Proceedings of ParlaCLARIN III @ LREC2022 pages 101–106
Marseille, 20 June 2022

Comparing Formulaic Language in Human and Machine Translation: Insight from a Parliamentary Corpus

Yves Bestgen
Laboratoire d'analyse statistique des textes
Université catholique de Louvain
Place Cardinal Mercier, 10 1348 Louvain-la-Neuve, Belgium
yves.bestgen@uclouvain.be

Abstract

A recent study has shown that, compared to human translations, neural machine translations contain more strongly-associated formulaic sequences made of relatively high-frequency words, but far less strongly-associated formulaic sequences made of relatively rare words. These results were obtained on the basis of translations of quality newspaper articles in which human translations can be thought to be not very literal. The present study attempts to replicate this research using a parliamentary corpus. The results confirm the observations on the news corpus, but the differences are less strong. They suggest that the use of text genres that usually result in more literal translations, such as parliamentary corpora, might be preferable when comparing human and machine translations.

Keywords: neural machine translation, human translation, parliamentary corpus, multiword unit

1. Introduction

Due to the success of neural machine translation systems, more and more research is being conducted to compare their translations to human translations. Most of these studies take a global view of quality (Popel et al., 2020; Läubli et al., 2018; Wu et al., 2016), but others focus on much more specific dimensions that could be further improved, such as lexical diversity and textual cohesion (De Clercq et al., 2021; Vanmassenhove et al., 2019).

In a recent study, Bestgen (2021) analyzed the frequency of use of a specific category of formulaic sequences, the "habitually occurring lexical combinations" (Laufer and Waldman, 2011), which are statistically typical of the language because they are observed "with markedly high frequency, relative to the component words" (Baldwin and Kim, 2010). To identify them, he used the CollGram technique which relies on two lexical association indices: mutual information (MI) and t-score, calculated on the basis of the frequencies in a reference corpus (Bernardini, 2007; Bestgen and Granger, 2014; Durrant and Schmitt, 2009). A discussion of two automatic procedures that at least partially implement this technique is given in Bestgen (2019). He showed that neural machine translations contain more strongly-associated formulaic sequences made of relatively high-frequency words, identified by the t-score, such as *you know*, *out of* or *more than*, but far less strongly-associated formulaic sequences made of relatively rare words, identified by the MI, such as *self-fulfilling prophecy*, *sparsely populated* or *sunnier climes*.

These observations can be linked with a series of studies that have shown similar differences in foreign language learning (Bestgen and Granger, 2014; Durrant and Schmitt, 2009) and which have proposed to interpret them in the framework of the usage-based model of language learning which "hold that a major determining force in the acquisition of formulas is the frequency of occurrence and co-occurrence of linguistic forms in the input" (Durrant and Schmitt, 2009). It is obviously tempting to use the same explanation for differences in translation, as neural models also seem to be affected by frequency of use (Koehn and Knowles, 2017; Li et al., 2020).

A competing explanation is however possible. All the texts analyzed in Bestgen (2021) were quality newspaper articles written in French and published in *Le Monde diplomatique*, and then translated in English for one of its international editions. However, as Ponomarenko (2019) pointed out following Bielsa and Bassnett (2009), "Translation of news implies a higher degree of re-writing and re-telling than in any other type of translation" (p.40) and "International news, however, tends to prefer domestication of information instead of translation accuracy, which also has its particular reasons" (p.35). As it is important in this kind of texts that the translated version is as relevant and interesting as possible for the target readers, often from another culture, lexical and syntactic modifications, deletions and additions are frequent. All these modifications make the translation of this kind of texts less literal than the one expected from a machine translation system and thus risk to affect the differences in the use of formulaic sequences.

In this context, parliamentary corpora are a perfectly justified point of comparison. Translation accuracy is the main objective. For example, the criteria that European parliamentary debates translations must meet include: "the delivered target text is complete (no omissions nor additions are permitted)" and "the target text is a faithful, accurate and consistent translation of the

News corpus

Original French

Au-delà du logiciel libre. Le temps des biens communs. Jusqu'où ira le droit de propriété?
Pour tout ce qui peut être représenté par de l'information, son extension semble ne devoir connaître aucune limite.

Human translation

The internet and common goods. Own or share? Is there any limit to property?
Current developments might suggest not, as property rights are steadily extended.

Machine translation (DeepL)

Beyond Free Software. The time of the commons. How far will property rights go?
For everything that can be represented by information, its extension seems to know no limits.

Parliamentary corpus

Original French

Nous savons que la Commission va bientôt réformer la politique commune de la pêche.
Cette proposition de règlement du Conseil est la première d' une série qui permettra d' élaborer cette réforme.

Human translation

We know that the Commission is going to reform the common fisheries policy shortly.
This proposal for a Council Regulation is the first in a series which will make it possible to draw up this reform.

Machine translation (DeepL)

We know that the Commission will soon reform the Common Fisheries Policy.
This proposal for a Council Regulation is the first of a series that will allow this reform to be developed.

Figure 1: Human and machine translation of a French excerpt from each of the two corpora.

source text" (Sosoni, 2011).

Figure 1 illustrates this difference in translation between these two types of texts. It shows a brief extract from each corpus in its original and translated versions. The excerpt from the news corpus shows the different translation strategies used by the human and the machine, with the human version showing several reformulations affecting both the lexicon and the syntax and the deletion of one constituent in the last sentence. Such differences are not present in the extract from the parliamentary corpus.

The objective of the present study is to determine whether machine translations of parliamentary texts differ from human translations in the use of phraseology, in order to confirm or refute the findings from the news corpus. The three following hypotheses are tested: compared to human translations, machine translations will contain more strongly-associated collocations made of high-frequency words, less strongly-associated formulaic sequences made of rare words and thus a larger ratio between these two indices. An positive conclusion, in addition to confirming the links between machine translation and foreign language learning, will suggest a way to make machine translations more similar to human translations, especially since the fully automatic nature of the analysis facilitates its large-scale use.

2. Method

2.1. Parliamentary Corpus

The material for this study is taken from the Europarl corpus v7 (Koehn, 2005) (Philipp Koehn, 2012) available at `https://www.statmt.org/europarl`. In order to have parallel texts of which the original is in French and the translation in English, information rarely directly provided in the Europarl corpus because it was not developed for this purpose (Cartoni and Meyer, 2012; Islam and Mehler, 2012), I employed the preprocessed version, freely available at `https://zenodo.org/record/1066474#.WnnEM3wiHcs`, obtained by means of the EuroparlExtract toolkit (Ustaszewski, 2019) (Ustaszewski, Michael, 2017).

Two hundred texts of 3,500 to 4,500 characters, for a total of 120,000 words, were randomly selected among all texts written in French and translated into English. These thresholds were set in order to have texts long enough for collocation analysis, but not exceeding the 5,000-character limit imposed by the automatic translators used so that the document could be translated in a single operation.

Between February 14 and 16, 2022, three neural machine translation systems were used to translate these texts into English: the online version of *DeepL* (`https://deepl.com/translator`) and *Google Translate* (`https://translate.google.com`) and the *Office 365* version of *Microsoft Translator*.

2.2. Procedure

All contiguous word pairs (bigrams) were extracted from the CLAWS7 tokenized version (Rayson, 1991) of each translated text. Bigrams, not including a proper noun, that could be found in the British National Corpus (BNC, `https://www.natcorp.ox.ac.uk`), were assigned an MI and a t-score on the basis of the frequencies in this reference corpus. Bigrams with $MI \geq 5$ or with $t \geq 6$ were deemed to be highly collocational (Bestgen, 2018; Durrant and Schmitt, 2009). On this basis, three GollGram indices were calculated for each translated text: the percentage of highly collocational bigrams for MI, the same percentage for the t-

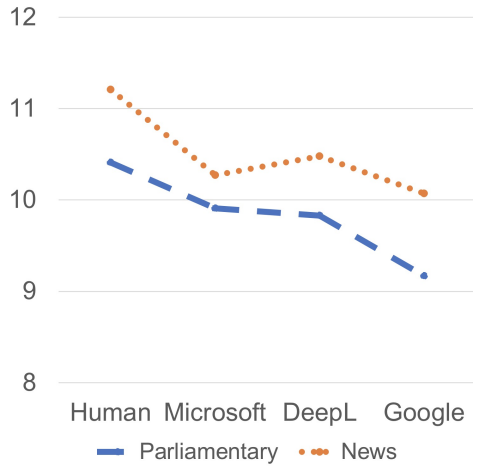

Figure 2: Mean percentages of highly collocational bigrams for MI.

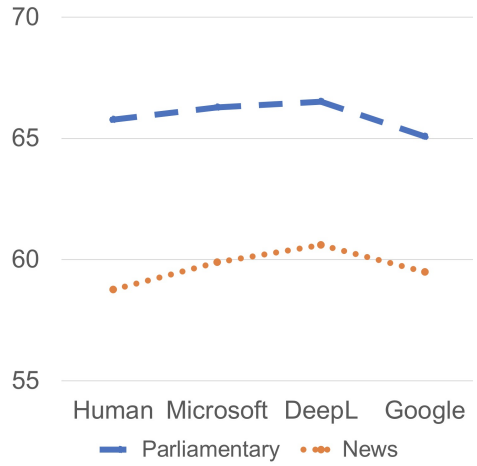

Figure 3: Mean percentages of highly collocational bigrams for t-score.

score and the ratio between these two percentages (%t-score / %MI).

This procedure is in all points identical to the one used in the analysis of the news corpus (Bestgen, 2021). The only difference is that the machine translations of the news corpus were undertaken in March-April 2021. According to Porte (2013) terminology, the present study is thus an *approximate* replication that "might help us generalize, for example, the findings from the original study to a new population, setting, or modality" (p.11).

3. Results

3.1. Parliamentary Corpus

The mean percentages of highly collocational bigrams for the MI and t-score and the mean ratio for the four translation type of the two genres of text are shown in Figures 2 to 4. The differences observed on the parliamentary corpus are very similar to those obtained with the news corpus. This section is focused on the anal-

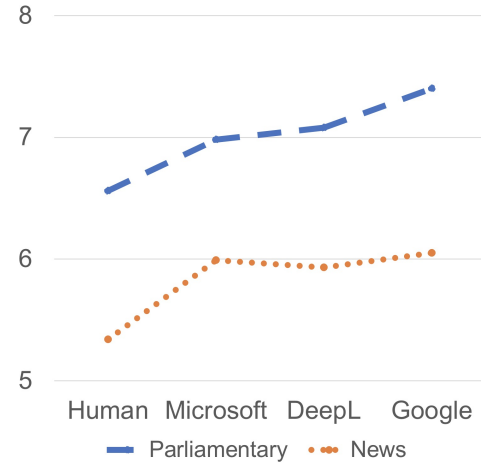

Figure 4: Mean ratio between t-score and MI.

		MI	Micro.	DeepL	Google
			MI		
Human	Di		-0.50	-0.59	-1.24
	d		0.38	0.44	0.89
	p		0.67	0.69	0.81
Micro.	Di			-0.09	-0.74
	d			0.11	0.67
	p			0.55	0.76
DeepL	Di				-0.65
	d				0.61
	p				0.71
			t-score		
Human	Di		0.51	0.76	-0.69
	d		0.15	0.31	0.26
	p		0.60	0.60	0.63
Micro.	Di			0.25	-1.20
	s			0.09	0.45
	p			0.55	0.80
DeepL	Di				-1.44
	d				0.72
	p				0.76
			Ratio		
Human	Di		0.42	0.52	0.84
	d		0.39	0.50	0.77
	p		0.69	0.70	0.81
Micro.	Di			0.10	0.42
	d			0.14	0.47
	p			0.56	0.71
DeepL	Di				0.31
	d				0.37
	p				0.66

Table 1: Differences (column translator minus row translator) and effect sizes for the two indices and the ratio in the four translation types. Di = Difference, d = Cohen's d, p = proportion of texts in which the mean effect is observed.

ysis of the Parliamentary corpus while the comparison with the news corpus is presented in the next section.

Table 1 presents the differences between the mean scores for every pairs of translators for the parliamentary corpus. The Student's t-test for non-independent measures was used to determine whether these mean differences were statistically significant. Due to the large number of tests performed (18), the Bonferroni procedure was used to compensate for the inflated Type I error rates, with the threshold for an alpha of 0.05 (two-tailed) set at 0.0027 (two-tailed). These tests indicate that all differences are significant, except for those between *Microsoft Translator* and *DeepL* for the three indices and the difference for the t-score between the human and *Microsoft* translations.

Table 1 also gives two effect sizes. Cohen's d informs about the size of the difference between the means as a function of its variability. It is usual to consider that a d of 0.20 indicates a small effect size, a d of 0.50 a medium effect and a d of 0.80 a large effect (Cohen, 1988). The second effect size is the proportion of texts for which the difference between the two translations has the same sign as the mean difference. The maximum value of 1.0 means that texts produced by a translator always have larger scores than those translated by the other translator while the minimum value of 0.50 indicates no difference for this measure between the two translators.

As these results show, the three hypotheses about the differences between human and machine translations are all confirmed by statistically significant differences, except for the difference in t-score between human and Microsoft translations, which is nevertheless in the right direction.

In these analyses, Cohen's d for MI and for the ratio are often medium and sometimes even large and the differences are present in a large proportion of texts. For the t-score on the other hand, all effect sizes are small. This could be explained by the fact that the collocations highlighted by this lexical association measure are mostly very frequent in the language and thus more easily learned by automatic systems (Koehn and Knowles, 2017; Li et al., 2020).

The comparison of the texts translated by *Microsoft Translator* and by *DeepL* does not show, as indicated above, any statistically significant difference and the effect sizes are very small. The outputs of *Google Translate* on the other hand contain fewer highly collocational bigrams for MI and for t-score than these two other machine translators, potentially suggesting less efficiency of this translator for collocation processing.

3.2. Comparison between the Two Genres

As shown in Figures 2 to 4, all the observed trends are very similar in the two corpora, indicating that the two genres of text lead to the same conclusions. However, there is a strong contrast in the mean values. The MI scores are lower in the parliamentary corpus while the t-scores are higher. This is the case for both human and machine translations. This observation is most probably explained by a difference between the text genres, a difference that should already be present in the original French texts.

The comparison of the effect sizes between the two types of translation (Table 2 in Bestgen (2021)) clearly indicates that the differences are much stronger in the news corpus. This observation is consistent with the hypothesis of a greater literalness of the human translations in the parliamentary corpus.

4. Conclusion

The analyses carried out on the parliamentary corpus have made it possible to replicate the conclusions obtained with a news corpus. The observed trends are very similar, but the differences are less strong in the parliamentary corpus. This observation seems to confirm the usefulness of the parliamentary genre for the comparison of human and machine translation. Indeed, one can think that the less literal nature of the translations in news (Ponomarenko, 2019; Sosoni, 2011) favors the identification of differences between the two types of translations. The differences observed in the parliamentary corpus thus seem to be more directly related to the translators effectiveness rather than to other factors.

Among the directions for future research, there is certainly the analysis of translations from English to French, but also between other languages. Here again, the Europarl corpus, as well as parliamentary debates in multilingual countries, allows for a great deal of experimentation. A potential difficulty is that the approach used requires a reference corpus in the target language. This problem does not seem too serious since it has been shown that freely available Wacky corpora (Baroni et al., 2009) can be used without altering the results of the CollGram technique (Bestgen, 2016). Confirming these results in other languages, but also in other genres of texts, would allow to take advantage of them to try to improve machine translation systems. A unanswered question is whether identical results would be obtained on the basis of a genre-specific reference corpus, which is very easy to obtain for European parliamentary debates. This corpus would be composed of documents originally produced in English. It would also be interesting to use other approaches to phraseology, especially more qualitative ones, to confirm the conclusions. Finally, the difference in mean values between the two text genres justifies an analysis of the original texts with the same technique.

5. Acknowledgements

The author is a Research Associate of the Fonds de la Recherche Scientifique (F.R.S-FNRS).

6. Bibliographical References

Baldwin, T. and Kim, S. N. (2010). Multiword expressions. In Nitin Indurkhya et al., editors, *Handbook of*

Natural Language Processing, pages 267–292. CRC Press.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43:209–226.

Bernardini, S. (2007). Collocations in translated language. combining parallel, comparable and reference corpora. In *Proceedings of the Corpus Linguistics Conference*, pages 1–16. Lancaster University.

Bestgen, Y. and Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, 26:28–41.

Bestgen, Y. (2016). Evaluation automatique de textes. Validation interne et externe d'indices phraséologiques pour l'évaluation automatique de textes rédigés en anglais langue étrangère. *Traitement automatique des langues*, 57(3):91–115.

Bestgen, Y. (2018). Normalisation en traduction : analyse automatique des collocations dans des corpus. *Des mots aux actes*, 7:459–469.

Bestgen, Y. (2019). Evaluation de textes en anglais langue ètrangère et séries phraséologiques : comparaison de deux procédures automatiques librement accessibles. *Revue française de linguistique appliquée*, 24:81–94.

Bestgen, Y. (2021). Using CollGram to compare formulaic language in human and machine translation. In *Proceedings of the Translation and Interpreting Technology Online Conference*, pages 174–180, Held Online, July. INCOMA Ltd.

Bielsa, E. and Bassnett, S. (2009). *Translation in Global News*. Routledge.

Cartoni, B. and Meyer, T. (2012). Extracting directional and comparable corpora from a multilingual corpus for translation studies. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2132–2137, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Erlbaum.

De Clercq, O., De Sutter, G., Loock, R., Cappelle, B., and Plevoets, K. (2021). Uncovering Machine Translationese Using Corpus Analysis Techniques to Distinguish between Original and Machine-Translated French. *Translation Quarterly*, 101:21–45.

Durrant, P. and Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching*, 47:157–177.

Islam, Z. and Mehler, A. (2012). Customization of the Europarl corpus for translation studies. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 2505–2510. European Language Resources Association (ELRA).

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Läubli, S., Sennrich, R., and Volk, M. (2018). Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium, October-November. Association for Computational Linguistics.

Laufer, B. and Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning*, 61:647–672.

Li, M., Roller, S., Kulikov, I., Welleck, S., Boureau, Y.-L., Cho, K., and Weston, J. (2020). Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728. Association for Computational Linguistics.

Ponomarenko, L. (2019). *Translating identities in multilingual news*. Ph.D. thesis, Universitat Autonoma de Barcelona.

Popel, M., Tomkova, M., Tomek, J., Kaiser, L., Uszkoreit, J., Bojar, O., and Zabokrtsky, Z. (2020). Transforming machine translation: A deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11:1–15.

Porte, G. (2013). Who needs replication? *CALICO Journal*, 30:10–15.

Rayson, P. (1991). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison.* Ph.D. thesis, Lancaster University.

Sosoni, V. (2011). Training translators to work for the EU institutions: luxury or necessity? *The Journal of Specialised Translation*, 16:77–108.

Ustaszewski, M. (2019). Optimising the Europarl corpus for translation studies with the Europarl Extract toolkit. *Perspectives*, 27:107–123.

Vanmassenhove, E., Shterionov, D., and Way, A. (2019). Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland, August. European Association for Machine Translation.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Łukasz Kaiser, 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. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation.

7. Language Resource References

Philipp Koehn. (2012). *European Parliament Proceedings Parallel Corpus 1996-2011*. available at https://www.statmt.org/europarl/.

Ustaszewski, Michael. (2017). *EuroparlExtract - Directional Parallel Corpora Extracted from the European Parliament Proceedings Parallel Corpus*. Zenodo, available at https://doi.org/10.5281/zenodo.1066474.