Anglia (2005).

Hutchins, W. and Hutchins, J. “Machine Translation: History of Research and Applications.” Routledge Encyclopedia of Translation Technology, Routledge (2015). http://www.hutchinsweb.me.uk/Routledge-2014.pdf.

Isozaki, H., Hirao T., Duh K., Sudoh K., and Tsukada H. “Automatic Evaluation of Translation Quality for Distant Language Pairs”, Association for Computational Linguistics 11 (2010): 944–52.

Jackson, P. “Introduction to Artificial Intelligence.” Dover Publications, Inc., New York, 2nd Edition (1985).

Jaech, A., Mulcaire G., Hathi S., Ostendorf M., and Smith N. "Hierarchical Character-Word Models for Language Identification.” In Proceedings of the Fourth International Workshop on Natural Language Processing for Social Media, Austin, TX, USA (2016): 84–93.

Jean, S., Firat, O., Cho, K., Memisevic, R., and Bengio, Y. “Montreal Neural Machine Translation Systems for WMT15.” In Proceedings of the Tenth Workshop on Statistical Machine Translation, Lisboa, Portugal. Association for Computational Linguistics (2015): 134–140,

Johnson, M., Schuster M., Le Q., Krikun M., Wu Y., Chen Z., Thorat N., Viégas F., Wattenberg M., Corrado G., Hughes M., Dean J. “Google’s multilingual neural machine translation system: Enabling zero-shot translation" CoRR, abs/1611.04558 (2016). URL http://arxiv.org/abs/1611.04558.

Kang, B. And Lee, J. “The Operating Principles of Neural Machine Translation and the Accuracy of Translation - Focusing on the Chinese-Korean Translation.” The Journal of Chinese Language and Literature, 73 (2018): 253-295.

Ki, Y. “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.” Chinese Studies, 74.0 (2018): 3-32.

Kim, S., and Lee, H. “Korean to English Translation of Embedded Sentences.” The Journal of Mirae English Language and Literature 22.4 (2017): 123-147.

King, M., Popescu-Belis, A., and Hovy, E. “FEMTI: Creating and Using a Framework for MT Evaluation.” In Proceedings of the Machine Translation Summit IX, New Orleans, LA (2003): 224–231.

Klubička, F., Toral A., and Sánchez-Cartagena V. "Fine-Grained Human Evaluation of Neural versus Phrase-Based Machine Translation.” (2017): 121–32. doi:10.1515/pralin-2017-0014.

Koehn, P. and Knowles, R. “Six challenges for neural machine translation.” In Proceedings of the First Workshop on Neural Machine Translation, Vancouver, August 2017, Association for Computational Linguistics (2017): 28–39. URL http://www.aclweb.org/anthology/ W17-3204

Koehn, P. and Monz, C. “Manual and automatic evaluation of machine translation between European languages.” In Proceedings of NAACL 2006 Workshop on Statistical Machine Translation, New York, (2006).

Koponen, M. “Comparing human perceptions of post-editing effort with post-editing operations." In Proceedings of the Seventh Workshop on Statistical Machine Translation, Montral, Canada, June, Association for Computational Linguistics (2012): 181–190.

Koponen, M. “Comparing human perceptions of post-editing effort with post-editing operations.” In Proc. of WMT, Montreal, Canada (2012).

Koponen, M. “Is Machine Translation Post-editing Worth the Effort? A Survey of Research into Post-editing and Effort.” The Journal of Specialised Translation (2016): 131-148.

Krings, H. “Repairing texts: empirical investigations of machine translation post-editing processes." Kent, OH. Kent State University Press (2001).

Kwak, J. And Han, S. “Revisiting Machine Translation Error Typology through Human Post-editing.” Journal of Interpretation & Translation Institute, 22.1 (2018): 1-25.

LDC. “Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Chinese-English Translations Revision 1.0." Technical report, Linguistic Data Consortium (2002). http://-www.ldc.upenn.edu/Projects/TIDES/Translation/-TransAssess02.pdf.

LeCun, Y., Bengio, Y., and Hinton, G. “Deep Learning”. Nature, 521(7553), 436 (2015).

Lee, C., Wang, M., Yen, S., Wei, T., Wu, I., Chou, P., Chou, C., Wang, M., and Yang, T. “Human vs. Computer Go: Review and Prospect.” (2016).

Lee, H., Kim J., Shin J., Lee J., Quan Y., and Jeong Y. “papago: A machine translation service with word sense disambiguation and currency conversion.” In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, Osaka, Japan. The COLING 2016 Organizing Committee (2016): 185–188.

Lee, S. “Transcreation, machine translation, and transcreator education. Interpretation and Translation.” Journal of Interpretation & Translation 18.2 (2016): 129-152.

Lee, S. “What can we learn from trainee translators’ post-editing?” Interpretation and Translation 19.3 (2017): 37-64.

Lee, S. “A Phenomenological Study of Undergraduate Students’ Experiences of Machine Translation Post-editing.” Journal of Interpretation & Translation Institute 22.1 (2018a): 117-143.

Lee, S. “Process research into post-editing - How do undergraduate students post-edit the output of Google Translate?” The Journal of Translation Studies, 19.3 (2018b): 259-286.

Leon, A., Davis L., and Kraemer H. “The Role and Interpretation of Pilot Studies in Clinical Research.” Journal of psychiatric research 45 (2010): 626-629. 10.1016/j.jpsychires.2010.10.008.

Levenshtein, V. “Binary codes capable of correcting deletions, insertions and reversals.” Soviet Physics Doklady, 10.8 (1966): 707-710.

Liddicoat, A. "Bilingualism: An Introduction.” (1991).

Liu, C., Dahlmeier D., and Tou Ng H. “Better Evaluation Metrics Lead to Better Machine Translation." EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference (2011): 375-384.

Lommel, A., and DePalma, A. “Europe’s Leading Role in Machine Translation.” Common Sense Advisory, Boston, USA (2016).

Lommel, A., Burchardt A., and Uszkoreit H. “Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics." Tradumàtica: tecnologies de la traducció (2014): 455-463. 10.5565/rev/tradumatica.77.

Mah, S. “An Empirical Investigation of Kor-Eng Machine Translation Post-Editing - Focused on the Analysis of MT-PE of Undergraduates.” Journal of Interpretation & Translation Institute 22.1 (2018): 53-87.

Maltarollo, G., Honório M., and da Silva F. “Applications of artificial neural networks in chemical problems.” in Artificial Neural Networks - Architectures and Applications, K. Suzuki, Ed. (2013).

Minsky, M. and Papert, S. “Perceptrons." MIT Press, Cambridge, MA (1969).

Mitchell, L. “The potential and limits of lay post-editing in an online community." In Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT2015), Antalya, Turkey, May 11-13 2015 (2015).

Naveen A., Bapna A., Firat O., Aharoni R., Johnson M., and Macherey W. “The missing ingredient in zero-shot neural machine translation.” OpenReview.net. (2018).

Niessen, S., Och J., Leusch G., and Ney H. “An evaluation tool for machine translation: Fast evaluation for MT research.” In Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000) (2000): 39–45.

OECD. “Patents in artificial intelligence technologies, 2000-15: Number of IP5 patent families, annual growth rates and top inventors' economies.” in Knowledge economies and the digital transformation, OECD Publishing, Paris (2017). https://doi.org/10.1787/sti_scoreboard-2017-graph7-en.

Olive, J. “Global Autonomous Language Exploitation (GALE)." DARPA/IPTO Proposer Information Pamphlet (2005).

Papineni, K., Roukos S., Ward T., and Zhu W. "BLEU: A Method for Automatic Evaluation of Machine Translation”, ACL (2002).

Papineni, K., Roukos S., Ward T., Henderson J., Reeder F., Freeder J., and Org M. “Corpus-based comprehensive and diagnostic MT evaluation: Initial Arabic, Chinese, French, and Spanish results.” (2003).

Paul, M., Finch A., and Sumita E. “How to Choose the Best Pivot Language for the Automatic Translation of Low-Resource Languages." Transactions on Asian Language Information Processing (TALIP) 12 (2013). 10.1145/2505126.

Pinnis, M., Kalnins R., Skadiņš R., and Skadina I. “What Can We Really Learn from Post-editing?” (2016).

Plitt, M. and Masselot, M. “A Productivity Test of Statistical Machine Translation.” The Prague Bulletin of Mathematical Linguistics, 93 (2010): 7-16.

Popescu-Belis, A. “Context in Neural Machine Translation: A Review of Models and Evaluations.” (2019).

Popović, M. “Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output.” The Prague Bulletin of Mathematical Linguistics 96 (2011): 59–68, October.

Popović, M. “Comparing Language Related Issues for NMT and PBMT between German and English.” The Prague Bulletin of Mathematical Linguistics 108 (2017): 209-220. doi: 10.1515/pralin-2017-0021.

Popović, M. “Language-related issues for NMT and PBMT for English–German and English–Serbian. Machine Translation.” (2018a). 10.1007/s10590-018-9219-5.

Popović, M. “Error Classification and Analysis for Machine Translation Quality Assessment." (2018b). 10.1007/978-3-319-91241-7_7.

Puchała-ladzińska, K. “Machine Translation: A Treat Or An Opportunity For Human translators?” (2016): 89–98, doi:10.15584/sar.2016.13.9.

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

Sánchez-Gijón, P. and Torres-Hostench, O.“MT post-editing into the mother tongue or into a foreign language? Spanish-to-English MT translation output post-edited by translation trainees.” Proceedings of the 11th Conference of the Association for Machine Translation in the Americas (2014). <http://www.amtaweb.org/AMT>.

Sanchez-Torron, M. and Koehn, P. “Machine translation quality and post-editor productivity.” AMTA 2016 (2016):16.

Sanders, G., Bronsart, S., Condon, S., and Schlenoff, C. “Odds of successful transfer of low-level concepts: A key metric for bidirectional speech-to-speech machine translation in DARPA‟s TRANSTAC program.” Proceedings of LREC 2008 (2008).

Schmidhuber, J. “Deep learning in neural networks: An overview.” Neural Networks 61 (2015): 85–117.

Schwenk, H. “Continuous space language models.” Computer Speech & Language (2007).

Skadina, I. and Pinnis, M. "NMT or SMT: Case Study of a Narrow-domain English-Latvian Post-editing Project.” (2017).

Snover, M., Dorr B., Schwartz R., Micciulla L., and Makhoul J. “A Study of Translation Edit Rate with Targeted Human Annotation.” In Proceedings of Association for Machine Translation in the Americas, Cambridge, Massachusetts, USA (2006): 223– 231.

Snover, M., Madnani N., Dorr J., and Schwartz R. “Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric.” In In Proceedings of the Fourth Workshop on Statistical Machine Translation (2009): 259–268.

Song, Y. "A Critical Look at Discourses on Machine Translation.” The Journal of Translation Studies 19.1 (2018): 119-145.

Specia, L., Raj D., and Turchi M. “Machine translation evaluation versus quality estimation.” Machine Translation 24.1 (2010): 39–50.

Sutskever, I., Vinyals O., and Le Q. “Sequence to Sequence Learning with Neural Networks.” Proceedings of the Neural Information Processing Systems (2014): 3104 - 3112.

Tatsumi, M. “Correlation Between Automatic Evaluation Metric Scores, Post-Editing Speed, and Some Other Factors.” (2009).

Toral, A. “Fine-Grained Human Evaluation of Neural Versus Phrase-Based Machine Translation.” The Prague Bulletin of Mathematical Linguistics 108 (2017): 121-132. 10.1515/pralin-2017-0014.

Toral, A. and Sánchez-Cartagena, V. “A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions.” In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics 1, Long Papers. Association for Computational Linguistics, Valencia, Spain (2017): 1063–1073. http://www.aclweb.org/anthology/E17-1100.

Toral, A., and Way, A. “What level of quality can neural machine translation attain on literary text?” In Translation Quality Assessment: From Principles to Practice, eds J. Moorkens, S. Castilho, F. Gaspari, S. Doherty (Berlin; Heidelberg: Springer) (2018).

Toral, A., Wieling M., and Way A. “Post-editing Effort of a Novel with Statistical and Neural Machine Translation.” Frontiers in Digital Humanities 5 (2018): 9.

Tubay, B. and Costa-Jussà, M. “Neural machine translation with the Transformer and multisource romance languages for the biomedical wmt 338 2018 task.” In Proceedings of the Third Conference on Machine Translation, Brussels, Belgium. Association for Computational Linguistics (2018).

Vasconcellos, M. “A comparison of MT post-editing and traditional revision.” In Proceedings of the 28th Annual Conference of the American Translators Association. Ed. K. Kummer, Medford, NJ: Learned Information (1987): 409-415.

Vaswani, A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A., Kaiser Ł., and Polosukhin I. “Attention is all you need.” In Advances in Neural Information Processing Systems (2017): 5998–6008.

Way, A. “Quality Expectations of Machine Translation.” (2018). http://arxiv.org/abs/1803.08409.

Weaver, W. “Translation.” Reprinted in William N. Locke and Andrew D. Booth (eds.) Machine Translation of Languages: Fourteen Essays, Cambridge, Massachusetts: Technology Press of the Massachusetts Institute of Technology (1949): 15-33.

White, J., O’Connell T., and Carlson L. “Evaluation of machine translation.” In Human Language Technology: Proceedings of the Workshop (ARPA) (1993): 206–210.

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., and Dean J. "Google’s neural machine translation system: Bridging the gap between human and machine translation." CoRR, abs/1609.08144 (2016).

Yngve, V. "Implications of Mechanical Translation Research.” in Proceedings of the American Philosophical Society, 108.4 (1964): 275-281.

Zhang, M., Liu Y., Luan H., and Sun M. “Adversarial training for unsupervised bilingual lexicon induction.” Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (2017).

Zhechev, V. “Analysing the Post-Editing of Machine Translation at Autodesk.” (2014).

Website

Amazon Web Services. “Train Neural Machine Translation Models with Sockeye.” Amazon Web Services (2017). https://aws.amazon.com/blogs/ai/train-neural-machine-translation-models-with-sockeye/. Accessed on 26 July 2018.

British Neuroscience Association. (2003). https://www.bna.org.uk/static/uploads/resources/BNA_English.pdf. Accessed on May 7, 2018.

European Language Industry Survey. “Expectations and concerns of the european language industry.” (2018). https://www.euatc.org/images/2018_Language_Industry_Survey_Report.pdf. Accessed on January 13, 2019.

iamtrask. https://iamtrask.github.io/2015/11/15/anyone-can-code-lstm. Accessed on May 7, 2018.

Kurenkov, A. (2015) https://www.andreykurenkov.com/writing/ai/a-brief-history-of-neural-nets-and-deep-learning/. Accessed on May 7, 2018.

Nimdzi 100. (2018). https://www.nimdzi.com/the-2018-nimdzi-100-3-market-analysis/. Accessed on November 22, 2018.

Stevens, K. “Neural Network Models and Google Translate.” (2016). http://ufal.mff.cuni.cz/mtm15/files/11-neural-network-models-and-google-translate-keith-stevens.pdf. Accessed on May 7, 2018.

TAUS. “Machine Translation Post-editing Guidelines.” (2010). https://www.taus.net/think-tank/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines. Accessed on August 4, 2017.

Uszkoreit, J. “Transformer: a novel neural network architecture for language understanding.” (2017). https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html. Accessed on December 30, 2018.

Yang, Z. and Le, Q. “Transformer-XL: unleashing the potential of attention models.” (2019). http://ai.googleblog.com/2019/01/transformer-xl-unleashing-potential-of.html. Accessed on March 6, 2019.

3blue1brown. https://www.3blue1brown.com. Accessed on December 30, 2018.

Survey. https://docs.google.com/forms/d/e/1FAIpQLScgdkrGmd64yYIlZD9kGye2X6IFUu5gwbOK2DIF1ttuFm_4vg/viewform

Appendix I. Source Text

El CIS otorga a Susana Díaz una amplia mayoría en las elecciones andaluzas

La encuesta sitúa a la coalición de Podemos e IU en segundo lugar y anticipa un empate técnico entre PP y Ciudadanos

La presidenta de la Junta de Andalucía, Susana Díaz, ganaría las próximas elecciones autonómicas con más de 10 puntos de ventaja, pero lejos de la mayoría absoluta (55 escaños de un total de 109), según la encuesta hecha pública este miércoles por el Centro de Investigaciones Sociológicas (CIS).

La candidata socialista obtendría el 37,41% de los votos y entre 45 y 47 escaños.

Esta última cifra coincide con el número de diputados que el PSOE tiene en la actualidad.

Con estos resultados, y si no hubiera bloqueos, Díaz podría aspirar a desarrollar lo que ella ha denominado "gobierno de banda ancha", que le permitiría dirigir la Junta en minoría con apoyos puntuales de distintas fuerzas políticas.

La batalla se juega en el terreno del segundo puesto, donde las tres formaciones principales, PP, Adelante Andalucía —la confluencia entre Podemos e IU— y Ciudadanos llegan muy igualadas, según el sondeo.

En cuanto a porcentaje de sufragios, la coalición de izquierdas liderada por Teresa Rodríguez se situaría en segunda posición, con un 19,34%;

pero mantendría 20 escaños, la cifra que actualmente suman los diputados de ambas formaciones en el Parlamento (15 Podemos y 5 IU).

Quienes mantienen un empate técnico son PP y Ciudadanos, los primeros con un 18,66% de voto y el segundo con un 18,55%, que se traduce en idéntico número de representantes en el Parlamento andaluz: una horquilla que va de los 20 a los 22.

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.

En los anteriores comicios autonómicos, el partido de extrema derecha quedó en novena posición, por detrás de UPyD, que vuelve a presentarse en estas elecciones en todas las provincias, el ahora disuelto Partido Andalucista y PACMA, que también tiene candidaturas en todas las circunscripciones.

La encuesta del CIS no ha despejado las dudas sobre quién ostentará el centro de la derecha en Andalucía, si Ciudadanos o PP.

Si bien Ciudadanos no acaba de dar su anhelado sorpasso, los datos del sondeo evidencian un avance muy significativo en relación con los resultados de las autonómicas de 2015, donde obtuvieron nueve escaños.

El partido de Albert Rivera obtendría representación en todas las provincias, con un avance muy significativo en Cádiz, donde pasaría de 1 a 3 o 4.

La lectura para el PP no es tan positiva ya que, de confirmarse la proyección del estudio preelectoral, el 2 de diciembre supondría una bajada de entre 13 y 11 diputados, un desplome del que se benefician Ciudadanos y Vox.

Los populares pierden escaños en las ocho circunscripciones.

La coalición de Adelante Andalucía se coloca como segunda fuerza política en porcentaje de voto, pero, sin llegar repetir la mala experiencia de la confluencia entre Podemos e IU a nivel nacional, en Andalucía ambas formaciones juntas no suman más que por separado y, de acuerdo con el CIS, mantendrían los 20 escaños que ya tienen en la actualidad, por lo que no sacarían rédito por el flanco de la izquierda del desgaste del PSOE.

Ni la última legislatura ni los casi 40 años de Gobiernos socialistas continuados parecen haber horadado las expectativas electorales del PSOE de cara a la cita electoral del 2 de diciembre, justo lo contrario que le ocurre al PP.

Díaz, no obstante, debería obtener el respaldo de Adelante Andalucía o de Ciudadanos para repetir mandato.

Hasta ahora, los de Rivera se han mantenido firmes en su intención de no volver a apoyar bajo ningún concepto a la dirigente socialista.

Desde la coalición de izquierdas también son claros en que no permitirán un Ejecutivo de derechas ante un eventual pacto PP y Ciudadanos, y mantienen el discurso beligerante contra el “régimen del PSOE y el susanismo”.

De acabar segundos el 2 de diciembre, serán determinantes para la elección de la presidenta.

Un 58,4% quiere un cambio de Gobierno

Pese a que el PSOE volvería a ganar las elecciones, la valoración de los votantes respecto de la gestión de los socialistas en esta última legislatura es mayoritariamente regular (39,8%), mala (27,9%) o muy mala (15,2%).

Poco más de la mitad de los andaluces, 58,4%, cree que sería bueno un cambio de Gobierno en la comunidad, después de casi cuatro décadas con el PSOE al mando.

Un 26% de los encuestados no tiene decidido aún a quién votará.

La campaña andaluza arrancará el viernes 16 de noviembre a las 00.00 con la encuesta preelectoral como telón de fondo y se antoja determinante para disipar las dudas que arroja el CIS.

Para los partidos estas elecciones son la antesala de lo que puede suceder en las próximas citas con las urnas (municipales, autonómicas, europeas e incluso generales).

Los líderes nacionales, incluido el de Vox, van a volcarse de lleno para arañar unos resultados que permitan mantener sus expectativas.

Para cuatro de cada 10 encuestados, los máximos responsables de los partidos influyen bastante a la hora de votar.

En estas últimas semanas y pese al interés de Díaz de que solo se hable de los problemas andaluces, todos han querido introducir los asuntos de la actualidad política en clave nacional.

Sin embargo, para el 56,4% de los electores de la comunidad, lo más importante a la hora de elegir la papeleta serán los temas propios de Andalucía.

Respecto de los líderes políticos, la candidata socialista y presidenta de la Junta, Susana Díaz, es la más conocida y la más valorada con una nota de un 4,1.

El resto de los aspirantes a dirigir la comunidad también suspenden.

El líder regional de Ciudadanos, Juan Marín, y el coordinador general de IU en Andalucía y número dos de Adelante Andalucía, Antonio Maíllo, obtienen una calificación de 3,5, seguidos de la cabeza de lista por la confluencia de izquierdas, Teresa Rodríguez, con un 3,4, y el candidato popular, Juan Manuel Moreno, 3,1.

Cuatro años después de su aparición en la política andaluza, Marín continúa siendo el político menos conocido.

El Parlamento andaluz está compuesto actualmente por 109 diputados repartidos entre los cinco grupos parlamentarios que lograron representación en las elecciones de 2015.

El PSOE logró 47 diputados;

le sigue el PP que consiguió 33.

Podemos obtuvo 15 diputados, mientras Ciudadanos e IU obtuvieron nueve y cinco respectivamente.

La campaña electoral finalizará el viernes 30 de noviembre.

La sesión constitutiva del Parlamento, tras los comicios, tendrá lugar el 27 de diciembre.

Londres elige entre vasallaje o caos

Ante la incapacidad de May de gestionar el Brexit, el intento de organizar otro referéndum es solo la vía de eludir la elección entre lo malo y lo peor

Dos años después de que los ingleses decidieran recuperar su soberanía —¿la habían perdido?—, disponen de solo cuatro meses y medio para elegir entre el vasallaje o el caos.

Es la alternativa a la que les ha llevado Theresa May, que acaba de sumar la octava dimisión en su Gobierno por la errática negociación sobre el Brexit.

La situación, sin embargo, puede empeorar porque eso es exactamente lo que ha ocurrido desde que los británicos votaron por salir de la UE.

El listado de malas noticias, en efecto, es interminable:

desde la profunda división del país en dos bloques al bandazo del Brexit duro al blando y la inestabilidad política, pasando por la voladura de las líneas rojas exhibidas por Londres frente a Bruselas o la incertidumbre, letal para la economía.

Así, la libra ya ha perdido el 12% de su valor y Reino Unido será el país europeo que menos crezca en 2019 y 2020.

La última pésima noticia ha sido la dimisión, el viernes pasado, del ministro de Transportes, Jo Johnson, autor de esa diatriba entre vasallaje y caos, es decir, entre el sometimiento a las reglas del mercado único sin voz ni voto de Londres (la oferta de May rechazada por todos) o la desastrosa salida sin acuerdo el 29 de marzo.

Hermano del también dimisionario Boris y político de gran prestigio y popularidad, Jo Johnson, que votó contra el Brexit, era el encargado de pactar el estratégico acuerdo sobre Transportes, un sector clave para las futuras relaciones del Reino Unido con la UE.

Su renuncia echa gasolina al incendio declarado en algunas aerolíneas que, alarmadas por el desastre, tratan como apestados a sus accionistas británicos y les amenazan con eliminar sus derechos de voto para demostrar a Bruselas una mayoría de propiedad comunitaria y operar así sin problemas en toda la UE.

Las llamas se observan también en Getlink, gestora del Eurotúnel, por el que anualmente cruzan sin controles fronterizos 1,6 millones de camiones, 2,5 millones de coches y 21 millones de personas.

Con destino a las factorías británicas de Toyota, por ejemplo, por ahí pasan a diario suspensiones alemanas, ruedas españolas, tubos de escape holandeses o cinturones de seguridad húngaros.

Muchos de esos camiones transportan entre 8.000 y 10.000 paquetes de objetos vendidos por comercio electrónico de entrega rápida.

Si el 29 de marzo no hay acuerdo, ¿cuál será el coste y el tiempo perdido en revisiones aduaneras y trámites burocráticos?

Ante la incapacidad de Londres de gestionar este proceso, el intento de organizar otro referéndum es solo la vía de eludir la elección entre lo malo y lo peor.

O de ganar tiempo prórroga tras prórroga.

Es decir, de evitar peores noticias cuando ya se ha cumplido ese teorema adjudicado al poeta estadounidense Allen Ginsberg:

“No puedes ganar.

No puedes empatar.

Y tampoco puedes abandonar el juego”.

O sea, juegas al Brexit.

Florida ordena un recuento en las elecciones a gobernador y senador

Una diferencia de menos del 0,5% de los votos vuelve a convertir al Estado en el epicentro de una polémica disputa legal tras unas elecciones

Dieciocho años después, Florida vuelve a ser el epicentro de una polémica y reñida disputa de votos.

Las elecciones del pasado martes a senador y gobernador tendrán que pasar por un recuento automático de los votos debido a la estrecha diferencia de votos entre los candidatos demócratas y republicanos, según decidió la autoridad electoral del Estado este sábado.

Los candidatos republicanos van en cabeza en el recuento, pero la diferencia es de menos del 0,5% de los votos.

A diferencia de las presidenciales de 2000 que se decidieron judicialmente tras una situación similar en Florida, esta vez no hay problemas con el diseño de las papeletas.

Pero sí se han repetido las protestas ante las oficinas electorales y el desembarco de abogados en medio de muchos nervios.

La noche del martes, los candidatos republicanos a senador por Florida, Rick Scott, y a gobernador, Ron DeSantis, clamaron victoria aunque oficialmente no se les había proclamado ganador porque quedaban por contar dos condados, de mayoría demócrata.

Uno de esos distritos examinó este viernes si algunas papeletas que se habían considerado defectuosas en realidad eran correctas.

Desde el martes, se ha ido achicando la diferencia de votos de los republicanos con sus rivales, el actual senador, el demócrata Bill Nelson, y el candidato a gobernador, Andrew Gillum.

El viernes, la ventaja de Scott era de solo el 0,18% de votos y la de DeSantis, del 0,44%.

Por ley, es obligatorio un recuento electrónico de papeletas si el sábado la diferencia de votos es de menos de 0,5 puntos.

Finalmente, el secretario del Estado (autoridad electoral) de Florida hizo oficial a mediodía del sábado que la diferencia de votos obliga a un recuento automático.

Los resultados deben estar listos el próximo jueves por la tarde.

Si después de ese recuento la diferencia sigue siendo de menos de 0,25 puntos, se hará otro recuento a mano.

Paralelamente, la elección ha sido judicializada a través de demandas de los dos aspirantes a senador.

Scott, que es el actual gobernador de Florida, y el presidente estadounidense, Donald Trump, han acusado sin pruebas a dos condados, de mayoría demócrata, de cometer fraude.

“Es una desgracia”, dijo Trump, quien en las presidenciales de 2016 atribuyó a irregularidades falsas su derrota, en el número total de votos, frente a la demócrata Hillary Clinton.

Scott logró el viernes que un juez obligara a la autoridad electoral del condado de Broward a entregar toda la información sobre los votos emitidos.

Nelson, su rival demócrata, le acusa de utilizar su potestad como gobernador para sugerir que puede pedir una investigación policial.

Las autoridades han dicho no tener pruebas de fraude alguno, aunque ese condado ha estado envuelto en polémicas en el pasado.

En unas elecciones de 2016, eliminó un archivo de papeletas físicas pero mantuvo uno digitalizado.

Tanto Broward como Palm Beach fueron dos condados clave en la disputa tras las elecciones presidenciales de 2000 entre el republicano George W. Bush y el demócrata Al Gore.

Florida acabó determinado quién se hacía con la presidencia.

Bush ganó oficialmente el Estado después de que el Tribunal Supremo paralizara un recuento de votos tras detectarse posibles irregularidades.

Entonces, ambos condados utilizaban papeletas en las que el votante hacía un agujero dentro de un círculo.

Esos formatos se cree que confundieron a muchos votantes y que podrían haberle costado la victoria a Gore.

También se encaminan a recuentos las elecciones a senador por Arizona y a gobernador de Georgia.

Según los últimos datos, la demócrata Kyrsten Sinema mantiene una mínima ventaja frente a la republicana Martha McSally, que defiende el escaño conservador en Arizona.

Una victoria demócrata en Arizona sería especialmente significativa del calibre de la respuesta contra los republicanos.

Ambas campañas llegaron este viernes a un acuerdo para esclarecer posibles irregularidades en el conteo de votos en zonas rurales.

Mientras en Georgia, la demócrata Stacey Abrams ha prometido tomar acciones legales para garantizar que todos los votos se cuenten correctamente.

Esas elecciones fueron polémicas desde el primer momento porque el candidato republicano, Brian Kemp, como secretario de Estado de Georgia, es el responsable de gestionar comicios y los demócratas le acusaron de tomar medidas para restringir el voto de la población negra, que iba a ser clave para Abrams, que sería la primera gobernadora afroamericana de EE UU.

Las elecciones legislativas, un referéndum sobre Trump

Los estadounidenses eligen este martes a congresistas y autoridades estatales y locales.

Pero los comicios son ante todo un juicio al «trumpismo« que determinará el futuro del presidente

Si esta semana alguien hubiera salido de un coma, podría pensar que despierta en el otoño de 2020 y que Donald Trump se juega su reelección.

El presidente de EE.UU. no está en las papeletas de las elecciones legislativas de este martes, pero sí lo está su figura y su futuro político.

La cita con las urnas es un juicio a su presidencia -formidable para unos, infame para otros, polarizante sin duda-, un referéndum al «trumpismo».

Sus resultados, además, determinarán su capacidad de maniobra en la segunda parte de su mandato y sentarán las bases de la reelección.

Todo explica que Trump apenas se haya quitado el traje electoral este otoño.

Se ha subido el escenario electoral casi una vez cada dos días.

Desde el 6 de septiembre hasta el 6 de noviembre, el día de las elecciones, habrá protagonizado treinta mítines.

«Yo no estoy en las papeletas, pero sí lo estoy, porque esto también es un referéndum sobre mí», dijo recientemente en un mitin en Misisipí.

«Pretended que estoy en la votación», pidió a sus seguidores.

Es una exigencia fácil de cumplir.

Trump ha monopolizado el discurso político desde que se presentó a la presidencia de EE.UU. en junio de 2015.

Una vez en la Casa Blanca, quienes confiaban -sobre todo, sus aliados republicanos- en que adoptaría un tono «presidencialista» se han topado con que Trump sigue en campaña de sí mismo.

Su lenguaje de confrontación, el insulto habitual, las referencias de tono racista y sexista y las filtraciones sobre el caos en la Casa Blanca han copado los medios, que aman odiar al presidente.

Su figura, además, domina a ambos partidos.

Los apoyos de Trump -con una opinión muy favorable de sus bases- han puesto y quitado candidatos republicanos, que no tienen más remedio que proclamar su adhesión al jefe.

El caso más claro es el de Ted Cruz:

como candidato a la presidencia llamó «mentiroso patológico» y «cobarde llorón» a Trump en 2016 y el mes pasado se vio obligado a invitarlo a un mitin para conservar su puesto como senador en Texas.

Los demócratas, por su lado, compiten por demostrar quién es más «anti Trump».

Los nombres que aparecen en las papeletas son los de los candidatos al Congreso y cientos de autoridades estatales, desde gobernadores de estados hasta concejales de distrito.

La Cámara de Representantes renueva por completo sus 435 miembros, mientras que el Senado elige a un tercio de sus cien legisladores.

En los primeros dos años de presidencia de Trump, las dos cámaras del Congreso tenían mayoría republicana, lo que no ha sido suficiente para que Trump impulsara puntos centrales de su agenda, como el desmantelamiento de la reforma sanitaria de Barack Obama o la financiación del muro con México, la estrella de su campaña.

Ahora, todo apunta a que el bloqueo legislativo será mucho mayor tras las legislativas.

Las encuestas dicen que los demócratas recuperarán la Cámara de Representantes, y la duda es cuál será la ventaja final.

En el Senado, el vuelco demócrata es mucho más difícil, porque buena parte de los escaños en juego están en territorio favorable para los republicanos.

De hecho, hay posibilidades de que los republicanos refuercen su mayoría, que hoy es mínima (51 senadores conservadores, por 49 demócratas).

Con un resultado así, los demócratas tendrán la posibilidad de torpedear la agenda política de Trump, pero se encontrarán con la patata caliente de qué hacer con uno de los temas que ha protagonizado su presidencia:

la investigación del supuesto complot de la campaña de Trump con Rusia, cuyos resultados no tardarán en llegar.

La Cámara de Representantes es la responsable de impulsar el 'impeachment' o juicio político del presidente.

Las corrientes más izquierdistas del partido demócrata optarán por ese camino, lo que convertiría la segunda mitad del mandato de Trump en un circo político que estimularía al electorado republicano en defensa de su presidente.

Las encuestas -como la historia se ha encargado de mostrar en los últimos años- pueden equivocarse.

Sobre todo tras un cierre de campaña virulento, agitado por la campaña del miedo desatada por Trump a raíz de la caravana de inmigrantes y por los episodios de violencia -los paquetes bomba a sus enemigos politicos y, sobre todo, la matanza en la sinagoga de Pittsburgh- que parte de la opinión pública relaciona con el mensaje agresivo del presidente.

Habrá que ver si todo ello está detrás de la alta participación que se anticipa para este martes.

La popularidad del presidente mejora cuando deja los exabruptos de lado y consigue resultados, como la bajada de impuestas o la confirmación de Brett Kavanaugh como juez del Tribunal Supremo.

No es el camino elegido por Trump en el final de la campaña, que ha preferido repetir el guión de 2016.

Nadie le puede negar que entonces funcionó.

Polémicas elecciones en las regiones separatistas del este de Ucrania

La Unión Europea, Estados Unidos y Kiev califican los comicios de «ficticios e ilegítimos»

Pese a las críticas y advertencias de la OSCE y la mayor parte de los países de Occidente, las autoridades rebeldes de las repúblicas separatistas del este de Ucrania, Donetsk y Lugansk, se disponen hoy a celebrar «comicios» para elegir a sus respectivos líderes y a los diputados de los llamados «Consejos Populares».

La Organización para la Seguridad y la Cooperación en Europa (OSCE) ha sido la última en condenar una decisión que viola los acuerdos de paz de Minsk.

Estas elecciones han sido también criticadas e incluso calificadas de «ficticias e ilegítimas» por la Unión Europa, EEUU y las autoridades de Kiev.

Según Enzo Moavero, presidente de turno de la OSCE y ministro de Exteriores italiano, los comicios en el este de Ucrania «van en contra de la letra y el espíritu de los acuerdos de Minsk».

El presidente ucraniano, Piotr Poroshenko, cree que «Rusia debería haber influido para evitar la celebración de las elecciones y ha hecho lo contrario, demostrando así que no quiere propiciar una solución pacífica».

Según su opinión, «los resultados de esta convocatoria jamás serán reconocidos por la Comunidad Internacional (...) y provocarán la adopción de nuevas sanciones contra Rusia».

Pero en Moscú piensan de otra manera.

El portavoz del Kremlin, Dmitri Peskov, dijo el martes que las elecciones organizadas por los separatistas «no vulneran los acuerdos de paz».

A juicio de Peskov, «quienes demuestran poco deseo de que se aplique lo pactado en Minsk son las autoridades de Kiev».

Lo firmado en Minsk, el 12 de febrero de 2015, contempla la devolución a Ucrania del control de la frontera con Rusia, en los tramos que corresponden a Donetsk y Lugansk, y la celebración en ambos territorios de elecciones realmente libres y democráticas con arreglo a la legislación ucraniana.

A cambio, Kiev debe conceder a los dos enclaves un sistema de autogobierno.

Pero la desconfianza mutua mantiene el proceso en punto muerto mientras los enfrentamientos armados se suceden de forma esporádica.

El «presidente» en funciones de Donetsk y favorito para obtener la mayoría de los votos, Denís Pushilin, explicó a comienzo de mes que la república «necesita celebrar estas elecciones» para dotarse de líder y asamblea local.

Prometió que transcurrirán «de forma transparente y cumpliendo todos los estándares internacionales», algo que no se cree nadie salvo Moscú.

El jefe de los Servicios de Seguridad de Ucrania (SBU), Vasili Gritsak, sostiene que las actas con los resultados de los comicios «ya están confeccionadas».

El pasado agosto, fue asesinado en un atentado todavía sin esclarecer el que había sido jefe de la autoproclamada República Popular de Donetsk (DNR) durante casi cuatro años, Alexánder Zajárchenko.

Pushilin se puso entonces al frente de la DNR de forma interina.

En la vecina también autoproclamada República Popular de Lugansk (LNR) se celebran igualmente elecciones y el que parte con ventaja es el que desempeña interinamente el cargo de máximo dirigente, Leonid Pasechnik.

Sustituyó hace justo un año a Ígor Plotnitski, que fue desplazado por un oscuro golpe de mano.

Donetsk y Lugansk se levantaron en armas contra el Gobierno ucraniano en abril de 2014, un mes después de que Rusia se anexionara Crimea.

Estalló entonces una guerra que Moscú atizó enviando armas, dinero y hombres en apoyo de los separatistas.

Desde entonces, según la ONU, el conflicto ha acabado con la vida de 10.000 personas.

Los acuerdos de Minsk, alcanzados bajo la mediación de Alemania y Francia, fueron un intento, por ahora fallido, de poner fin definitivamente al enfrentamiento armado entre el Ejército ucraniano y las milicias rebeldes.

Misión 2019: evitar que los euroescépticos utilicen fondos europeos para boicotear la UE

Pese a que las instituciones de la UE han reforzado sus mecanismos de control de cuentas, partidos euroescépticos y sus fundaciones han decidido explotar las oportunidades europeas en lugar de renunciar a todo lo relacionado con la Unión

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.

A lo largo de la última década, las instituciones de la UE han reforzado sus anticuerpos para enfrentarse a prácticas abusivas con dinero europeo.

Aun así, los partidos euroescépticos y las fundaciones políticas han decidido explotar las oportunidades europeas en lugar de boicotear todo lo relacionado con la Unión.

El portavoz del Parlamento Europeo, Jaume Duch, también trata de lidiar con el auge de los partidos populistas y euroescépticos:

«No se trata de dar a ciertos países una cantidad determinada de dinero, sino saber en qué políticas invertir.

El Eurobarómetro dice que los temas más importantes son el empleo, el crecimiento económico, la gestión de la migración, el cambio climático, medio ambiente y protección social», afirma a ABC.

El problema, según el portavoz de la UE, es que para satisfacer estas necesidades los Estados miembros deberían asumir una contribución mayor, pero los países no están dispuestos a gastar más dinero en el presupuesto común una vez que el Brexit se consuma y aleje a la Unión de un contribuyente fundamental como Reino Unido.

«Pero el populismo no puede ser la solución a estos problemas», advierte.

Desde el lleno de Vistalegre, el partido español VOX, inspirado en los consejos del exasesor de Donald Trump Steve Bannon y el éxito del antiguo Frente Nacional en Francia o la Liga en Italia, aspira a conseguir representación en el Parlamento Europeo gracias a la circunscripción electoral única, lo que supondría un notable estímulo tanto a nivel de exposición pública como para sus finanzas, tal como hizo Podemos en 2014.

«Es curioso ver cómo algunos partidos antieuropeos cuando entran en el Gobierno moderan su lenguaje, pero en otros países ocurre todo lo contrario:

Italia, Hungría y Polonia son diferentes», considera Paul Schmidt, editor del capítulo austriaco del libro «El futuro de Europa - visiones desde las capitales», presentado en el Real Instituto Elcano.

Aunque esos partidos van a ser más influyentes, el responsable del apartado español del mismo título, Ignacio Molina (Real Instituto Elcano), cree que será más difícil articular una posición común.

«No existe un solo partido euroescéptico, existen veintitantos, cada uno con su sensibilidad.

De forma divertida, el partido más euroescéptico extremista húngaro es antieslovaco y el eslovaco, antihúngaro;

no va a ser tan sencillo ponerlos en común», agrega.

En términos de ingresos fijos, cada eurodiputado recibe un salario de alrededor de 6.600 euros netos al mes, 4.400 para gastos de oficina sin justificar y alrededor de 24.000 para asesores.

Además, cada parlamentario tiene ingresos mensuales variables (viajes por día) y una pensión de jubilación del 3.5% del salario por un año completo.

«El problema real es cómo controlar este dinero:

si no va a la sesión plenaria, debería tener menos dinero:

si asiste a menos del 51% de los votos en las sesiones plenarias, obtendrá la mitad del recorte salarial.

Luego está la partida pensada para gastos de la oficina, como una nueva impresora o iPad, pero que carece de control.

Eurodiputados de los verdes y socialdemócratas dicen exactamente «pagué 200 euros por un traje», pero no están obligados a hacerlo.

Además, hay muchas personas poderosas de fuera de Europa que financian dando enormes subvenciones a estos partidos y tenemos que evitarlo», dicen fuentes del Parlamento Europeo a ABC.

Según la Declaración 11 del Tratado de Niza de 2001, la financiación de la UE para los partidos políticos europeos no puede utilizarse para financiar, ni directa ni indirectamente, campañas nacionales.

«Mientras que los subsidios europeos debían reforzar el sistema político de la UE y abordar el déficit democrático, los euroescépticos han utilizado principalmente sus recursos para apoyar a sus partidos nacionales», escribe Wouter Wolfs, investigador del Instituto de Gobernanza Pública de KU Leuven, para el Foro contra la corrupción y la integridad de la OCDE, donde resalta la evolución de las formaciones euroescépticas desde hace veinte años, pasando del total boicot a una participación muy intensa en las instituciones y elecciones europeas.

En 2016, Morten Messerschmidt, eurodiputado danés del ahora desaparecido Movimiento por una Europa de Libertades y Democracia (MELD) que utilizó fondos europeos para apoyar campañas nacionales, se vio obligado a devolver este dinero.

«Con Marine Le Pen, los servicios del Parlamento han demostrado claramente que usaron fondos europeos para trabajar en asuntos nacionales.

En ese caso, la ley es que tienen que devolver este dinero», dice la italiana Mercedes Bresso, vicepresidenta del grupo del Partido Socialista Europeo en el Parlamento.

Junto con el alemán Rainer Wieland (Partido Popular Europeo), Bresso lideró una propuesta de resolución sobre la financiación de los partidos políticos y las fundaciones políticas a nivel europeo.

La oficina antifraude (OLAF) ya ha llevado casos como este a los tribunales.

Además, el Comité Europeo de Control Presupuestario descubrió que el grupo Europa de las Naciones y la Libertad, donde se agrupan partidos de extrema derecha como el antiguo Frente Nacional francés, la Liga (Italia), el Partido de la Libertad de Austria (FPÖ, socio del Gobierno de Sebastian Kurz en Austria) y el Partido de la Libertad (PVV, Holanda), presuntamente gastó ilegalmente 427.000 euros de fondos de la UE en 2016.

Lo emplearon en bebidas caras (champagne), cenas y regalos para empleados.

La Comisión de Control Presupuestario del Parlamento Europeo (CONT) recomendó recuperar el dinero a partir de un recorte proporcional en futuros fondos para el grupo Europa de las Naciones y la Libertad.

«No creo que el Parlamento Europeo esté ahora mejor preparado porque el sistema siempre ha funcionado.

Ha destapado fraudes y ahora ha habido un caso en el Tribunal en el que se ha pedido devolver el dinero», dijo Laurin Berresheim, quien prepara el trabajo de Arndt Khon (eurodiputado de los socialdemócratas alemanes) en el Comité de Control Presupuestario.

Este organismo supervisa cómo se gasta el dinero europeo:

no solo cómo lo emplean miembros de las instituciones, sino también en lo que se refiere a proyectos europeos.

Según Gabriel Richard-Molard, asistente de Bresso y que también trabajó en esta moción, «hace diez años no teníamos un registro de transparencia adecuado ni un reglamento eficaz, tampoco una dirección general de finanzas.

Tuvimos que adaptarnos, sí, definitivamente estamos mejor preparados que antes».

Para Berresheim, los eurodiputados tienen todas las mismas reglas sobre cómo gastar su dinero;

«pero las reglas no deberían ser más estrictas para los representantes de determinado color político:

sigue siendo una democracia».

El 13 de septiembre de 2017, la Comisión Europea adoptó una propuesta para enmendar las normas sobre el estatuto y la financiación de los partidos políticos europeos y las fundaciones políticas europeas y con la que se pretende revisar el actual reglamento de 2014, previo a las elecciones europeas de 2019, para atajar varias lagunas legales.

«Las enmiendas propuestas se centran en proporcionar más transparencia, mejorar la legitimidad democrática y fortalecer la aplicación de la ley», según el informe sobre estatuto y el financiación de partidos políticos y las fundaciones europeas redactado por Wieland y Bresso.

Esta moción, pendiente de la votación en el plenario, establece que particulares ya no pueden patrocinar el registro de un partido político europeo, solo partidos políticos.

«La financiación ahora se vinculará con los resultados reales de las votaciones, haciendo que los partidos políticos europeos sean más europeos», dijo el relator Rainer Wieland en una sesión informativa.

En 2018, según el informe de Bresso y Wieland, se ha asignado un total de 32,44 millones de euros para subvenciones a los partidos políticos europeos dentro del presupuesto de la UE, y 19,32 millones de euros para subvenciones a fundaciones políticas europeas.

Un donante puede donar un máximo de 18.000 euros al año, sin embargo, los nombres de los donantes solo deben ser publicados para cualquier donación por encima de los 3.000 euros.

Como se hizo eco el portal especializado «EUObserver», dos partidos ultranacionalistas en el Parlamento Europeo ya no tendrán derecho a recibir fondos de la UE después de no cumplir con las condiciones mínimas.

La autoridad independiente que supervisa los partidos y fundaciones anunció a finales de septiembre que la Alianza de los Movimientos Nacionales Europeos (AEMN) y la Alianza para la Paz y la Libertad (APF), donde se encuentra el UKIP británico, uno de los partidos más denunciados por abusos de los fondos europeos, han sido eliminadas de su lista de registro.

A este último, ya disuelto, se le impuso la devolución 1,1 millones de euros al Parlamento Europeo por uso indebido de dinero europeo.

El Parlamento Europeo está discutiendo el nuevo marco financiero plurianual para los años 2021-2027, donde está sobre la mesa la opción de otorgar más capacidad económica a la Autoridad para controlar el uso de los Fondos de la UE.

«En este momento tenemos tres actores diferentes que verifican el uso de los Fondos de la UE: Dirección General de Finanzas (DGFIN), el Comité Independiente de Personas Eminente y también la Autoridad para los Partidos Políticos Europeos y las fundaciones políticas europeas (EUPPF) (...)

Pero aparte de la cuestión de tener una regulación adecuada para evaluar el uso indebido de los fondos europeos para actividades políticas, el punto principal es poder analizar de manera eficiente la forma en que utilizan los Fondos de la UE.

Tenemos muy pocos recursos humanos aquí», concluye Bresso.

Prevén alta tasa de participación en las elecciones estadounidenses

Estados Unidos celebra el martes 6 elecciones de mitad de periodo.

El estado de Vermont fue el primero en abrir las puertas de sus colegios electorales, a las 5 de la madrugada, hora local.

Las votaciones comienzan de este a oeste de forma consecutiva por los distintos husos horarios, y los comicios finalizarán al cerrar las urnas en Hawái a las 11 de la noche.

Estos comicios renovarán los 435 escaños de la Cámara de Representantes, 35 de los 100 escaños del Senado y 36 de los 50 gobernadores.

Se prevé una tasa de votación extraordinariamente elevada, pues para el pueblo estadounidense es la ocasión de evaluar los primeros dos años de Donald Trump como presidente.

Si bien la competencia es reñida, los expertos prevén que el Partido Demócrata renovará la mayoría en la cámara baja por primera vez en ocho años -suponiendo un gran obstáculo para Trump, que busca la relección en 2020-, mientras que estiman que el Senado mantendrá el predominio republicano.

Investigan al gobernador de Jeju por presunta violación de la ley electoral

El gobernador de Jeju, Won Hee Ryong, fue interrogado por la Policía sospechoso de haber violado la Ley de elecciones a cargos públicos.

En concreto, la acusación contra el político se centra en no respetar el periodo oficial de campaña al presentar sus promesas electorales y visiones antes de que el mismo empezara, y por divulgar falsa información al comentar sin pruebas objetivas que en la afluencia masiva de capital chino en el desarrollo urbano de Jeju estuvieron implicados, tanto el político que fue su rival en las elecciones, como un ex gobernador de Jeju.

El interrogatorio duró nueve horas y Won regresó a casa a las 3:00 de la madrugada del sábado 29.

El presidente Moon Jae In sigue perdiendo apoyo

La tasa de apoyo al presidente Moon Jae In y a su partido, The Minjoo, cayó al nivel más bajo desde mayo de 2017, mes en que se celebraron elecciones presidenciales prematuras tras la destitución de la ex presidenta Park Geun Hye.

Este es el resultado que arrojó la encuesta realizada por la agencia de estudios de opinión pública Realmeter a 1.005 surcoreanos mayores de edad entre el 13 y el 14 de agosto.

En ella el mandatario registró un apoyo popular del 55,6%.

El porcentaje es el más bajo en lo que va de su administración.

El descenso se debe, según Realmeter, a la reforma del sistema de pensión nacional y a la confianza de la población en el actual Gobierno, que decrece tras ser declarado inocente el ex gobernador de Chungcheong del Sur, Ahn Hee Jung, en su juicio por abuso sexual.

Asimismo el partido en el poder, The Minjoo, perdió apoyo al contar con un 37% de respaldo ciudadano, unos tres puntos porcentuales menos que hace una semana.

El sondeo presenta un margen de error de 3,1 puntos porcentuales y un nivel de confianza del 95%.

Las regionales de 2018 registran la 2ª mayor participación de la historia

Las elecciones regionales del 13 de junio registraron una participación del 60,2%, según datos provisionales de la Comisión Nacional Electoral.

Eso significa que de los 42.900.000 ciudadanos con derecho de sufragio, 25.840.000 votaron.

Dicha tasa es además la segunda más elevada de la historia después del 68,4% de los primeros comicios regionales celebrados en 1995.

Este resultado contradice los pronósticos planteados antes de las elecciones, que vaticinaban una participación relativamente baja al desviarse la atención por la cumbre entre Corea del Norte y Estados Unidos, así como la ya anticipada supremacía del oficialismo sobre la oposición.

No obstante, se analiza que el interés de la población en la política, que empezó a aumentar a partir de las protestas contra el gobierno de la ex presidenta Park Geun Hye, su destitución, y las elecciones presidenciales prematuras organizadas en mayo de 2017;

se tradujo en altas tasas de participación electoral para las regionales, superando incluso el 20% en la votación anticipada.

En tanto, en las legislativas parciales, que por primera vez en la historia tuvieron lugar simultáneamente a las regionales, la participación marcó el 60,7%, rebasando en gran medida el 53,9% de las últimas celebradas el 12 de abril del año pasado.

Comienza el recuento electoral

El recuento de votos para las séptimas elecciones regionales comenzó a las 18:30 h del miércoles 13, tras cerrar los centros electorales a las 18:00 h.

Según la Comisión Nacional Electoral, 78 mil contadores de votos, 12 mil observadores, 105 mil administradores y 2.500 máquinas para el recuento automatizado de papeletas han sido movilizados para el escrutinio a nivel nacional.

Las urnas para la votación anticipada, tanto en persona como por correo, ya fueron trasladadas a los centros de escrutinio escoltadas por los observadores, integrantes del comité electoral y agentes de la policía tras cerrar la votación, mientras que las urnas de hoy también fueron llevadas para el recuento bajo supervisión, tras ser selladas.

Asimismo, un representante del comité electoral afirmó que para elevar la transparencia, los contadores verifican la cantidad de papeletas contabilizadas por las máquinas, y todo el proceso de escrutinio es supervisado por observadores seleccionados entre ciudadanos comunes con derecho a voto.

Se estima que, en algunas circunscripciones, se podrá saber quien ocupará los cargos a partir de las 22:30 h.

Appendix II. System Translation of Google Translate

CIS는 Susana Díaz에게 안달루시아 선거에서 대다수를 준다.

이 설문 조사는 Podemos와 IU의 연합을 두 번째 장소에두고 PP와 Citizens 사이의 기술적 인 관계를 예상합니다

유다 데 안달루시아 (Susana Díaz) 의장은 다음 주 지방 선거에서 10 점 이상의 우위를 점할 것으로 예상했지만, 이번 수요일에 센터에 의해 공개 된 설문 조사에 따르면, 절대 다수 (109 개 가운데 55 석) 사회학 연구 (CIS).

사회당 후보는 37.41 %의 표와 45 ~ 47 표의 표를 얻는다.

이 마지막 숫자는 PSOE가 현재 보유한 대리인의 숫자와 일치합니다.

이러한 결과와 함께 장애가없는 경우 Díaz는 그녀가 "광대역 정부"라고 부른 것을 발전 시켜서 그녀가 다른 정치 세력으로부터 특별한 지원을받는 소수 민족위원회를 이끌 수있게 할 수있었습니다.

이 전투는 2 위의 지형에서 진행되는데 PP, Adelante Andalucía (Podemos와 IU- 시민들의 합류는 매우 평평하다.

투표율에 대해서는 테레사 로드리게스 (Teresa Rodríguez)가 이끄는 좌파 연합이 19.34 %로 2 위에 올랐다.

그러나 의회의 두 대의 대의원이 현재 추가로 (15 개의 Podemos와 5 개의 IU) 추가하는 20 석을 유지할 것입니다.

기술적 인 관계를 유지하는 사람들은 PP와 시민이며, 18.66 %의 투표권을 가진 첫 번째 사람과 18.55 %의 사람인 두 번째 사람은 안달루시아 의회 의원과 같은 수의 사람으로 번역됩니다. 머리말은 *20 22*시에

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.

이전의 지방 선거에서 맨 오른쪽 파티는 UPyD 뒤에서 9 번째 위치에 있었고, 모든 주에서 선거로 돌아 왔고, 지금은 해체 된 Andalusian Party와 PACMA도 모든 선거구에서 후보가되었습니다.

CIS 설문 조사는 시민 또는 PP 여부에 관계없이 안달루시아에서 권리의 중심을 누가 잡을지에 대한 의구심을 해소하지 못했습니다.

시민들이 열광한 sorpasso를 제공하는 것만은 아니지만 설문 조사 데이터는 2015 년 자치 결과 9 개를 차지한 결과와 관련하여 매우 중요한 진전을 보였습니다.

앨버트 리베라 (Albert Rivera)의 당은 카디스에서 매우 중요한 진보와 함께 모든 지역에서 대표권을 얻었으며 카디스는 1에서 3 또는 4로 넘어갔습니다.

예비 선거 연구 계획이 확정되면 12 월 2 일에 시민과 복스에게 혜택을주는 붕괴가 13 명에서 11 명으로 감소 할 것이기 때문에 PP에 대한 독서가 그렇게 긍정적이지는 않습니다.

인기는 8 개 지구의 좌석을 잃어 버린다.

Adelante Andalucía 연합은 투표의 비율로 두 번째 정치 세력으로 배치되지만 안달루시아에서 Podemos와 IU 간의 합류에 대한 나쁜 경험을 되풀이하지 않고 두 가지 형성은 함께 개별적으로 추가하지 않으며 CIS와 함께, 그들은 이미 가지고있는 20 석을 유지할 것이므로 PSOE 찬용의 왼쪽 측면에 대해 돋음받지 않을 것입니다.

마지막 입법부와 거의 40 년 동안 지속 된 사회주의 정부 모두 12 월 2 일 선거를 앞두고 PS에 대한 선거 기대를 뚫지 못했다.

그러나 디아즈는 아 델란 테 안달루시아 (Adelante Andalucía) 또는 시우 다다 노 (Ciudadanos)의 지원을 받아 위임을 반복해야한다.
지금까지 리베라는 어떤 식 으로든 사회주의 지도자를지지하지 않을 것이라는 확고한 견해를 유지해 왔습니다.
좌파 연립 정부로부터 그들은 최종 PP와 시민 협정 전에 권리 집행을 허용하지 않으며 "PSOE 정권과 수사 니즘"에 대한 호전적인 담론을 유지할 것임을 분명히했다.
12 월 2 일에 두 번째를 마치면 대통령 선거에 결정적인 역할을 할 것입니다.
58.4 % 정부 변화를 원한다.
PSOE가 다시 선거에서 승리 할지라도,이 마지막 기간에 사회 주의자들의 관리에 관한 유권자 평가는 대부분 규칙적 (39.8 %), 나쁜 (27.9 %) 또는 매우 나쁘다 (15. 2 %).
안달루시아 인 중 절반 이상 (58.4 %)은 PSOE를 담당 한 거의 40 년 후 공동체에서 정부의 변화가 좋을 것이라고 생각한다.
응답자의 26 %는 투표 대상자를 아직 결정하지 않았습니다.
안달루시안 캠페인은 11 월 16 일 금요일 00:00에 예비 선거 조사를 배경으로 시작되며 CIS가 제기 한 의구심을 없애는 결정적인 요인으로 보인다.
당사자들에게이 선거는 여론 조사 (시, 지역, 유럽 및 심지어 일반)에서 예정된 약속에서 일어날 수있는 일의 서곡입니다.
Vox를 포함한 전국 지도자들은 자신들이 기대치를 유지할 수 있도록 결과를 뒤집기 위해 전복하려고합니다.
응답자 10 명 중 4 명에 대해서는 당사자 대표가 투표 할 때 큰 영향력을 행사합니다.
지난 몇 주간 그리고 디아즈가 안달루시아 문제에 대해서만 이야기하는 것에 관심을 갖고 있음에도 불구하고 모든 사람들은 정치적 시사 문제를 국가 핵심 용어로 소개하기를 원했습니다.
그러나 지역 사회 유권자의 56.4 %가 안달루시아의 대상이 될 것입니다.
정치 지도자들에 관해서, 사회주의 후보자이자 이사회 의장 인 Susana Díaz는 4.1 등급으로 가장 잘 알려져 있고 가장 가치가 있습니다.
공동체를 이끌어 나갈 다른 유망주들도 일시 중지합니다.
후안 마린과 안달루시아 IU 총지배인 인 아델란테 안달루시아 (Andelucía), 안토니오 메이 올로 (Antonio Maíllo)의 지역 지도자는 3.5 점을 얻었고 왼쪽의 합류점 인 테레사 로드리게스 (Teresa Rodríguez) 3.4, 인기 후보 인 후안 마누엘 모레노 (3.1).
안달루시안 정치에 출마 한 지 4 년 후 마린은 계속 정치인으로 알려지지 않았다.
안달루시안 의회는 현재 2015 년 선거에서 대의원을 얻은 5 명의 의원들로 구성된 109 명의 의원들로 구성됩니다.
PSOE는 47 명의 대리인을 확보했습니다.
그 다음에는 33 점을 얻은 PP가 나옵니다.
Ciudadanos와 IU가 각각 9 점과 5 점을 얻은 동안 우리는 15 명의 대리인을 얻을 수 있습니다.
선거 운동은 11 월 30 일 금요일에 끝납니다.
선거 이후 의회의 구성 회의는 12 월 27 일에 열릴 예정이다.
런던은 계략 또는 혼돈 중 하나를 선택합니다.

5 월에 Brexit을 관리 할 수 없다는 점을 감안할 때 다른 국민 투표를 조직하려는 시도는 악과 최악의 선택을 피하는 유일한 방법 일뿐입니다
영국군이 주권을 되찾기로 결정한 지 2 년 후 - 그들은 그것을 잃었을 까? - 그들은 단지 4 개월 반 밖에 걸리지 않았기 때문에 계략이나 혼란 중에서 선택했다.
테레사 수녀가 취한 대안은 브리태츠 정부의 불협화음에 대한 8 번째 사임을 방금 추가 한 것입니다.
그러나 영국이 유럽 연합 (EU)을 떠난 지 투표 한 이후 발생한 상황이기 때문에 상황이 악화 될 수도있다.
사실 나쁜 소식 목록은 끝이 없습니다.
브뤼셀 앞에서 런던이 전시 한 빨간 선이 폭발하거나 경제에 치명적인 불확실성으로 인해 브리짓이 강경 한 브리짓 (Br Brexit)이 부드럽고 정치적으로 불안정한 상태에 이르기까지 두 블록에 걸친 국가의 깊은 분파에서 일어났습니다.
따라서 파운드는 이미 가치의 12 %를 잃었으며 영국은 2019 년과 2020 년에 성장하지 않는 유럽 국가가 될 것입니다.
최악의 소식은 지난 금요일, 교통 장관, 조 존슨 (Jo Johnson)의 사퇴, 즉 매매와 혼돈 사이의 비난 즉, 목소리 나 런던의 투표가없는 단일 시장의 규칙에 대한 제출 사이다. 3 월 29 일에 합의없이 모든 사람이 거부 할 수 있음) 또는 비참한 출발.
부시 대통령과 블레치 총리에게 투표 한 조 존슨 (Jo Johnson) 대통령은 보리스 (Boris)를 사임하고 영국과 EU의 미래 관계를위한 핵심 분야 인 교통에 관한 전략적 협상을 담당했다.
그들의 사임은 재난으로 경계를당한 일부 항공사들에 의해 선언 된 화재에 휘발유를 붓고 그들의 영국 주주들을 괴롭 히고 브뤼셀에 대다수의 공동체 재산을 보여주고 투표권을 없애겠다고 위협했다. EU
Eurotunnel 매니저 인 Getlink에서도 매년 160 만 대의 트럭, 250 만 대의 자동차 및 2 천 1 백만 명의 사람들이 국경 통제없이 교차하고 있습니다.
예를 들어, 도요타의 영국 공장으로 향하는 곳에서 독일의 서스펜션, 스페인 바퀴, 네덜란드 배기관 또는 헝가리 안전 벨트가 매일 통과합니다.
이 트럭의 대부분은 빠른 배송 전자 상거래로 판매되는 8,000 ~ 10,000 개의 패키지를 운송합니다.
3 월 29 일 합의가 없으면 세관 검토 및 관료 절차에서 손실되는 비용과 시간은 어떻게됩니까?
런던이이 과정을 관리 할 능력이 없다는 점을 감안할 때 다른 국민 투표를 조직하려는 시도는 악과 최악의 선택을 피하는 유일한 방법 일뿐입니다.
또는 연장 후 시간 연장을이기는 것.
즉, 미국 시인 앨런 긴스 버그 (Allen Ginsberg)에게 주어진 정리가 이미 성취되었을 때 더 나쁜 소식을 피하기 위해서입니다.
"당신은 이길 수 없습니다.
당신은 묶을 수 없습니다.
그리고 게임을 떠날 수도 없다. "
즉, Brexit을 연기합니다.
플로리다 주 주지사 선거에 대한 재검 표 명령
득표율의 0.5 %에도 미치지 못하는 점이 다시 선거 이후 논쟁의 여지가있는 법적 분쟁의 진원지가된다.
18 년 후, 플로리다는 다시 논쟁의 여지가 많은 분쟁 투표 대회의 진원지가되었습니다.

지난 토요일 상원 의원과 총재 선거는 이번 주 토요일의 선거 결정 권한에 따라 민주당과 공화당 후보들의 투표 차이가 좁아서 자동 투표기를 거쳐야한다.

공화당 후보들은 그 수위에서 앞서 있지만, 그 차이는 득표 수의 0.5 % 미만이다.

플로리다에서 유사한 사태가 발생하여 사 법적으로 결정된 2000 년 대선 때와는 달리 이번에는 투표 용지 설계에 문제가 없다.

그러나 선거 사무실 앞에서 여러 차례의 항의와 많은 신경들 가운데에서 변호사들의 상륙이 반복되어왔다.

플로리다 주 상원의 원인 릭 스콧 (Rick Scott) 상원 의원과 론 데 산티 (Ron DeSantis) 주지사는 화요일 밤에 민주당 다수당을 두 군데 보유하고 있기 때문에 공식 승리자가 아니었지만 승리했다고 주장했다.

결함이 있다고 여겨지는 투표 용지가 실제로 맞으면 이번 금요일에 해당 지구 중 한 곳을 조사했습니다.

화요일 이후, 라이벌, 현재의 상원의 원인 민주당 원 빌 Nelson 및 주지사, Andrew Gillum 후보자와 가진 공화당 원의 투표 다름은 줄어들고있다.

금요일 Scott의 리드는 0.18 %, DeSantis는 0.44 %에 그쳤다.

법으로 토요일 투표의 차이가 0.5 점 미만인 경우 전자 투표 수는 필수입니다.

마지막으로, 플로리다 주 (장관) 선거 당국은 토요일에 정오에 투표 결과의 차이가 자동 재계 표를 요구했다고 공식 발표했다.

결과는 목요일 오후에 준비해야합니다.

이 카운트 후에 차이가 여전히 0.25 포인트보다 작 으면 다른 핸드 카운트가 만들어집니다.

병행하여, 선거는 두 명의 야심 찬 상원 의원의 요구를 통해 사법 화되었습니다.

현재 플로리다 주지사 인 스콧 (Scott)과 도널드 트럼프 (Donald Trump) 미국 대통령은 사기를 저지른 두 군, 대부분 민주당인데도 불구하고 기소했다.

"그것은 불명예 스럽다." 2016 년 대통령 선거에서 민주당 힐러리 클린턴에 반대하여 총 투표 수에서 거짓 부정에 대한 자신의 패배를 돌린 트럼프는 말했다.

Scott은 지난 금요일에 판사가 Broward 카운티의 선거 권한으로 투표 득표에 대한 모든 정보를 넘겨 줄 것을 요구하면서 성공했습니다.

그의 민주당 경쟁자 인 넬슨 (Nelson)은 주지사로 그의 권력을 사용하여 그가 경찰 수사를 요청할 수 있다고 제안했다고 비난했다.

당국은 과거에 논란에 휘말린 상태이지만 사기에 대한 증거가 없다고 말했습니다.

2016 년 선거에서 그는 실제 투표 용지 파일을 삭제했지만 하나는 디지털화했습니다.

브로 워드와 팜 비치는 공화당 조지 부시 대통령과 민주당 앨 고어 대통령 간의 2000 년 대선 이후 분쟁의 핵심 카운티였다.

플로리다는 누가 대통령직을 맡았는지 결정했다.

부시 대통령은 대법원이 부작용 가능성을 발견 한 후 득표 수를 마비시킨 이후 공식적으로 국가를 승리로 이끌었다.

그런 다음 두 카운티는 유권자가 원안에 구멍을 뚫은 투표 용지를 사용했습니다.

이 포맷은 많은 유권자를 혼란스럽게 만들었으며 Gore에게 승리를 요구할 수 있습니다.

애리조나 상원 의원과 조지아 주지사 선거 역시 중요하게 고려된다.

최신 데이터에 따르면 민주당의 키스 텐 시네마 (Keithsten Sinema)는 아리조나의 보수당 자리를 지키고있는 공화당의 마사 맥 샐리 (Martha McSally)에 대해 최소한의 우위를 유지하고있다.

애리조나에서의 민주당 승리는 공화당에 대한 대응의 구상에서 특히 중요 할 것이다.

두 캠페인은 이번 주 금요일 농촌 지역 투표 수의 불규칙성을 분명히하기 위해 합의했습니다.

조지아 주에서 민주당의 스테이시 에이 브람스 (Stacey Abrams)는 모든 득표가 올바르게 집계 될 수 있도록 법적 조치를 취할 것을 약속했습니다.

그 선거는 공화당 출신 후보 인 브라이언 켐프 (Brian Kemp)가 조지아 국무 장관으로 선거 운영에 책임이 있고 민주당이 흑인 인구의 투표를 제한하는 조치를 취했다고 비난했기 때문에 처음부터 논란이 많았다. 미국 최초의 아프리카 계 미국인 총재가 될 아브람에게 열쇠

입법 선거, 트럼프에 대한 국민 투표

미국인들은 화요일에 국회의원과 주 및 지방 당국을 선택합니다.

그러나 선거는 무엇보다도 대통령의 미래를 결정할 "트럼프즘"에 대한 재판이다.

이번 주 누군가가 혼수 상태에서 벗어났다면 2020 년 가을에 그가 깨어나고 도널드 트럼프 (Donald Trump)가 그의 재선을 노리고 있다고 생각할 것입니다.

미국 대통령 화요일에 입법 선거에 투표하지는 않았지만 그의 인물과 그의 정치적 미래는 있습니다.

투표소와의 약속은 그의 대통령 대행의 재판이다. 다른 사람들에게는 악명이 높으며 의심의 여지없이 "트럼프즘"에 대한 국민 투표가 실시된다.

그들의 결과는 또한 위임장의 두 번째 부분에서 기동성을 결정하고 재선거를위한 기반을 마련 할 것입니다.

이번 가을에 트럼프가 그의 선거를 막 제거했다고 모든 것이 설명됩니다.

선거 장면은 거의 2 일에 1 번 제기되었습니다.

선거 당일 9 월 6 일부터 11 월 6 일까지 그는 30 번의 회의를 개최 할 예정이다.

미시시피 주에서의 집회에서 그는 "나는 투표 용지가 아니지만 나는 국민 투표이기도하다.

«내가 투표 한 척해라»그는 추종자들에게 물었다.

성취하기 쉬운 요구 사항입니다.

트럼프는 미국 대통령직에 출마 한 이후 정치 담론을 독점했다. 2015 년 6 월

백악관에서 한 번 공화당 동맹국 인 공화당 동맹국들에 대해 "대통령 주의적 입장"을 채택할 것을 신뢰 한 사람들은 트럼프가 여전히 스스로 선거 운동을하고 있음을 발견했다.

그의 대립 언어, 통상적 인 모욕, 인종 차별주의 및 성 차별주의의 언급, 백악관의 혼돈에 대한 누출이 언론을 점령 해 대통령을 싫어한다.

그의 인물은 또한 양당을 지배합니다.

트럼프의 지원은 공화당 후보들을 두었다. 공화당 후보들은 입후보자의지지를 선언하는 것 외에는 선택의 여지가 없다.

가장 분명한 사례는 테드 크루즈의 것입니다 :

대통령 후보로 그는 2016 년에 트럼프에게 "병적 인 거짓말 쟁이"와 "겁쟁이 족제비"라고 부르며 지난달 텍사스에서 상원 의원 자격을 유지하기 위해 그를 집회에 초대해야했다.

반면에 민주당은 누가 더 "트럼프 반대"인지를 증명하기 위해 경쟁합니다.

투표 용지에 표시되는 이름은 의회 후보자와 주 정부 당국에서 지방 공무원에 이르는 수백 개 주 당국의 이름입니다.

하원은 435 명의 회원을 완전히 갱신하고 상원은 100 명의 입법 원 중 3 분의 1을 선출합니다.

대통령 트럼프의 첫 2 년 동안, 의회의 두 집은 트럼프 충분하지 않은 공화당의 대부분은, 예를 들면 건강 개혁 버락 오바마의 해체 또는 벽 자금으로, 그의 의제의 중심 포인트를 재촉했다 그의 캠페인의 스타 인 멕시코와.

이제 입법 봉쇄에 대한 모든 내용은 입법 후에 훨씬 커질 것입니다.

여론 조사는 민주당 원들이 하원 의원을 되찾을 것이고, 의문점은 최종 우위가 될 것이라고 말했습니다.

상원 의석에서 민주당 전복은 훨씬 어려워진다. 공화당 의원들 대부분이 공화당의 유리한 영역에 있기 때문이다.

실제로, 공화당이 대다수를 강화할 가능성이있다. 공화당은 49 명의 민주당 원에 대해 보수당 의원 (상원 의원 51 명)이 최소한이다.

이러한 결과로 민주당은 트럼프의 정치적 의제를 어지럽 힐 수있는 가능성을 갖게 될 것입니다. 그러나 대통령 당선자를 이끌었던 문제들 중 무엇을 해야할지에 대한 고구마를 발견할 것입니다.

주장 음모 캠페인 트럼프와의 조사를 그의 결과는 오래 가지 않을 것이다.

하원은 대통령 탄핵 또는 탄핵 촉진을 담당합니다.

민주당의 가장 좌파 전류는 대통령의 방어 공화당 유권자를 자극 할 정치적 서커스에서 트럼프의 위임 하반기을 만들 것이라고 경로를 선택합니다.

최근 몇 년 사이에 역사가 보여 주듯이 설문 조사는 잘못 될 수 있습니다.

특히 악의적 인 선거 운동 종료 후, 이민자들의 캐러밴과 폭력 사건의 결과로 트럼프가 쏟아내는 두려움에 대한 캠페인에 동요 됨. - 정치적 원수 폭파, 특히 무엇보다도 시나고그의 학살 피츠버그 - 여론의 일부는 대통령의 공격적인 메시지와 관련이 있습니다.

이 모든 것이 이번 화요일에 예상되는 높은 투표율 뒤에 있는지 확인해야합니다.

부시 대통령의 인기는 그가 버드 대금을 제쳐두고 떠날 때 개선되고 세금 감면 또는 대법원 판사 브렛 카바 너 (Brett Kavanaugh)의 확인과 같은 결과를 얻는다.

캠페인이 끝나면 트럼프가 선택한 경로가 아니며 2016 년 스크립트를 반복하는 것이 좋습니다.

아무도 그 경로가 작동한다고 부인할 수 없습니다.

우크라이나 동부 분리주의 지역에서의 논란이 많은 선거

유럽 연합 (EU)과 미국, 키예프는 선거를 "허구와 불법"이라고 부른다.

OSCE와 대부분의 서방 국가에 대한 비판과 경고에도 불구하고 우크라이나 동부 분리주의 반군 당국 인 도네츠크와 루간 스크는 오늘 각자의 지도자를 선출하기 위해 "선거"를 개최 할 준비를하고있다. 소위 "인민위원회"의 대리인.

유럽 안보 협력기구 (OSCE)는 민스크 평화 협정을 위반 한 결정을 비난 한 마지막 일이다.

이 선거는 유럽 연합, 미국 및 우크라이나 당국에 의해 "허구와 사생아"로 비판 받고 심지어 비난을 받아왔다.

OSCE 의장 겸 이탈리아 외무 장관 인 Enzo Moavero에 따르면 우크라이나 동부의 선거는 "민스크 협약의 서신과 정신에 어긋난다"고 말했다.

우크라이나의 Piotr Poroshenko 대통령은 "러시아는 선거를 막기 위해 영향을 미쳐야하고, 반대 입장을 취한 것으로 평화적 해결책을 홍보하고 싶지 않다는 것을 보여주고있다"고 믿는다.
그의 견해로는 "이 호의 결과는 국제 사회 (...)에 의해 인정되지 않으며 러시아에 대한 새로운 제재의 채택으로 이어질 것"이라고 말했다.
그러나 모스크바에서 그들은 다르게 생각합니다.
드미트리 페스 코프 (Dmitri Peskov) 러시아 대변인은 화요일 분리 주의자들에 의해 조직된 총선은 평화 협정을 위반하지 않는다고 밝혔다.
Peskov에 따르면, "민스크에서 협약을 적용하려는 욕구가 거의없는 사람들은 키예프 당국이다."
2015 년 2 월 12 일 민스크에서 서명 한이 협정은 도네츠크와 루 간츠크에 해당하는 부분에서 러시아와의 국경 통제에 대한 우크라이나로의 회귀와 그에 따른 자유와 민주주의 선거의 두 영역 모두에서의 축하를 고찰한다. 우크라이나 입법.
그 대가로 키에프는 두 영토를 자치 체제로 만들어야합니다.
그러나 상호 불신은 무력 충돌이 산발적으로 발생하는 동안 중립적 인 과정을 유지합니다.
도네츠크의 대통령과 대다수의 투표를 얻으려는 데니스 푸 스틸 린 (Denis Pushilin)은 이달 초 공화당이 "선거를 개최해야한다"고 설명하면서 지도자와 지방 의회가되었다.
그는 투명한 방식으로 모든 국제 표준을 준수 할 것이라고 약속했다. 모스코바 만 믿는 사람은 없다.
우크라이나 보안 서비스 총괄 책임자 인 바실리 그 리삭 (Vasili Gritsak)은 선거 결과에 대한 회의록은 이미 준비가되어 있다고 주장했다.
지난 8 월, 알렉산더 자 자르 첸코 (Alexander Zajárchenko)는 자신이 주장한 도네 치크 인민 공화국 (Donetsk People 's Republic, DNR)의 머리로 4 년 가까이 머물렀다는 점을 아직도 분명히 모르는 공격에서 암살 당했다.
Pushilin은 잠정적으로 DNR을 담당하게됩니다.
이웃 국가들에서도 스스로 자칭 된 LUMR (Lugansk Republic of Lugansk) 선거도 개최되며 유리한 방향으로 이끌어가는 당은 일시적으로 최고 지도자 인 Leonid Pasechnik의 입장을 취하고있다.
그는 불과 1 년 전에 이고르 플롯 니츠키 (Igor Plotnitski)를 대신했고, 그는 어두운 손으로 타격을 입었습니다.
도네츠크와 루간 스크는 러시아가 크림을 합병 한 지 한 달 후인 2014 년 4 월에 우크라이나 정부에 대항하여 무기를 들었다.
모스크바는 분리 주의자들을 지원하기 위해 무기와 돈, 그리고 사람들을 보냄으로써 전쟁을 일으켰다.
그 이후 유엔에 따르면 분쟁으로 1 만 명이 사망했다.
독일과 프랑스의 중재 아래 도달 된 민스크 합의는 우크라이나 군과 반군 민병대 간의 무장 한 대립을 결정적으로 끝내려는 시도였다.
미션 2019 : EU의 불매 운동에 유럽의 기금 사용을 금지
EU기구가 통제 메커니즘을 강화했다는 사실에도 불구하고 Eurosceptic 당사자와 그들의 재단은 EU와 관련된 모든 것을 포기하는 대신 유럽의 기회를 이용하기로 결정했다.
10 월 17 일 수요일에 발표 된 마지막 유로 로바 미터 (Eurobarometer)는 영국에서도 유럽 연합의 지지자들이 Brexit의 그것들을 능가하는 유럽주의의 증가를 보여준다.
그러나 현재 연합에 의해 설계된 마스 트리 히트 조약 발효 25 주년과 유럽 단일 통화 인 유럽이 방금 만났을 때 EU는 다가오는 유럽 선거에서 역사상 가장 큰 도전 중 하나에 직면해있다. :

처음으로 전통적인 중도 좌파와 중도 좌파의 유럽 연합 블록은 50 % 이하로 떨어질 것으로 예상되며 여론 조사원들은 유럽과 유럽의 가치관을 어지럽히도록 노력하는 국민 강대국당이 3 분의 1의 좌석을 차지할 것으로 추정하고있다. 역설적이게도 그들은 유니언 기금을 이용하여 재정을 강화했습니다.
지난 10 년 동안 유럽 연합 기관들은 유럽의 돈으로 학대적인 관행을 다루기 위해 항체를 강화했습니다.
그럼에도 불구하고 Eurosceptic 정당과 정치 기반은 EU와 관련된 모든 것을 보이콧하는 대신 유럽의 기회를 이용하기로 결정했습니다.
유럽 의회 (Jaume Duch)의 유럽 대륙 의회의 대변인은 또한 대중 주의자와 유로 지성인들의 부상에 대처하기 위해 노력한다.
"특정 국가에 특정 금액을 제공하는 것이 아니라 투자 할 정책을 알고있는 것입니다.
유로 바로 메타는 가장 중요한 문제는 고용, 경제 성장, 이주 관리, 기후 변화, 환경 및 사회 보호라고 말했습니다.
EU 대변인에 따르면이 문제는 회원국들이 더 큰 기여를해야한다는 요구를 충족시키지 못하지만 Brexit이 완성되고 유럽 연합에서 벗어나면 국가들은 공통 예산에 더 많은 돈을 쓰려고하지 않는다. 영국과 같은 근본적인 납세자 연합.
"그러나 포퓰리즘은 이러한 문제에 대한 해결책이 될 수 없다"고 그는 경고했다.
전직 도널드 트럼프 컨설턴트 인 스티브 반넌 (Steve Bannon)의 충고와 프랑스의 국가 전선 (National Front) 또는 이탈리아 리그의 성공에 힘 입어 Vistalegre (Vistalegre)가 가득한 이래로 VOX는 선거구 선거구 덕분에 유럽 의회에서 재판을받는 것을 목표로합니다. Podemos가 2014 년에했던 것처럼 공개 노출 수준과 재정 측면에서 중요한 자극이 될 것입니다.
«반유럽 정당이 정부에 들어갈 때 자신들의 언어를 온건하게 비판하는 방법을 알고 싶지만 다른 국가에서는 그 반대가 발생합니다.
이탈리아, 헝가리, 폴란드는 다르다 "고 Elcano Royal Institute에서 발표 한"유럽의 미래 - 수도의 전망 "이라는 오스트리아 지부 편집장 Paul Schmidt가 고려한다.
비록 그 당사자들이 더 영향력이 있지만, 같은 제목의 스페인어 부문 책임자 인 Ignacio Molina (Real Instituto Elcano)는 공동의 입장을 분명히하기가 더 어려울 것이라고 믿는다.
"유로 이탈리즘 당이 하나도 없으며, 감성이있는 스무 명의 사람들이 있습니다.
재미있는 방법으로, 대부분의 Eurosceptic 극단적 인 헝가리 당은 반대로 슬로바키아어이고 반대로 헝가리 인 슬로바키아어이다;
공통점을 두는 것은 그렇게 쉬운 일이 아닙니다. "라고 그는 덧붙입니다.
고정 수입 측면에서, 각 MEP는 월 6,600 유로의 순이익을, 정당화되지 않은 사무 경비는 4,400, 고문은 약 24,000의 급여를받습니다.
또한 각 국회의원마다 월별 소득 (일별 출장)이 다양하고 연중 연봉의 3.5 %에 달하는 퇴직 연금이 있습니다.
"진짜 문제는이 돈을 제어하는 방법이다 :
본회의에 참석하지 않으면 적은 돈을 벌 수 있습니다 :
본회의에서 51 % 미만의 표결에 출석하면 급여 삭감의 절반을 받게됩니다.
그런 다음 새로운 프린터 나 iPad 같은 사무 비용을 위해 고안된 게임이 있지만 통제가 없습니다.
녹색당과 사회 민주당 원들의 MEP들은 정확하게 "나는 200 유로를 지불했다"고 말하지만 그렇게 할 의무는 없다.

게다가 외부 유럽 출신의 많은 강력한 사람들이이 당사자들에게 거대한 보조금을 제공하고 우리는이를 피할 필요가있다 "고 유럽 의회 소식통은 ABC에 말했다.

2001 년 좋은 조약 (Nice Treaty of 2001)의 선언에 따르면, 유럽 연합 정당에 대한 EU 의 기금은 국가 캠페인에 직접 또는 간접적으로 자금을 제공하는 데 사용할 수 없습니다.

쿠루 벤 (KU Leuven) 공공 거버넌스 연구소의 연구원 인 우스터 울프 스 (Wouter Wolfs)는 "유럽의 보조금은 EU 정치 체제를 강화하고 민주적 인 적자를 다루어야하지만 Eurosceptics는 자국의 정부를 지원하기 위해 주로 자원을 사용했다" OECD의 부패와 청렴성 반대 포럼 - 전체 보이콧에서부터 유럽 제도와 선거에 대한 매우 치열한 참여에 이르기까지 20 년 동안 우로 지성 형성인 지하를 강조합니다.

2016 년에, Morten Messerschmidt, 자유주의와 민주주의 (MELD)의 국가 운동을 지원하기 위하여 유럽 기금을 사용한 지금 무효 한 운동의 덴마크 MEP는,이 돈을 돌려 보내야했다.

"마린 르 펜 (Marine Le Pen)을 통해 의회의 서비스는 유럽의 자금을 사용하여 국가 문제를 해결한다는 것을 분명히 보여주었습니다.

이 경우 법안은이 돈을 갚아야한다는 것 "이라고 의회의 유럽 사회당 당원 인 이탈리아 메르세데스 브레소 (Mercedes Bresso) 부회장은 말했다.

독일인 Rainer Wieland (유럽 인민당)와 함께 Bresso는 유럽 차원에서 정당 및 정치 재단 자금 조달에 대한 결의안을 주도했습니다.

부정 방지 사무소 (OLAF)는 이미 이런 경우를 법원에 제기했습니다.

또한, 유럽 예산 관리위원회 (European Committee of Budgetary Control)는 극단적인 권리 그룹이 구 프랑스 국가 전선 (French National Front), 리그 (Italy), 오스트리아 자유당 (FPÖ, 오스트리아 세바스티안 쿠르츠 정부의 파트너)와 자유 당 (네덜란드 PVV)은 2016 년 EU 자금 427 000 유로를 불법적으로 보냈다고합니다.

그들은 고가의 음료 (샴페인), 저녁 식사 및 직원 선물로 사용했다.

유럽 의회의 예산 통제위원회 (CONT)는 EU 국가 및 자유 그룹을위한 미래 기금의 비례 삭감으로부터 자금을 회수 할 것을 권고했다.

"나는 시스템이 항상 일 때문에 유럽 의회는 지금 더 잘 준비라고 생각하지 않습니다.

그것은 사기를 밝혀 냈고 지금은 재판소에 돈을 돌려 줄 것을 요청한 사례가있다 "라고 예산 통제위원회의 Arndt Khon (독일 사회 민주주의 자들의 MEP)의 작업을 준비하는 Laurin Berresheim은 말했다.

이기구는 유럽의 돈이 어떻게 소비되는지를 감시합니다.

기관의 구성원들에 의해 사용되는 방법뿐만 아니라 유럽 프로젝트들에 대해서도 사용됩니다.

Bresso의 조수 인 Gabriel Richard-Molard와이 운동에 또한 참여한 사람은 "10 년 전에 적절한 투명성이나 효과적인 규제 또는 일반적인 금융 방향에 대한 기록이 없었습니다.

우리는 적응해야했습니다. 예, 우리는 이전보다 확실히 준비가 잘되어 있습니다.

Berresheim에게있어 MEP는 돈을 쓰는 법에 대해 모두 동일한 규칙을 가지고 있습니다. “

«그러나 어떤 정치적 색채의 대표자들에게는 규칙이 엄격 해져서는 안된다.

그것은 여전히 민주주의입니다. "

2017 년 9 월 13 일 유럽 집행위원회는 유럽 정당 및 유럽 정치 기금에 관한 법규 및 기금에 관한 규칙을 개정하고 선거 이전에 현행의 2014 년 규정을 개정하고자하는 제안을 채택했습니다 여러 법적 허점을 다루기 위해 2019 년 유럽인.

"제안 된 개정안은 Wieland와 Bresso에 의해 작성된 정당과 유럽 재단의 법령과 기금에 관한 보고서에 따르면,보다 투명성을 제 공하고, 민주적 정당성을 향상시키고, 법률 적용을 강화하는 데 중점을 둡니다.

총회에서 투표를 기다리는이 동의안은 개인이 더 이상 유럽 정당의 정당 등록을 후원 할 수 없다는 것을 확증합니다.

브리핑에서 레이더 위 랜드 (Raidereur Rainer Wieland)는 "현재 자금 조달은 실제 투표 결과와 연결되어 유럽의 정당을 더욱 유럽 적으로 만든다"고 말했다.

Bresso와 Wieland 보고서에 따르면 2018 년 EU 예산 내에서 유럽 정당 보조금으로 324.4 백만 유로, 유럽 정치 재단 보조금으로 1,932 만 유로가 할당되었습니다. .

기증자는 연간 최대 18,000 유로를 기부 할 수 있지만, 기부자의 이름은 3,000 유로 이상의 기부금으로 만 게시되어야합니다.

전문 포털«EU Observer»의 말처럼 유럽 의회에 소속 된 2 개의 초 민족 주의자들은 최소 조건을 충족시키지 못하면 더 이상 EU 기금을받을 자격이 없다.

유럽 연합 운동 연합 (AEMN)과 UKIP가 소재한 평화 자유 연맹 (APF)이 9 월 말에 발표한 정당과 재단을 총괄하는 독립적 인 당국은 유럽의 자금 남용으로 비난 받아 등록 목록에서 삭제되었습니다.

이미 해체 된 후자는 유럽의 돈을 오용하여 유럽 의회에 110 만 유로를 상환했다.

유럽 의회는 2021 년에서 2027 년까지의 새로운 다년간 금융 체제에 대해 논의 중이며 EU 당국이 EU 자금 사용을 통제 할 수있는 경제적 능력을 더 많이 부여 할 수있는 방안이 논의되고있다.

"현재 EU 재무 기금 (DGFIN), 저명한 인물위원회 (Independent Committee of Eminent Persons), 그리고 유럽 정당 및 유럽 정치 재단 (EUPPF) 당국의 EU 기금 사용을 검증하는 3 가지 행위자가있다. ( )

그러나 정치 자금에 대한 유럽 자금의 오용을 평가할 적절한 규제가있는 문제는 별개로 EU 자금 사용 방법을 효율적으로 분석 할 수있는 것이 주요 포인트입니다.

우리는 여기에 인적 자원이 거의 없다 "고 Bresso는 결론 지었다.

미국 선거의 높은 참여율 방지

미국은 화요일 6 차례 중간 선거를 실시한다.

버몬트 주 (州)는 현지 시간으로 오전 5시에 처음으로 투표소 문을 열었습니다.

서로 다른 시간대를 거쳐 동서로 투표가 연속적으로 시작되며 하와이 여론 조사가 밤 11시에 끝나면 선거가 끝납니다.

이 선거는 하원 의원 435 석, 상원 100 석 중 35 석, 상원 50 석 중 36 석을 갱신한다.

미국 국민의 경우 도널드 트럼프 (Donald Trump)의 첫 2 년을 대통령으로 평가할 기회이기 때문에 매우 높은 투표율이 예상됩니다.

경쟁이 치열한 상황에서 전문가들은 8 년 만에 처음으로 민주당이 하원에서 다수를 갱신할 것으로 예상한다. 트럼프는 2020 년 재선을 모색하는데 큰 걸림돌이된다. 공화당 우세

선거법 위반 혐의로 제주 총독을 조사한다.

제주도 원희룡 총재는 공직 선거법을 위반 한 것으로 의심되는 경찰이 의문을 제기했다.

특히 정치인에 대한 비난은 공식 선거 운동 기간을 존중하지 않는 데 초점을두고있다. 선거전의 약속과 비전을 시작하기 전에 제시하고, 중국 자본의 대규모 유입 선거에서 경쟁자였던 정치인이자 전 총재직이었던 제주도의 도시 개발에 참여했다.

심문은 9 시간 동안 지속되었고, 원 총리는 토요일 29시에 아침 3시에 귀국했다.

문재인 회장, 계속지지를 잃다.

대통령 문재인 일행의 Minjoo에 대한 지원의 비율 월 2017 년 이후 가장 낮은 수준으로 떨어졌습니다, 전 대통령 박근혜의 축출 이후 한 달 조기 대통령 선거가 일어났다.
이는 여론 조사 기관인 Realmeter가 8 월 13 일에서 14 일 사이에 거주하는 1,005 명의 한국인을 대상으로 실시한 설문 조사의 결과입니다.
그것에서 대통령은 55.6 %의 대중적인 지원을 등록했다.
그 비율은 행정부에서 지금까지 최저치이다.
Realmeter에 따르면, Realmeter에 따르면, 국민 연금 제도의 개혁과 현 정부의 국민에 대한 신뢰 (무죄로 선언 된 후 감소 함), 남 충의 전 총리 안희정 (An Hee Jung)은 학대에 대한 재판에서 성적인
또한 권력자 인 민주 (Minjoo)는 일주일 전보다 약 3 % 포인트 낮은 37 % 시민 지원으로 지지를 잃었다
설문 조사의 오차 범위는 3.1 % 포인트이고 신뢰 수준은 95 %입니다.
2018 지역 주민이 역사상 2 번째로 큰 참여를 기록합니다.
6 월 13 일의 지방 선거는 전국 선거위원회 임시 자료에 따르면 60.2 %의 참여율을 기록했다.
투표권이있는 42,900,000 명의 시민 중 25,840,000 명이 투표했습니다.
이 비율은 또한 1995 년에 열린 첫 번째 지방 선거의 68.4 % 이후 역사상 두 번째로 높습니다.
이 결과는 선거 이전의 예측과 반대되는 것으로, 북한과 미국 간의 정상 회담과 반대파에 대한 여당의 예상이 이미 우세한 상황에서 상대적으로 낮은 참가율을 예상했다.
그러나 2017 년 5 월 조직 된 박근혜 (朴 槿 惠) 전 (前) 대통령 정부의 해임과 해임과 대통령 선거에 대한 항의 시위에서부터 시작된 정치인들의 관심이 증가한 것으로 분석된다.
지방 선거 투표율이 높았고 조기 투표에서 20 %를 능가했다.
한편, 역사상 처음으로 지역과 동시에 일어난 부분 입법부에서는 참여율이 60.7 %로 4 월 12 일의 53.9 %를 크게 웃돌았다 과거의
선거 재개가 시작됩니다.
제 7 회 지방 선거의 투표 수는 오후 6시에 투표소 폐쇄 후 수요일 13시 6시 30 분에 시작되었습니다.
전국 선거위원회 (National 선거위원회)에 따르면 78000 명의 투표자, 12000 명의 옵서버, 105000 명의 관리자 및 2,500 대의 자동 투표기가 전국 조사를 위해 동원되었습니다.
개인 투표와 우편 투표로 투표소를 닫은 후에 옵서버, 선거 위원 및 경찰관이 호송 한 계산 센터로 이미이 투표 용지를 옮겨 놓은 투표 용지가 이미 옮겨졌습니다. 봉인 된 후 감독하에 카운팅하기 위해.
또한 선거위원회 대표는 투명성을 높이기 위해 회계사들이 기계로 계산 한 투표 수를 확인하고 투표권을 가진 일반 시민들 중에서 선출 된 옵서버가 전체 조사 과정을 감독한다고 밝혔다.
일부 지구에서는 오후 10시 30 분에 시작하는 직책을 누가 차지할 지 알 수 있습니다.