Using CollGram to Compare Formulaic Language in Human and Machine Translation*

Yves Bestgen[1][0000−0001−7407−7797]

Université catholique de Louvain, 10 place Cardinal Mercier, Louvain-la-Neuve, 1348, Belgium yves.bestgen@uclouvain.be

Abstract. A comparison of formulaic sequences in human and neural machine translation of quality newspaper articles shows that neural machine translations contain less lower-frequency, but strongly-associated formulaic sequences, and more high-frequency formulaic sequences. These differences were statistically significant and the effect sizes were almost always medium or large. These observations can be related to the differences between second language learners of various levels and between translated and untranslated texts. The comparison between the neural machine translation systems indicates that some systems produce more formulaic sequences of both types than other systems.

Keywords: Neural machine translation · Multiword unit · Lexical association indice.

1 Introduction

Neural machine translation (NMT) systems are currently considered to bridge the gap between human and machine translation [22, 26]. However, little research has been done to determine whether NMT systems are also very effective in processing multiword units [20, 27], whereas the importance of preformed units in language use is now well established, including in foreign language learning and translation [1, 21, 24]. The present study addresses this issue by comparing formulaic language in human and neural machine translation. It focuses on a specific category of multiword units, the "habitually occurring lexical combinations" [17], such as *dramatic increase*, *depend on*, *out of*, which are not necessarily semantically non-compositional, but are considered statistically typical of the language because they occur "with markedly high frequency, relative to the component words or alternative phrasings of the same expression" [2]. These formulaic sequences (FSs) are analyzed by means of a technique proposed by [4], improved by [13] and automated by [10] under the name *CollGram*[1].

* The author is a Research Associate of the Fonds de la Recherche Scientifique - FNRS (Fédération Wallonie Bruxelles de Belgique). He would like to warmly thank Sylviane Granger and Maïté Dupont for access to the PLECI corpus.

[1] As one reviewer pointed out, this automation is not an "easily available plug-and-play implementation". However, there is a freely available system that implements it

Translation and Interpreting Technology Online, pages 174–180,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_020

CollGram rests on two lexical association indices that measure the strength of attraction between the words that compose a bigram, mutual information (MI) and t-score [14], calculated on the basis of the frequencies of occurrences in a reference corpus [10, 13]. These two indices are complementary, MI favoring lower-frequency, but strongly-associated, FSs such as *self-fulfilling prophecy*, *sparsely populated* or *sunnier climes* while the t-score favors high-frequency bigrams such as *you know*, *out of* or *more than*. A series of studies have shown that, compared to native speakers, English as a foreign language learners tend to underuse collocations with high MI scores, while overusing those with high t-scores and that exactly the same differences are observed between advanced learners and intermediate learners [10, 13]. These observations are in agreement with usage-based models 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" [13]. It is worth noting that the same differences were observed between translated and untranslated texts, but the proposed explanation relies on a tendency towards normalization in translation [5, 8]. Since neural models also seem to be affected by frequency of use [15, 19], the hypothesis tested in the present study is that the same effects could be observed when comparing human translations (HTs) and NMTs, namely that NMTs will underuse high MI FSs and overuse high t-score FSs.

2 Method

2.1 Translation Corpus

The texts used are taken from the journalistic section of the PLECI corpus (uclouvain.be/en/research-institutes/ilc/cecl/pleci.html). It is a sentence-aligned translation corpus of quality newspaper articles written in French and published in *Le Monde diplomatique* and in English in one of the international editions of this same newspaper. Two hundred and seventy-nine texts, published between 2005 and 2012, were used for a total of 570,000 words in the original version and of 500,000 words in the translation.

All original texts were translated into English by three well-known NMT systems: DeepL (deepl.com/translator), Google Translate (translate.google.com) and Microsoft Translator (microsoft.com/translator). Online translators were used for the first two, while the version available in *Office 365* was used for the third. All these translations were performed between March 24 and April 6, 2021.

(http://collgram.pja.edu.pl) [18, 25]. Some of the indices used can also be easily obtained with the TAALES software [16] which allows the automatic analysis of many other lexical indices. TAALES presents however an important limitation because it only takes into account bigrams that occur at least 51 times in the reference corpus [9], a value much too high for the MI at the heart of the CollGram approach.

2.2 Procedure

Each translated text was tokenized and POS-tagged by CLAWS7 [23] and all bigrams were extracted. Punctuation marks and any character sequences that did not correspond to a word interrupted the bigram extraction. Each bigram, which did not include any proper name or number according to CLAWS, was then searched for in the 100 million word British National Corpus (BNC[2], www.natcorp.ox.ac.uk). When it is present, the corresponding MI and t-score were used to decide whether it is highly collocational or not. Based on [8] and [13], bigrams with a score greater than or equal to 5 for the MI and 6 for the t-score were considered highly collocational. The last step consisted in calculating, for each text and for each association index, the percentage that the bigrams considered as highly collocational represent compared to the total number of bigrams present in the text.

3 Analyses and Results

Table 1 shows the average percentages of highly collocational bigrams for the MI and t-score in the four type of translations. The four means for each measure of association were analyzed using the Student's test for repeated measures since the same texts, which are the unit of analysis, were translated by the four translators. All these comparisons were statistically significant ($p < 0.0001$).

Table 1. Average percentages of highly collocational bigrams for the two indices in the four translation types.

Measure	Human	DeepL	Google	Microsoft
High MI	11.21	10.48	10.07	10.27
High t-score	58.76	60.60	59.49	59.89

Table 2 presents the differences between the means as well as two effect sizes. The first is Cohen's d, which expresses the size of the difference between the two means as a function of the score variability. According to [11], a d of 0.50 indicates a medium effect and that a d of 0.80 a large effect. The second effect size indicates the percentage of texts for which the difference between the two translations has the same sign as the mean difference. A value of 100 means that all texts produced by a given translator have a higher score than those translated by the other one and a value of 50 means that there is no difference.

As shown in these tables, both hypotheses are verified. Compared to HTs, texts translated by the three neural systems contain a significantly smaller percentage of highly collocational bigrams for the MI and a larger percentage

[2] [6] showed that CollGram produces the same results if another reference corpus, such as COCA (corpus.byu.edu/coca) or WaCKy [3], is used.

of highly collocational bigrams for the t-score. Cohen's *d*s are almost always medium or large and the percentages of texts for which differences are observed are greater than 70% except in one case.

Table 2. Differences (row translator minus column translator) and effect sizes for the two indices in the four translation types.

	Human			DeepL			Google		
	D	Es	%	D	Es	%	D	Es	%
	High MI								
DeepL	-0.72	0.59	73.48						
Google	-1.14	1.00	84.33	-0.41	0.65	74.91			
Microsoft	-0.94	0.77	81.00	-0.21	0.35	62.72	0.20	0.36	64.52
	High t-score								
DeepL	1.83	0.84	80.65						
Google	0.72	0.32	62.37	-1.11	0.98	84.59			
Microsoft	1.13	0.51	71.33	-0.70	0.62	70.97	-0.41	0.42	69.18

An analysis of the passages in which the differences between HT and NMT are the largest suggests that the origin lies at least partially in the less literal nature of human translations (see Table 3 for an example).

Table 3. Example of the four translation types and percentages of highly collocational bigrams for the two indices.

Type	Phrase	%High MI	%High t-score
Original	A raison de huit heures par jour		
Human	In an eight-hour day	67	33
DeepL	At eight hours a day	25	100
Google & Microsoft	At the rate of eight hours a day	14	100

The differences between the three NMT systems are smaller, but still statistically significant. However, they require a different interpretation. When NMTs are compared to HTs, the patterns of differences are reversed according to the MI or the t-score, as expected. For the NMT systems, these patterns are identical for both types of collocation. The average percentages of highly collocational bigrams (see Table 1) are always higher in texts translated by DeepL than in those translated by Microsoft and also higher in the latter than in those translated by Google. Only a detailed qualitative analysis could determine whether these results indicate a difference in effectiveness.

4 Discussion and Conclusion

The reported analyses confirm the hypotheses and thus suggest that, compared to HTs, NMTs more closely resemble texts written by intermediate learners than by advanced learners of English as a foreign language, a result that could be interpreted in the context of a usage-based model of language learning [13]. The NMTs also resemble translated texts more than untranslated texts, but it is not clear that this can be explained by a normalization process. Statistically significant differences, but smaller in terms of effect size, were also observed between the three NMT systems.

It is important to keep in mind that the present study only considers global quantitative properties of MWUs. At no point is the appropriateness in context of the MWUs assessed. It is therefore a very partial approach. However, it has the advantage of not requiring a human qualitative evaluation that is often complicated and cumbersome to set up. Moreover, it is likely that the appropriateness of a MWU is much more important for non-compositional expressions than for the habitually occurring lexical combinations studied here [12].

Another important feature of the approach is that it relies on a native reference corpus to identify highly collocational bigrams for both indices. As already mentioned, research on foreign language learning, but also on the comparison of translated and untranslated texts, has shown that the use of other large reference corpora such as COCA or WaCKy [3] did not change the results [5, 6]. One can also wonder whether the use of a comparable reference corpus, rather than a generic one, would have returned different results. In the case of the comparison of translated and untranslated texts, [5] observed that the use of a journalistic corpus, the *Corpus Est Républicain* (115 million words) made available by the Centre National de Ressources Textuelles et Lexicales, produced differences similar to those obtained with the WaCKy corpus.

Before considering taking advantage of these observations to try to improve NMT systems, a series of complementary analyses must be conducted. Indeed, this study has many limitations, such as focusing only on a subcategory of MWUs [20], on a single language pair, and on a single genre of texts. Moreover, a thorough qualitative analysis is essential to better understand the results and evaluate the proposed explanations. As it has been shown in foreign language learning [7], it would also be interesting to verify that the observed effects are not explained by differences in single-word lexical richness. Finally, the differences between the three NMT systems also require further analysis.

References

1. Baker, M.: Patterns of idiomaticity in translated vs. non-translated text. Belgian Journal of Linguistics **21**, 11–21 (2007)
2. Baldwin, T., Kim, S.N.: Multiword expressions. In: Indurkhya, N., Damerau, F.J. (eds.) Handbook of Natural Language Processing, pp. 267–292. CRC Press (2010)
3. Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E.: The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. Language Resources and Evaluation **43**, 209–226 (2009)

4. Bernardini, S.: Collocations in translated language. combining parallel, comparable and reference corpora. In: Proceedings of the Corpus Linguistics Conference. pp. 1–16. Lancaster University (2007)
5. Bestgen, Y., Granger, S.: Collocation et traduction. analyse automatique au moyen d'indices d'association. In: Kauffer, M., Keromnes, Y. (eds.) Theorie und Empirie in der Phraseologie / Approches théoriques et empiriques en phraséologie, pp. 101–113. Stauffenburg Verlag (2019)
6. Bestgen, Y.: 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 (2016)
7. Bestgen, Y.: Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. System **69**, 65–78 (2017)
8. Bestgen, Y.: Normalisation en traduction : analyse automatique des collocations dans des corpus. Des mots aux actes **7**, 459–469 (2018)
9. Bestgen, Y.: 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 (2019)
10. Bestgen, Y., Granger, S.: Quantifying the development of phraseological competence in L2 English writing: An automated approach. Journal of Second Language Writing **26**, 28–41 (2014)
11. Cohen, J.: Statistical power analysis for the behavioral sciences. Erlbaum (1988)
12. Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., Todirascu, A.: Survey: Multiword expression processing: A Survey. Computational Linguistics **43**(4), 837–892 (dec 2017)
13. Durrant, P., Schmitt, N.: To what extent do native and non-native writers make use of collocations? International Review of Applied Linguistics in Language Teaching **47**, 157–177 (2009)
14. Evert, S.: Corpora and collocations. In: Lüdeling, A., Kytö, M. (eds.) Corpus Linguistics. An International Handbook, pp. 1211–1248. Mouton de Gruyter (2009)
15. Koehn, P., Knowles, R.: Six challenges for neural machine translation (2017)
16. Kyle, K., Crossley, S., Berger, C.: The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. Behavior Research Methods **50**, 1030–1046 (2018). https://doi.org/10.3758/s13428-017-0924-4
17. Laufer, B., Waldman, T.: Verb-noun collocations in second language writing: A corpus analysis of learners' English. Language Learning **61**, 647–672 (2011)
18. Lenko-Szymanska, A., Wolk, A.: A corpus-based analysis of the development of phraseological competence in EFL learners using the collgram profile. In: Paper presented at the 7th Conference of the Formulaic Language Research Network (FLaRN) (2016)
19. Li, M., Roller, S., Kulikov, I., Welleck, S., Boureau, Y.L., Cho, K., Weston, J.: Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4715–4728. Association for Computational Linguistics (2020)
20. Monti, J., Seretan, V., Pastor, G.C., Mitkov, R.: Multiword units in machine translation and translation technology. In: Mitkov, R., Monti, J., Pastor, G.C., Seretan, V. (eds.) Multiword Units in Machine Translation and Translation Technology, pp. 2–37. John Benjamins (2018)
21. Pawley, A., Syder, F.H.: Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In: Richards, J.C., Schmidt, R.W. (eds.) Language and Communication. Longman (1983)

22. Popel, M., Tomkova, M., Tomek, J., Kaiser, L., Uszkoreit, J., Bojar, O., Zabokrtsky, Z.: Transforming machine translation: A deep learning system reaches news translation quality comparable to human professionals. Nature Communications **11**, 1–15 (2020). https://doi.org/10.1038/s41467-020-18073-9
23. Rayson, P.: Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. Ph.D. thesis, Lancaster University (1991)
24. Sinclair, J.: Corpus, Concordance, Collocation. Oxford University Press (1991)
25. Wołk, K., Wołk, A., Marasek, K.: Unsupervised tool for quantification of progress in L2 english phraseological. In: Proceedings of the 2017 Federated Conference on Computer Science and Information Systems. pp. 383–388 (2017)
26. 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., Dean, J.: Google's neural machine translation system: Bridging the gap between human and machine translation (2016)
27. Zaninello, A., Birch, A.: Multiword expression aware neural machine translation. In: Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020). pp. 3816–3825. European Language Resources Association (ELRA) (2020)