recibido 01-12-2019 revisado 14-01-2020 aceptado 17-01-2020

Can *translationese* features help users select an MT system for post-editing?

¿Pueden ayudar las características del traduccionés *a los usuarios a seleccionar un sistema de TA para posedición?*

Nora Aranberri
IXA Group - University of the Basque Country
nora.aranberri@ehu.eus

Abstract: This work explores the possibility of using translationese features as indicators of machine translation quality for users to select an MT system for post-editing assuming that a lower level of translationese will reveal a reduced need for editing. Results reveal that translationese and automatic metrics rank systems differently, opening an avenue for further research into the information each provides.

Keywords: Machine translation, translationese, quality evaluation

Resumen: Este trabajo explora la posibilidad de utilizar las características del traduccionés como indicadores de calidad de traducción automática para ayudar a los usuarios a seleccionar un sistema de TA para posedición asumiendo que un nivel más bajo de traduccionés revela una menor necesidad de edición. Los resultados apuntan a que el traduccionés y las métricas automáticas clasifican los sistemas de manera diferente, abriendo nuevas vías de investigación sobre la información que aporta cada métrica.

Palabras clave: Traducción automática, traduccionés, evaluación de la calidad

1 Introduction

Translation research has long debated the existence of *translationese*, that is, a set of features common to translated texts that differentiates them from texts originally written in their respective language. Whereas the specific features in which translations and original texts differ is still debated, authors seem to agree that differences exist (Laviosa, 2002), and what is more, that original and translated texts can be automatically distinguished (Baroni and Bernardini, 2005; Volansky, Ordan, and Wintner, 2013). Research suggests that translated texts show signs of reduced richness and seem *'abnormally normal'* (Tirkkonen-Condit, 2002, p. 217). When post-editing machine translation, these translationese features seem to get amplified (Toral, 2019) probably because post-editors are primed by the MT output (Green, Heer, and Manning, 2013), which inherits a significant closeness to the source text and covers as much of the target language the training resources and techniques allow.

While translationese does not intend to refer to poor translated versions that result out of lack of translation skills but rather to features that are intrinsically present in translations and that might even be unavoidable (Tirkkonen-Condit, 2002), when exacerbated, as it seems to be the case when machine translation is involved, the differences between the translation and the original language could indicate a variation in translation quality. To be more precise, we wonder whether translation universals, namely, simplification, normalisation, explicitation and interference (Baker, 1993; Toury, 2012) could provide machine translation users with information about output quality.

So far, in research and professional environments, information about the overall quality has been reported via automatic metrics, and when feasible given the time and cost involved, human evaluations. If proven useful, translationese features would offer an advantage over automatic metrics in that they do not require a reference translation of the segment to be assessed, but rather rely on the difference between the translation of such segment, and standardised features and ratios of the source and target languages.

Recent work on exploring translationese features in post-edited (PE) and human translations has shown evidence that (1) PE

ISSN 1135-5948. DOI 10.26342/2020-64-11

texts are simpler in terms of lexical use, (2) sentence length is closer to the source text length, and (3) grammatical units in PE texts tend to preserve typical sequences of the source language to a larger extent that in human translations (Toral, 2019). This seems to indicate that, as Green, Heer, and Manning (2013) claim, translators are primed by the MT output. This being so, there are grounds to believe that an MT output with a lower level of translationese might benefit users who intend to edit the MT output.

In this work, we aim to analyse whether differences in translationese are present in the output of different MT systems and whether this could be used to help a user select the system that suits a particular text better. To that end, we set the experiment within the current scenario in the Basque Country, where multiple MT systems have been made available to the public in a relative short period of time. We collect the output of the systems and study the relation between translationese features and automatic metrics. Results suggest that they do not always point at the same MT system as the best performing. This outcome opens up a new avenue for further investigation about the type and usefulness of the information provided by translationese features, specially for selecting a system for post-editing.

2 *The Basque context*

According to a market study conducted by Langune, the Basque Association of Language Industries, in the Basque Autonomous Region alone, the translation industry market reached 40 million euros in 2017 with the translation of around 500 million words, when considering the demand of both the public and private sectors [1]. Needless to say that this study did not consider the additional translation needs that are addressed without resorting to professional services. Translations between the two official languages of the region, namely, Basque and Spanish, are commonplace in both work environments and the private sphere.

Publicly available systems are emerging. In less than a year, three different systems have joined Google Translate[2] in offering Spanish-Basque neural machine translations, namely, Modela[3], Batua[4], and the neural system offered by the Basque Government[5][6].

The systems are made available under the assumption that they provide translation quality useful for users. These users might be professional translators that are willing to try the system to see whether a paid subscription or a customised version would be worth investing in, or regular users who find themselves having to translate texts in their daily lives or even professional settings, and given the free nature of the systems, decide to try them out. All systems include warning messages about the potential quality issues of the output and specify that it should be used carefully. However, they do not provide any further indication as to the quality the system offers. Indeed, this is very difficult to evaluate, as a system's performance might vary considerably depending on the text used as input, its syntactic complexity, its topic, its register, that is, how suitable the text is for the particular system to translate given the information used during its training.

From a user's perspective, what is the difference between the systems in terms of quality? Should each user pick a system after attempts of trial and error? Is it possible to provide users with some pointers as to which one to select depending on their needs? This work is a first step towards studying the output of the available systems.

3 *Experimental set-up*

The following subsections describe the data sets used and the MT systems tested in this work.

3.1 MT systems

The MT systems used in this study are five freely available systems, four neural systems and a rule-based system, that can translate from Spanish into Basque. It is worth noting that we are not aware of the data used to train the different systems and that it is possible that part of the data sets used in the experiments were included during training. However, we believe that this does not pose a threat to this study because it is not

[1] http://www.langune.eus

[2] Available at: https://translate.google.com

[3] Available at: https://www.modela.eus/eu/itzultzailea

[4] Available at: https://www.batua.eus/

[5] Available at: http://www.euskadi.eus/itzultzailea/

[6] Note that following the completion of this research, in December 2019, the freely available NMT system *itzultzailea* was launched by Elhuyar.

our objective to discover the systems' absolute translation quality but rather compare the results of translationese features and automatic metrics.

- **Google Translator.** This is the multilingual multidirectional MT system service developed by Google. It is not clear from the documentation available if the queries for Spanish–to–Basque translations are handled by a "zero-shot" neural system, by linking two neural systems, that is, Spanish-English, English-Basque, or with phrase-based statistical systems. The number of words to be translated freely through the web is unlimited but a user may not translate more than 5,000 words at a time, which can be input into the system window or uploaded as a document.
- **Modela.** It is a bidirectional Spanish-Basque neural MT system developed in a research project funded by the Elkartek scheme of the Basque Government and the Basque Business Development Agency during 2016-2017 (Etchegoyhen et al., 2018). The consortium consisted of ISEA (Coordinator of innovation projects for MondragonLingua), the IXA research group of the University of the Basque Country, the Vicomtech research centre, the Ametzaigaina technological agent and the Elhuyar Foundation. In November 2018, the consortium agreed to have a baseline system publicly available and users may translate up to 2,000 words per month free of charge.
- **Batua.** It is a bidirectinal Spanish-Basque neural MT system powered by Vicomtech and sponsored by the telecommunications group Euskaltel and the linguistic services group MondragonLingua. It was released in 2019. The number of words to be translated freely is unlimited but a user may not translate more than 1,000 words at a time, which must be input into the system window. It is stated that the system is in beta version. It provides the user with the possibility to edit and correct the translations, which the system will use to improve the quality of its output.
- **EJ-NMT.** It is a bidirectional Spanish-Basque neural MT system made available by the Department of Culture and Language Policy of the Basque Government, *Eusko Jaurlaritza*, *EJ* for short. The system was developed in collaboration with departments within the EJ, the Basque Network for Science and Technology and other organisations with a strong focus on language, such as the Basque radio and television broadcaster EITB. Additionally, the system has been customised with the translation memories created over the last 20 years at the Basque Institute for Public Administration. It was released on October 16, 2019. The number of words to be translated freely is unlimited but a user may not translate more than 4,000 characters at a time. The text can be typed into the system window or a URL may be provided. The page displays a warning stating that the system is in beta version.
- **EJ-RBMT.** It is a rule-based MT system made available by the Department of Culture and Language Policy of the Basque Government, *EJ*. The system has been running since 2010 powered by Lucy and supports translation between Spanish and Basque, and English and Basque. The number of words to be translated freely is unlimited but a user may not translate more than 1,500 characters at a time. The text can be typed into the system window or a URL may be provided. The page includes the option of displaying multiple translations for ambiguous words.

The use of the five systems will allow us to study whether differences between the systems in terms of translationese are considerable, and also analyse the behaviour of translationese-related metrics in NMT and RBMT systems. Translations for all systems and all data sets were collected over the first two weeks of November 2019.

3.2 Data sets

The data sets to be used must comply with a number of criteria to be adequate for the study. Firstly, we require data sets from different domains in order to check whether systems perform differently depending on the topic. We aimed at using existing publicly available sets whenever possible to facilitate replication. However, due to the limited sets for Basque used in research, we also compiled two new sets. Nonetheless, as can be seen in the description below, the new sets are accurately described to allow easy identification. Secondly, to use automatic metrics, the texts must be available in Spanish and Basque, as a reference translation is necessary.

Linked to this, it was considered important to check if the source text was originally written in Spanish or whether it was already a translation, as this may introduce a degree of translationese in the original text. Similarly, it was considered important to track the original source language used to obtain the Basque translations, as this might, once again, introduce a degree of translationese. Also, it was important that the Basque translation was not a post-edited version, as this would bias the results in favour of the MT system that output a proposal closer to the original system used to create the reference.

When dealing with a minority language such as Basque, finding data sets that meet all the required criteria can prove challenging. We finally opted for the five data sets presented below; three publicly released data sets and two sets specifically created for this study. All five belong to different domains and have been translated into Basque without the aid of an MT system. Unfortunately, the original language of the texts and the source language for the translations was not always traceable.

Given the relatively large size of the corpora and the use restrictions of the MT systems (see subsection 3.1), we extracted 100 segments per set. Please remember that a segment might include, depending on the corpus, one or several sentences.

- **QTLeap IT Corpus.** The QTLeap corpus[7] consists of 4,000 question and answer pairs in the domain of computer and IT troubleshooting for both hardware and software. The text was gathered by an IT support company through their chat support service. As a result, the corpus consists of naturally occurring utterances produced by users while interacting with a service. Both Spanish and Basque are professional translations of the English text, which, in turn, is a professional translation of the original Portuguese text.

- **QTLeap News Corpus.** The QTLeap News corpus[8] is a sample of the News Commentary corpus created as training data resource for the Conference for Statistical Machine Translation Evaluation Campaign[9]. It consists of political and economic news crawled from the Project Syndicate site. The QTLeap News corpus consists of 1,104 sentences made available by the WMT 2012 and 2013 translation tasks, where the Spanish version was distributed (a manual translation of the original English text), and the Basque version was obtained through professional translation of the original English text.

- **TED Talks Corpus.** This data set was released by the 2018 IWSLT Evaluation for one of the two official translation tasks: Low Resource MT of TED talks from Basque to English Speech Translation of lectures, and is available at the Web Inventory Transcribed and Translated Talks[10] (Cettolo, Girardi, and Federico, 2012). We further cleaned the corpus to fix a number of alignment mismatches. The final set available consists of 6,649 parallel sentences. The corpus includes a range of miscellaneous talks given by distinguished experts. According to the TED Talks translation initiative, transcribed talks are translated by volunteers from their original language (mainly English) into the target languages. Therefore, both the Spanish and Basque translations would result from English source texts.

- **GuggenSet.** This set was compiled by extracting the text from the web page of Guggenheim Bilbao. It consists of the texts corresponding to the Essentials of the Collection section[11] where the main authors and their work are described. Therefore, the set belongs to the area of art, and formal and specialised register. Whereas the language of the original text is not specified, given the socio-linguistic context of the region, it is highly likely that this was Spanish and that the translation was carried out from Spanish to Basque by professional translators. This data set consists of 100 sentences.

- **AdminSet.** This set includes an extract of a law passed by the Basque Government.

[7] https://metashare.metanet4u.eu/

[8] https://metashare.metanet4u.eu/

[9] http://www.casmacat.eu/corpus/news-commentary.html

[10] https://wit3.fbk.eu/mt.php?release=2018-01

[11] https://www.guggenheim-bilbao.eus/en/the-collection

Data set	Domain	Source sentences	Avr. sent. length	Source words
QTLeap IT	IT	114	10 (min. 2 – max. 28)	1,249
QTLeap News	news	101	21 (min. 3 – max. 70)	2,127
TED Talks	miscellaneous	149	12 (min. 1 – max. 50)	1,832
GuggenSet	art	100	32 (min. 1 – max. 75)	3,274
AdminSet	law	133	33 (min. 1 – max. 233)	4,474

Table 1: Features of the data sets

Specifically, it contains the first 100 segments (133 sentences) of the Law 10/2019, of June 27, on Territorial Planning of Large Commercial Establishments (1). The set belongs to the area of administration and law, and highly formal and specialised register. The text was originally drafted in Spanish and professionally translated into Basque by the in-house translation service of the Basque Government.

Table 1 shows a summary of the main features of the data sets. As we can see, sets belong to a specific domain (IT, news, miscellaneous, art and law) and include between 100-150 sentences. As expected given their spontaneous nature, the QTLeap IT corpus and the Ted Talks corpus consist of shorter sentences, 10-12 words on average, whereas the specialized GuggenSet and AdminSet consist of considerably longer sentences, 32-33 words on average, with the AdminSet including sentences of up to 233 words. We can see that the total number of words included in the sets varies from 1,249 words for the QTLeap IT corpus, to 4,474 words for the AdminSet, the shortest and longest respectively.

4 Translationese experiments

In this section we compare the system outputs with respect the four translationese features, namely, simplification, normalisation, explicitation and interference, measured in the form of lexical variety, lexical density, length ratio and part-of-speech (PoS) sequence, respectively.

4.1 Lexical variety

Lexical variety indicates how rich the vocabulary used in a text is. As is standard in the field, we measure the difference in type/token ratio (TTR) as an approximation for this feature. In this study, we compare the TTRs of the system outputs. It is assumed that the lower the ratio, the more reduced the vocabulary in the translation is, and therefore, the use of the target language is poorer. This could mean that the translations lack precision and that lexical richness is not fully exploited. Therefore, a user should be inclined to use systems with higher type/token ratios.

A word of caution is in order here. This measure must be taken with caution when MT output is involved as the risk exists that an increased number of types is achieved through the incorrect translation of words.

Results in Table 2 show that the lexical variety in terms of type/token ratio is very similar for all systems, which tend to be within a range of 0.02 points. However, there are two systems that consistently score highest, namely, the EJ-EBMT system and Google Translator.

4.2 Lexical density

Lexical density intends to measure the amount of information present in the text. To that end, we calculate the ratio between the number of content words (nouns, adjectives, verbs and adverbs) and the total number of words in the texts. It is assumed that the higher the number of content word density, the higher the information transferred from the source text will be, which is, in principle, what we aim for.

Results are displayed in Table 2. These reveal that except for the QTLeap IT corpus, where EJ-RBMT scores the highest lexical density, Google Translator obtains the best scores across all data sets.

4.3 Length ratio

MT systems tend to produce sentences of a similar length to the source text because they often lack the capacity to distance the output from the source pattern. When this is the case, the length ratio tends to be low. Again, we should warn that significant differences in sentence length could also be explained by incorrect translation outputs such as incomplete translations, a feature that has been observed in NMT systems in particular.

	Text origin	QTLeap IT	QTLeap News	TED talks	GuggenCorpus	AdminCorpus
Lex. variety	Google	0.47522	**0.63087**	0.51355	**0.62278**	0.47308
	Modela	0.48206	0.61169	0.50295	0.59064	0.40965
	Batua	0.47144	0.61246	0.52962	0.60364	0.48930
	EJ-NMT	0.47203	0.62361	0.52583	0.61040	0.47638
	EJ-RBMT	**0.48866**	0.62801	**0.53732**	0.61320	**0.49466**
	Reference	0.43769	0.61413	0.50996	0.61165	0.45208
Lex. density	Google	0.86486	**0.84899**	**0.79267**	**0.83114**	**0.80833**
	Modela	0.85313	0.81768	0.76180	0.82054	0.76571
	Batua	0.84994	0.83740	0.77645	0.82107	0.79485
	EJ-NMT	0.84340	0.83803	0.76999	0.82438	0.79179
	EJ-RBMT	**0.87010**	0.84088	0.78592	0.82945	0.50605
	Reference	0.81256	0.83715	0.79188	0.82693	0.79800
Length ratio	Google	**0.01069**	0.05998	**0.00318**	0.05436	0.10815
	Modela	0.00345	0.09586	0.01784	0.05831	0.08619
	Batua	0.01908	0.02496	0.05242	0.00128	0.08527
	EJ-NMT	0.01677	0.03920	0.04912	**0.00103**	0.08690
	EJ-RBMT	0.06298	**0.00921**	0.06452	0.03770	**0.05400**
	Reference	0.10919	0.03398	0.00819	0.00383	0.00755
Perplexity	Google	4.75425	**4.91998**	5.22633	4.67304	4.26184
	Modela	**5.0867**	4.89808	**5.39553**	**4.87391**	**5.22066**
	Batua	4.91754	4.45909	5.08883	4.84463	4.73409
	EJ-NMT	4.72987	4.75369	4.92077	4.71945	4.75306
	EJ-RBMT	3.00027	4.4454	4.82782	4.34122	3.32085
	Reference	4.70072	4.61032	5.26604	4.99856	4.63802

Table 2: Results for the translationese features for the different data sets and MT systems where the best results for each data set are shown in bold

We calculate the absolute difference in length (words) between the source text and the translations, normalised by the length of the source text. Results in Table 2 show that for the QTLeap IT corpus and the TED Talks corpus, Google Translator outputs the lowest length ratios, whereas it is EJ-RBMT that obtains the best scores for the QTLeap News corpus and the AdminSet. EJ-NMT, closely followed by Batua, scores best for the GuggenSet.

4.4 PoS sequence

Interference aims to account for the source language patterns that are kept in a translation when these do not necessarily belong in naturally occurring text in the target language. Toral (2019) proposes to observe the PoS patterns as an indicator of interference and calculate the difference between the perplexities of the translation's PoS sequence when compared against a language model of PoS sequences in the source and target languages. As a low perplexity indicates that the translation is similar to the language model, the higher the difference in perplexity between the source and language model, the more the translation will differ from the original language and the closer it will be from the target model.

To train the language models, we first compiled the corpora for Basque and for Spanish, which included general text retrieved from the web, news and administrative texts, of 3 and 3.2 million sentences, respectively. We then used ixa-kat (Otegi et al., 2016) to annotate the Basque corpus and ixa-pipes (Agerri, Bermudez, and Rigau, 2014) to annotate the Spanish corpus. Both tools use the naf specification (Fokkens et al., 2014) for PoS classification, which consists of 9 tags and identify common nouns (N), proper nouns (R), adjectives, (G), verbs (V), prepositions (P), adverbs (A), conjunctions (C), determiners (D) and others (O). Finally, we built the models with Modified Kneser-Ney smoothing and no pruning, considering n-grams up to $n = 6$ but had to assign default parameters to singletons, even when they are not present in the PoS-annotated corpus.

The results in Table 2 show that the translations with the lowest interference from the source are provided by Modela for all data sets, even when other systems are close to its results. Interestingly, the EJ-RBMT output is the system displaying the highest level of similarity toward the source language and

	QTLeap IT		QTLeap News		GuggenCorpus		TED Talks		AdminCorpus	
MT system	**BLEU**	**TER**	**BLEU**	**TER**	**BLEU**	**TER**	**BLEU**	**TER**	**BLEU**	**TER**
Google	**18.75**	**63.05**	12.58	71.90	15.16	67.01	**36.18**	**41.85**	19.05	59.63
Modela	16.09	69.07	13.46	**69.07**	**28.64**	72.53	17.71	67.15	17.20	70.10
Batua	17.55	64.77	**17.33**	69.74	21.93	62.62	29.11	47.36	22.31	58.81
EJ NMT	18.43	64.43	16.60	69.58	22.14	**61.66**	29.11	48.56	**23.39**	**58.36**
EJ RBMT	10.16	82.30	07.46	79.38	10.23	76.15	19.05	59.63	07.50	76.79

Table 3: Automatic metric results

distances most from the naturally occurring pattern of the target language.

5 *Automatic quality experiments*

This section analyses the output of the five MT systems under study when measured by automatic metrics. Specifically, we calculated BLEU and TER scores. Whereas they are both string-based, precision-oriented metrics that compare the MT output against a reference translation, it could be argued that BLEU is more directed to providing an overall quality measurement whereas TER bears more relationship with the editing work required by the output.

Results from Table 3 show interesting trends. We observe that systems tend to follow the same trend in variation when moving from one domain to the next, that is, in general, the quality of the output either increases or worsens for the whole group of systems. We can observe that, on average, systems perform worse on the QTLeap IT and QTLeap News corpora. The AdminSet obtains better results, followed closely by the GuggenSet. Overall, the set with best scores is the TED Talks corpus.

However, there are two aspects to highlight here. Firstly, we can observe that for each domain or data set, it is a specific MT system that stands out as best. For example, Google obtains the highest BLEU score for the QTLeap IT corpus and TED Talks corpus, while Batua performs best for QTLeap News corpus, Modela for the GuggenSet and EJ-NMT for the AdminSet. In turn, the RBMT system lags behind for all domains. Note that the TER scores point at different systems for three out of the five data sets.

Secondly, it is interesting to see that systems not only vary in output quality in unison, they also display BLEU scores within the same ranges for each domain, except at times in the specific sets in which each system stands out or lags behind. The NMT systems revolve around the range of 17.18 BLEU points for the QTLeap IT corpus, around 13-17 for the QTLeap News corpus, around 22-28 for the GuggenSet, around 29-36 for the TED Talks, and around 18-23 for the AdminSet. TER scores, in turn, also remain in ranges that vary from 2 to 11 points.

From the scores obtained, we could conclude that TED texts within the TED Talk style are best translated by Google Translate, administrative texts are best handled by EJ-NMT, formal art-related texts will require fewer edits if translated by EJ-NMT (even when the overall quality might be better with Modela), news are best translated with Batua, and Google Translate performs best with IT-related material. Needless to say that these conclusions must be approached with caution for two main reasons: (1) our test sets are small; and (2) these systems are not static but rather they are being improved and retrained with additional data and by using different techniques.

The limitations mentioned above, however, do not prevent us from comparing automatic metric results and translationese features. Firstly, we observe that it is not the same system that consistently performs best with regards translationese. This raises the question of what the implications for post-editing each of the features are and which could be more or less relevant for users. Further analyses of post-editing effort would be necessary to establish the impact each translationese feature has on MT output use.

Secondly, by comparing the rankings assigned by the different features and metrics, we observe that automatic metrics and translationese do not always point at the same MT system as the best performer. The same question as above emerges here, that is, which type of information is more useful for a user who wants to post-edit the output? A closer look at the rankings proposed by BLEU and TER reveals that, whereas different between them, none of them is in more agreement with translationese results.

Interestingly although not unexpectedly, it is worth noting the presence of the EJ-RBMT system in high positions of the ranking by certain translationese features, whereas automatic metrics consistently show that its output quality is considerably poorer.

6 Conclusions

It is believed that translations display a set of shared features that distinguishes them from texts written in the original language, referred to as Translation Universals, which results in translationese. In this work, we set to explore the possibility of using such features as indicators of MT quality for users to select an MT system for post-editing assuming that a lower level of translationese will reveal a reduced need for editing. To that end, we compared the results obtained from translationese features and automatic metrics for five data sets and MT systems. Whereas further experiments using large data sets and variations of the approaches to measure the features should be performed to gather conclusive data, and contrast these results with user post-editing performance, results seem to indicate that the two sets of metrics rank systems differently, opening an avenue for research into the information each provides.

Acknowledgements

This research was partially supported by the Spanish MEIC and MCIU (UnsupNMT TIN2017-91692-EXP and DOMINO PGC2018-102041-B-I00, co-funded by EU FEDER), and the BigKnowledge project (BBVA foundation grant 2018).

Bibliography

Agerri, R., J. Bermudez, and G. Rigau. 2014. Ixa pipeline: Efficient and ready to use multilingual nlp tools. In *LREC*, volume 2014, pages 3823–3828.

Baker, M. 1993. Corpus linguistics and translation studies: Implications and applications. *Text and technology: In honour of John Sinclair*, 233:250.

Baroni, M. and S. Bernardini. 2005. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274.

Cettolo, M., C. Girardi, and M. Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the EAMT*, pages 261–268, Trento, Italy, May.

Etchegoyhen, T., E. Martínez Garcia, A. Azpeitia, G. Labaka, I. Alegria, I. Cortes Etxabe, A. Jauregi Carrera, I. Ellakuria Santos, M. Martin, and E. Calonge. 2018. Neural machine translation of basque. In *Proceedings of the 21st Annual Conference of the EAMT, 28-30 May, Alacant, Spain*, pages 139–148.

Fokkens, A., A. Soroa, Z. Beloki, N. Ockeloen, G. Rigau, W. R. Van Hage, and P. Vossen. 2014. Naf and gaf: Linking linguistic annotations. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 9–16.

Green, S., J. Heer, and C. D. Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 439–448. ACM.

Laviosa, S. 2002. *Corpus-based translation studies: theory, findings, applications*, volume 17. Rodopi.

Otegi, A., N. Ezeiza, I. Goenaga, and G. Labaka. 2016. A Modular Chain of NLP Tools for Basque. In *Proceedings of the 19th International Conference of Text, Speech, and Dialogue, Brno, Czech Republic, September 12-16.* pages 93–100.

Tirkkonen-Condit, S. 2002. Translationese - a myth or an empirical fact? a study into the linguistic identifiability of translated language. *Target. International Journal of Translation Studies*, 14(2):207–220.

Toral, A. 2019. Post-editese: an exacerbated translationese. In *Proceedings of MT Summit XVII, 19-23 August, Dublin, Ireland*, pages 273–281.

Toury, G. 2012. *Descriptive translation studies and beyond: Revised edition*, volume 100. John Benjamins Publishing.

Volansky, V., N. Ordan, and S. Wintner. 2013. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.