Multi-word Units in Machine Translation: why the Tip of the Iceberg Remains Problematic – and a Tentative Corpus-driven Solution

Jean-Pierre Colson

Université catholique de Louvain
Louvain School of Translation and Interpreting,
place Cardinal Mercier 14, Louvain-la-Neuve, Belgium
jean-pierre.colson@uclouvain.be

Abstract. Neural machine translation (NMT) has recently made considerable progress in the improvement of the quality of the produced texts. Among the new features of NMT are the fluidity of the translations and their successful treatment of multi-word units. In this paper we first report the results of a global and automated evaluation of the percentage of phraseology in the translations produced by Google Translate and DeepL. A corpus-based approach makes it possible to estimate that both NMT systems succeed in producing an average percentage of phraseology that is quite natural and may sometimes even be higher than in natural language production by native speakers. Taking a closer look at some problematic cases, however, reveals that the phraseological value of NMT systems may be deceptive, as they are often inable to cope with contextual complexity and low-frequency idioms.

Keywords: Neural Machine Translation, Phraseology

1   Introduction: Lingering Doubts about Neural Machine Translation

Neural machine translation (NMT) is now generally considered as a major breakthrough in machine translation (MT). As pointed out by Loock [1], the growing success of NMT since 2015 is now such that in 2018, and for the first time, more than half of all translation companies and services in Europe reported using MT to some extent. In particular, the results obtained by DeepL with NMT have impressed the general public (as shown by the many newspaper articles on DeepL's web page), as well as language professionals and researchers.

According to DeepL's official Press Release, their system outperforms all existing ones, both on the basis of human evaluation as against the Bleu score [2]. Google Translate, however, has also been switching from statistical machine translation (SMT) to NMT and is reporting better and better results, with a much larger number of language pairs.

145

Apart from DeepL's own references, a limited number of studies have been devoted to an in-depth evaluation of the system, and of NMT as a whole. An important theoretical remark should be made in that respect. As NMT and MT in general have been improving their results in recent years, they have attracted not only researchers in NLP, but also language professionals and translators in particular. However, communication between these two worlds is difficult, because the norms for quality checking differ very much between for instance researchers in NMT and professional translators. It is therefore crucial that the evaluation of the performance of NMT should be carried out in various ways. In the past, MT algorithms were mainly evaluated by means of automated metrics such as the BLEU score [2] or the Meteor score [3]. In recent years, those metrics have been completed by other techniques including human evaluation and the recourse to a 'Challenge Set', a selection of difficult linguistic constructions or structures submitted to the MT systems. Thus, Isabelle et al. [4] propose a set of 108 short English sentences that are used as a challenge set for measuring the efficiency of MT systems. The results were checked by 3 human evaluators. Interestingly, Pierre Isabelle published on the Web an addendum to the paper, in which the authors submitted the challenge set to the latest version of DeepL. It is striking to note that DeepL turned out to perform much better (with an overall correction score of 84 %) than the other systems.

This enthusiasm for the progress made by NMT in general, and DeepL in particular, is not really shared by Loock [1], who uses a methodology derived from Corpus-based Translation Studies [5]. The study compares a corpus of original French with a corpus of translations from English into French, carried out by 2 NMT systems: DeepL on the one hand, and the system used by the European Commission, MT@EC/eTranslation on the other. The results clearly show that, in spite of the apparent fluidity reached by NMT in the target texts, they systematically display a number of flaws as compared to original texts. Among the linguistic structures targeted by the study (such as the frequency of the French word for "thing"), both systems display an abnormal overrepresentation. This makes the end product, the target text produced by the two NMT systems, very poor from the point of view of translation professionals, and it therefore requires a lot of post-editing. In other words, the target texts produced by NMT may be artificial, because they miss some key features, especially idiomaticity, that make them natural in the eyes of native speakers.

As all evaluation methods only address part of this complex issue, our point of view is that a wide variety of approaches should be used in the study of the performance of NMT systems, which will also contribute to a better understanding of the future needs of post-editing. In this paper, two research questions address this problem from the point of view of multi-word units or phraseology [6]:

1. What is the rough proportion of phraseology in texts produced by NMT, as compared to original texts in the target language, belonging to the same register?

2. Does phraseology still pose a major problem to NMT, and if so why?

2   Phraseology and NMT: an Experiment

As regards the first research question mentioned at the end of the preceding section, it should first be pointed out that there is no standard measure of phraseology in a text. Phraseology as defined by Burger et al. (2007) [6] can be considered in the broad sense, including weakly idiomatic constructions (such as lexical collocations, e.g. *harsh criticism* and grammatical collocations, e.g. *again and again*) or in the strict sense (with very idiomatic constructions such idioms, e.g. *spill the beans*). Formulaic language as defined by Wray [7] includes formulas in the broadest sense, many of which have a communicative function (e.g. *What time is it*?). Finally, construction grammar or CxG [8] [9] [10] considers that language as a whole consists of *constructions* in the sense of (partly) arbitrary pairings of form and meaning, at various degrees of abstraction and schematicity.

A construction may be a word in the traditional sense (e.g. *book*), a bound morpheme (*pre-, -ing*), an idiom (*spill the beans, take the rough with the smooth*), a partially filled idiom (*take X into account*), but also an abstract construction such as the ditransitive construction or the passive. At the intersection of phraseology (in the broad sense) and CxG, many discontinuous and partly idiomatic constructions (such as *the more… the more*, *for me to* VERB CLAUSE BE *the work of an instant / moment*) are particularly difficult to extract by means of automated methods.

Extracting phraseology (in the broad sense) from running text is a particularly daunting challenge, as demonstrated by the results of the Parseme 2018 shared task for the extraction of verbal multiword expressions. The best overall system (TRAVERSAL) taking part in the shared task reached, across the 19 languages, an F1-measure (token-based) of 59.67 percent, but a closer look at individual languages and specific categories of multi-word units reveals a number of weaknesses. Thus, for the English dataset (and in spite of the obviously immense linguistic resources for English), TRAVERSAL only reaches a general F1-score (token-based) of 30.15 percent (with recall as low as 20.33). Besides, the very central category of verbal idioms (as in *paint the town red*) receives with TRAVERSAL an even more disappointing F1-score of 4.22 (with precision at 41.67 but with recall at only 2.22). Of all the English verbal idioms that should have been extracted from the running texts, the best overall system was just able to extract correctly 2.22 percent of them. The very inspiring Parseme 2018 shared task illustrates again how difficult it is to establish a gold set of multi-word units (in this case by two human annotators), and also to derive very idiomatic structures from sparse data.

We have proposed [11] [12] the *cpr-score* for measuring the association strength between n-grams of size 2 to 12. It is an adaptation of metric clusters, based on the average distance between the component grams of an n-gram, measured in large corpora of at least 200,000 tokens. This simple metric makes it possible to extract phraseology in the broad sense, formulaic language or idiomatic constructions, with an acceptable precision and recall.

It is particularly difficult to measure precision and recall for a complete phraseology extraction method in a running text, because it is very hard to establish the gold set of constructions that have to be extracted. As explained in Colson [12], a reliable way of

implementing a general extraction method for phraseology is to test it against Chinese segmentation, as Chinese constructions make less strict differences between (Western) notions of words, collocations or idioms.

The precision and recall scores for segmenting Chinese with the *cpr-score* were measured in Colson [12] and reached an F-measure of 0.70 with one of the standard Chinese datasets. Recall is particularly high with the *cpr-score*: 0.749 for Chinese segmentation, which is almost the average degree of agreement for segmentation between Chinese native speakers (0.75).

The *cpr-score* has also been tested against the English dataset of the Parseme 2018 shared task for the extraction of verbal multiword expressions [13]. It should be noted that the gold set for this task was established by just 2 linguists and displayed a number of debatable cases. For precision, the results yielded by the *cpr-score* were less good than those produced by the best system for English: TRAPACC, based on neural networks [14], but the recall score (token-based) was much better with cpr: 0.5225, as opposed to 0.2898 with TRAPACC. These results indicate that there is still room for improvement in any extraction method, but they also highlight the limitation of deep learning (DL). Focusing on recurrent patterns in the training set, DL has a limited recall capacity, while a method such as cpr, based on huge linguistic data, does not require any training set and is able to reach high recall scores. The lower precision scores obtained with cpr at Parseme also derive from the fact that a much broader view of phraseology is taken, as opposed to the more restricted view of the annotators and the often debatable decisions they made. This partly explains why the *cpr-score* reaches much higher precision and recall scores for Chinese segmentation, as the average agreement among native speakers for segmentation is higher than for multiword units / phraseology / formulaic language / idiomatic constructions, all notions that are largely unknown to the average native speaker.

The *cpr-score* has been implemented in a web application: the *IdiomSearch* tool. As explained in Colson [11], the tool may always be improved, but it can be used for raising the phraseological awareness of language students. It also enables translators to get a general idea of the most frequent multiword units in the source text or in the target text. Experiments with *IdiomSearch* have indicated that this methodology makes it possible to extract phraseology in the broad sense and formulaic language, providing much richer results than those produced by manual extraction on the basis of reference books and stricter criteria. Thus, Dupal [15] showed that *IdiomSearch* confirmed results obtained manually for the difference between CLIL (*Content and Language Integrated Learning*) and non-CLIL students, while providing far more instances of phraseology and formulaic language.

In spite of its shortcomings as an ongoing project, the *IdiomSearch* tool, based on the *cpr-score*, offers, for any given text, a neutral and global view of the presence of phraseology in the broad sense / formulaic language / idiomatic constructions.

We have therefore used the *IdiomSearch* tool for finding an answer to the first research question mentioned above, namely the difference in phraseology between texts produced by NMT on the one hand and original texts on the other.

For our experiment, we checked by means of the *IdiomSearch* tool the average percentage of phraseology (in the broadest sense, including formulaic language and idiomatic constructions) in:

- 10 original articles from English newspapers, on the general topic of Brexit;
- 10 French newspaper articles on Brexit, translated into English by 2 NMT systems (Google Translate and DeepL).

The complete list of the chosen articles can be found in the Appendix at the end of this paper. The texts had to be of a non-specialised nature, and the length of each paper was about the same: 5,000 characters, including blanks, for the French articles. As translation from French into English results in a smaller number of words, the comparable original English articles had to be slightly shorter, and their length was therefore set at 4,500 characters; if necessary, longer texts were slighty shortened.

Table 1 below displays the average percentage of phraseology, as measured by *IdiomSearch*, for the original English newspaper articles. Table 2 shows the same percentage for the 10 French articles translated into English by Google Translate and by DeepL.

Table 1. Average percentage of phraseology in 10 English newspaper articles

English Newspaper Articles	
List	% Phraseology
Article 01	34.28
Article 02	32.22
Article 03	45.39
Article 04	41.41
Article 05	54.67
Article 06	46.65
Article 07	49.06
Article 08	43.73
Article 09	42.94
Article 10	55.23
Average	**44.56**

As shown by Table 1, the average percentage of phraseology, measured by *IdiomSearch*, varies a little from one article to the other, in spite of the fact that the number of characters per article was strictly the same (4,500) and the general topic identical. These figures suggest a phraseology range from 30 to 55 percent, with just 2 articles above 50 percent. We should take into account the fact that the *cpr-score* displays a margin of error, does not extract all phraseological units, and is based on the presence or absence of those units in corpora. We may then consider that these figures roughly confirm John Sinclair's idiom principle [16], according to which about 50 percent of any text consists of phraseology in the broad sense.

Table 2. Average percentage of phraseology with Google Translate and DeepL (translations from French into English)

List	% Phraseology	
English Translations of French Newspaper Articles		
	Google Translate	DeepL
Article 01	48.46	47.84
Article 02	47.12	46.62
Article 03	48.07	45.32
Article 04	41.97	40.08
Article 05	55.06	53.73
Article 06	42.15	45.01
Article 07	48.64	46.59
Article 08	39.04	38.06
Article 09	41.10	43.96
Article 10	47.85	48.27
Average	**45.95**	**45.55**

The most striking result shown by a comparison between Table 1 and Table 2 is that there is (slightly) more phraseology in the 10 English translations of French articles, produced by both Google Translate and by DeepL, than in 10 original English newspaper articles on the same subject.

This interesting finding confirms that NMT pays a lot of attention to phraseology, as all recurrent sequences of n-grams in multilingual corpora are trained by the models. Table 2 further indicates that, contrary to what might have been expected, there is no significant difference in phraseology (as measured by our methodology) between the translations by Google on the one hand, and by DeepL on the other. Google Translate even seems to be slightly more phraseological than DeepL, with just 3 translations of DeepL being more phraseological than the corresponding translations by Google.

The answer to our first research question, on the basis of this experiment, is pretty clear: there is (at least) as much phraseology in the English translations of French texts produced by both Google Translate and by DeepL, as in original English texts belonging to the same domain.

This is therefore an indication of the progress made by NMT with respect to the thorny issue of phraseology: translations produced by Google Translate and by DeepL are so fluid and natural, that the phraseological differences with original texts written by native speakers tend to disappear.

In seeking an answer to our second research question, however, we wish to take a closer look at some problematic cases.

3   Phraseology and NMT: a closer look at problematic examples

As illustrated by the preceding section, phraseology as a whole may be pretty rich in the translations produced by NMT systems, but this is no guarantee that those translations will in the first place fulfil their main role: conveying the correct meaning of the source text.

In order to address our second research question (*Does phraseology still pose a major problem to NMT, and if so why?*), we therefore adopted a more classical case-study methodology. There is indeed (yet) no reliable automated way of checking the semantic adequacy between a source text and a target text.

Barreiro et al. [17] showed that Google Translate was wrong in the translation of phraseology in about 40 percent of the cases. Since then, NMT was introduced into the system, and competing systems such as DeepL have been developed. However, translation practice reveals that there remain many errors in the texts produced by NMT. A survey of all cases of wrong translations, or an extensive overview of the performance of NMT across languages falls beyond the scope of the present contribution.

In addition to automated methodologies as illustrated in the preceding section, we agree with Loock [1] that a careful manual analysis of examples that remain problematic for MT is of the essence. In this case, we have randomly selected from the news a number of examples of sentences with rich phraseology, which posed a real problem to both Google Translate and DeepL. The crux of the matter is to determine what are the common features between those very problematic sentences, and what conclusions could be drawn from them.

Example 1: *UK car industry in brace position ahead of Brexit deadline* (The Guardian, 9 August 2019).

Google Translation (English-French): *L'industrie automobile britannique en bonne voie pour la fin du Brexit.*

DeepL Translation (English-French): *L'industrie automobile britannique en position de force avant l'échéance de Brexit.*

Example 1 is an interesting case, because both NMT translations are rich in phraseology, as they contain two nice French idioms (*être en bonne voie, être en position de force*). However, both translations are a complete misinterpretation of the source text: the English sentence means that the UK car industry is fearing the worst (and placing itself in a defensive position, as people protecting themselves against an airplane crash). The two French translations, on the contrary, convey the opposite meaning: the UK car industry is right on track (*en bonne voie*) or in a position of power (*en position de force*). Clearly, the NMT systems were unable to train sufficient contexts of the phrase *in brace position*, so that they could not distinguish between positive and negative connotation.

Example 2: *How Trump's economic chickens are finally coming home to roost* (Forbes.com, 13 February 2019).

Google Translation (English-French): *Comment les poulets économiques de Trump sont enfin rentrés à la maison.*

DeepL Translation (English-French): *Comment les poulets économiques de Trump rentrent enfin au bercail.*

In Example 2, we notice, as in example 1, a rich phraseology in both French translations (especially for the very idiomatic expression *rentrer au bercail*, to return home). However, both NMT systems have completely missed the idiomatic meaning of the phrase *the chickens have come home to roost:* some day you have to pay for your past mistakes. The literal translation of this phrase does not make any sense in French. It should be noted that the canonical form of the phrase, *the chickens have come home to roost*, is known to Linguee, the parallel corpora database underlying DeepL, and receives a number of correct French translations (*payer les pots cassés, refaire surface*). What is apparently problematic here is the number of figurative contexts, as well as the variants of the phrase (in this case, a modifying adjective before *chickens*, the adverb *finally* and the present progressive instead of present perfect).

Example 3: *Zwietracht bei der öffentlichprivaten Partnerschaft. Da haben wir gegenüber dem Ballungsraum keine Chance und fallen immer durch den Rost* (Parseme 2018 German dataset, id = . . newscrawl-19).

Google Translation (German-English): *Discord in the public-private partnership. Since we have no chance against the metropolitan area and always fall through the rust.*

DeepL Translation (German-English): *Disagreement in the public-private partnership. We have no chance against the conurbation and always fall through the rust.*

In Example 3, both NMT systems were unable to detect the German idiom *durch den Rost fallen* (be discarded), and therefore produced a wrong literal translation.

Example 4: *Max Buset dacht dat ik er het bijltje bij neer zou leggen", zegt Eyskens daarover in zijn memoires. "Maar dat was buiten de waard gerekend! Na het herhaalde weigeren van de socialisten en de liberalen besloot ik zonder meer een minderheidsregering te vormen."* (hln.be, 8 December 2018).

Google Translation (Dutch-English): *Max Buset thought that I would give up, "Eyskens says about this in his memoirs. "But that was not counted! After the repeated refusal of the Socialists and the Liberals, I decided to form a minority government."*

DeepL Translation (Dutch-English): *Max Buset thought I'd put my foot down," says Eyskens in his memoirs. "But that was out of the question! After the repeated refusals of the socialists and the liberals, I decided to form a minority government.".*

In Example 4, the Dutch verbal idiom (*buiten de waard rekenen*, which means to make a miscalculation, to overlook sg) was wrongly translated by Google (by means of a literal verb) and also by DeepL. In the case of DeepL, the machine translation is again phraseological (*to be out of the question*) but the meaning is not correct: the sentence means that people who thought that minister Eyskens would call it a day were wrong in thinking so; the sentence does not mean that it was out of the question for him to resign. The Dutch idiom *buiten de waard rekenen* has been ignored (Google) or wrongly interpreted (DeepL) by the training systems of NMT.

Example 5: 相反地，他第一次在大学辍学是为了报读自己更有兴趣的室内设计，第二次辍学则是因为他成功为第一套房子做了室内设计，加上无法适应理工学院的学习环境，所以决定破釜沉舟放手一搏。[*Xiāngfǎn dì, tā dì yī cì zài dàxué chuòxué shì wèile bào dú zìjǐ gèng yǒu xìngqù de shìnèi shèjì, dì èr cì chuòxué zé shì yīnwèi tā chénggōng wèi dì yī tào fángzi zuòle shìnèi shèjì, jiā shàng wúfǎ shìyìng*

lǐgōng xuéyuàn de xuéxí huánjìng, suǒyǐ juédìng pòfǔchénzhōu fàngshǒu yī bó.] (zaobao.com, 30 July 2019).

Google Translation: *On the contrary, he first dropped out of college to enroll in the interior design that he was more interested in. The second time he dropped out of school was because he successfully designed the interior of the first house, and could not adapt to the learning environment of the Polytechnic. So, I decided to let go and let go.*

In Example 5, the last sentence is wrongly translated by NMT (in this case, only Google because DeepL does not offer Chinese). In addition to the wrong pronoun (*I* instead of *he*), the Chinese phrase 破釜沉舟 [*pò fǔ chén zhōu*] (a case of chéngyǔ, a sort of idiomatic expression linked to Chinese culture), literally "break cauldrons sink boats", means *to cross the Rubicon, to make a big decision*. The correct translation could have been "*so he decided to take a big step and call it a day*".

It cannot be over-emphasized that, in cases such as Examples 1 to 5, NMT often creates the illusion of phraseologically correct translations, while being completely wrong. Far from being exceptions, such examples are easy to find in any domain, and in all languages. It may be pointed out that NMT is constantly being improved, and that the models will get even better in the future.

While this remains an open question, we would like to report the results obtained by our corpus-based approach in cases such as Examples 1-5. Contrary to what might be inferred from the poor results of NMT in those cases, the phraseological units used in the examples can be easily extracted from any large corpus by means of the *Idiom-Search* tool and the *cpr-score* mentioned in section 2, as shown in table 3.

Table 3. Association scores and frequency for the phrases in Examples 1-5

Association scores for phrases in Examples 1-5		
List	cpr	Freq
brace position	0.25	10
the chickens have come home to roost	0.73	8
durch den Rost fallen	0.81	29
buiten de waard gerekend	0.73	41
破釜沉舟[pò fǔ chén zhōu]	1.00	45.55

As shown in table 3, the problematic phrases that were wrongly translated by the NMT systems under Examples 1-5 could be easily identified by the *cpr-score* and large corpora of about 1 billion tokens. The question then arises why such cases remain problematic to NMT.

In addition to the results obtained with the Parseme 2018 shared task mentioned in section 2, our data suggest that neural networks have difficulties in learning from training data multiword units with a low frequency. Future work might therefore consider a hybrid approach, in which the research carried out within the framework of computational and corpus-based phraseology is taken into account.

4   Conclusion

While current NMT systems reach an impressive percentage of phraseology in the translations produced, the tip of the phraseological iceberg remains problematic for them. Our examples show that complex contextual features are hard to cope with in the NMT models, and that very idiomatic but fairly infrequent combinations may fall through the cracks.

A combination of NMT and a corpus-based approach may in the future shed new light on the feasibility of solving most phraseological issues. In the meantime, the methodology presented here for checking the global phraseological percentage, as well as the phraseological character of specific combinations, may be used in the manual post-editing by professional translators.

References

1. Loock, R.: Traduction automatique et usage linguistique : une analyse de traductions anglais-français réunies en corpus. Meta, Journal des traducteurs / Translators' Journal, 63(3):786-806. (2018) https://doi.org/10.7202/1060173ar.
2. Papineni, K., Roukos, S., Ward, T., et al. Bleu: a method for automatic evaluation of machine translation. In Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, pp 311-318 (2002)
3. Denkowski, M., Lavie, A.: Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In Proceedings of the EACL 2014 Workshop on Statistical Machine Translation, pp 376-380 (2014)
4. Isabelle, P., Cherry, C., Foster, G.: A Challenge Set Approach to Evaluating Machine Translation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp 2486–2496 (2017)
5. Laviosa, S.: Corpus-Based Translation Studies: Theory, Findings, Applications. Rodopi, Amsterdam/New York (2002)
6. Burger, H., Dobrovol'skij, D., Kühn, P., Norrick, N.: (eds.) Phraseologie / Phraseology. Ein internationales Handbuch der zeitgenössischen Forschung / An International Handbook of Contemporary Research. De Gruyter, Berlin / New York (2007)
7. Wray, A.: Formulaic Language: Pushing the Boundaries. Oxford University Press, Oxford (2008)
8. Croft, W.: Radical Construction Grammar: Syntactic Theory in Typological Perspective. Oxford University Press, Oxford (2001)
9. Goldberg, A.: Constructions at Work. Oxford University Press, Oxford (2006)
10. Hoffmann, T., Trousdale, G.: (eds.) The Oxford Handbook of Construction Grammar. Oxford University Press, Oxford/NewYork (2013)
11. Colson, J.-P.: The IdiomSearch Experiment: Extracting Phraseology from a Probabilistic Network of Constructions. In Mitkov R (ed.), Computational and Corpus-based phraseology, Lecture Notes in Artificial Intelligence 10596. Springer International Publishing, Cham, pp 16-28 (2017)
12. Colson, J.-P.: From Chinese Word Segmentation to Extraction of Constructions: Two Sides of the Same Algorithmic Coin. In Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), ACL, pp 41-50 (2018)

13. Savary, A. Ramisch, C., Hwang, J.D., Schneider, N., Andresen, M., Pradhan, S., Petruck, M.R.L.: (eds.) Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018). ACL, Santa Fe (2018)
14. Stodden, R., QasemiZadeh, B., Kallmeyer, L.: TRAPACC and TRAPACCS at PARSEME Shared Task 2018: Neural Transition Tagging of Verbal Multiword Expressions. In Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), ACL, pp 268-274 (2018)
15. Dupal, J.: Investigating the Phrasicon of CLIL and NON-CLIL Students : a corpus-based comparative analysis using IdiomSearch. Thesis, Université catholique de Louvain, Louvain-la-Neuve (2018)
16. Sinclair, J.: Corpus, concordance, collocation. Oxford University Press, Oxford (1991)
17. Barreiro, A., Monti, J., Batista, F., Orliac, B.: When multiword go bad in machine translation. In Proceedings of the workshop on multi-word units in machine translation and translation technologies, 14th Machine Translation Summit, Nice (2013)

Appendix: List of articles

French articles:
- Brexit : le chaos et des pénuries prévisibles au Royaume-Uni en cas de « no deal » (Le Monde, 19 August 2019)
- Brexit : Boris Johnson prépare l'UE et le Royaume-Uni à un « no deal » (Le Monde, 21 August 2019)
- Brexit, « gilets jaunes », sommet du G7…, Macron aborde les dossiers de la rentrée (Le Monde, 21 August 2019)
- Brexit : à Londres, la bataille du « no deal » a commencé (Le Monde, 28 August 2019)
- Suspension du Parlement britannique : le coup de poker de Boris Johnson (Libération, 28 August 2019)
- Brexit : l'« outrage constitutionnel » de Boris Johnson (Le Monde, 29 August 2019)
- Royaume-Uni : « La piste d'un gouvernement d'union nationale semble de moins en moins plausible » (Le Monde, 29 August 2019)
- En suspendant le Parlement britannique, Boris Johnson aggrave la crise du Brexit (Le Monde, 29 August 2019)
- Parlement suspendu : Johnson prend son temps (Libération, 29 August 2019)
- Les Pays-Bas tirent profit de « l'effet Brexit » (Le Monde, 30 August 2019)

English articles :
- Eurocrats know Boris Johnson well, making no-deal Brexit more likely (The Economist, 15 August 2019)
- What Britain's release of an Iranian tanker says about its post-Brexit foreign policy (The Economist, 22 August 2019)
- Britain to be 'energetic partner' after Brexit (The Guardian, 23 August 2019)
- Johnson Walks Tightrope at G7, as Trump Pledges 'Very Big Trade Deal' for U.K. (The New York Times, 25 August 2019)
- Jeremy Corbyn agrees to prioritise legislation to stop no-deal Brexit (The Guardian, 27 August 2019)

12

- Boris Johnson is trashing the democracy fought for with the blood of our ancestors (The Guardian, 28 August 2019)
- Johnson has wrongfooted opponents of no deal. An election surely looms (The Guardian, 28 August 2019)
- Boris Johnson suspends Parliament, causing uproar (The Economist, 29 August 2019)
- Has Brexit destroyed party loyalty? Ruth Davidson's resignation will show us (The Guardian, 29 August 2019)
- Boris Johnson's Parliament Suspension Prompts Growing Backlash (The New York Times, 29 August 2019)