DeepL vs Google Translate: who's the best at translating MWEs from French into Polish? A multidisciplinary approach to corpora creation and quality translation of MWEs

Emmanuelle Esperança-Rodier, Damian Frankowski

DeepL vs Google Translate: who's the best at translating MWEs from French into Polish? A multidisciplinary approach to corpora creation and quality translation of MWEs

Emmanuelle Esperança-Rodier
Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP*, LIG, 38000 Grenoble, France
emmanuelle.esperanca-rodier@univ-grenoble-alpes.fr

Damian Frankowski
Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP*, LIG, 38000 Grenoble, France
frankowskidmail@gmail.com

Abstract

This article proposes a multidisciplinary approach to the creation of parallel French and Polish corpora annotated with multi-word expressions and the analysis of neural machine translation errors of annotated MWEs, from French into Polish. Our work first aims at building parallel FR-PL corpora from a French News corpus, taken from the WMT 2010 corpora of 40 000 words, by automatically translating them into Polish using the main commercial systems DeepL and Google Translate. The French source corpus has already been manually annotated with MWE, using the typology developed at LIDILEM (Tutin and Esperança-Rodier 2019). In a second step, the quality of the MWE translation of the FR-PL parallel corpora has been evaluated by annotating the translation errors creating a new error typology based on MQM-DQF (Lommel & al., 2018a) and the linguistic features of MWE translations using the ACCOLÉ platform (Brunet-Manquat and Esperança-Rodier, 2018). As a start, we have selected 154 sentences (4 332 French words) from the already MWE annotated French Document and translated them into Polish using DeepL (3 599 Polish words) and Google Translate (3 519 Polish words). The first general result shows that for the French to Polish language pair, DeepL translates MWE better than Google Translate even though it uses English as a pivot.

1. Introduction

Numerous works and workshops have focused on multiword expressions (MWE), that can be defined as recurrent word combinations in which the general meaning does not correspond to the literal meaning carried by its individual lexical items (Firth 1957 and Sag et al. 2002). Some of those works entailed the creation of several corpora as described in Constant et al. (2017). Among those, we can cite two corpora of nominal and adverbial MWEs (Laporte et al. 2008a; Laporte et al. 2008b) which do not provide a typology. In addition, there is the French Treebank (Abeillé et al. 2003) which contains various MWEs, including verbal ones, but only on continuous expressions. As far as Polish is concerned, several works on MWEs have been conducted, and if only one should be mentioned, the work of Savary (2001) presented the named entity annotation subtask of the National Corpus of Polish. Finally, the ANR project PARsing and Multiword Expression (PARSEME - Project ANR-14-CERA-001), noting the lack of linguistic resources related to this topic, has constituted corpora of syntactic annotations of MWEs (especially verbal and nominal) (Candito et al. 2017) as well as tools related to the analysis of MWEs.

MWE research is an active and topical research topic as seen with the many conferences held including the Multiword Expressions (MWE) workshops, organized since 2007 by the Special Interest Group on the Lexicon (SIGLEX) of the Association for Computational Linguistics (ACL) and supported by the Global Wordnet Association (GWA).

This article proposes first, to create a parallel French and Polish corpora annotated with

MWEs, in order to further analyze NMT errors of annotated MWEs, from French into Polish. We are providing the Tutin's typology (Tutin and Esperança-Rodier 2019) that describes a wide range of MWEs allowing to annotate non continuous expressions, as well as an MT error typology adapted to MWE translations from the MQM-DQF typology (Lommel et al, 2018a).

After a short state-of-the-art presentation, we first address the methodology used by describing the typology of MWE, the typology of Machine Translation (MT) errors, the corpus itself and the evaluation tool. Secondly, we present the results of the evaluation with examples and discuss those results to answer our question.

2. State of the Art

As Sag et al. (2002) already noticed, MWEs represent a real challenge for Natural Language Processing and Neural Machine Translation (NMT) have to face up to this demanding task even though they reach record quality levels.

Furthermore, Castilho et al. (2017) figured out that even if NMT systems achieved brilliant results, the human evaluations were not as enthusiastic as the automatic metrics, especially on adequacy and post-editing efforts. Koehn and Knowles (2017) also demonstrated that although NMT has achieved some great success, it still faces various challenges, notably out-of-domain performance and limited resources.

Since entire sentences are converted to vectorial representations in NMTs (Rikters and Bojar, 2017), and because of the absence of phrasal segmentation in NMTs (Zaninello and Birch, 2020), MWEs are particularly difficult to identify. Colson (2020) reports that in about 40% of MWE translations Google Translate did make a mistake.

Automatic metrics are often difficult to interpret and do not identify the main translation problems as Vilar et al (2006) mentioned while offering a MT error typology in order to offer an analysis of the translation errors. In the literature, error typologies are used for assessing MT outputs as Popović (2018) mentioned and Lommel (2018b) offered a proposal of standardization of error classification with MQM-DQF.

The study we describe hereafter, positions itself among those works, offering a way of annotating all the MWE types, those which are continuous and those which are non-continuous, and proposing a standardized translation error typology adapted to MWEs.

3. Methodology

We first aim at building parallel FR-PL corpora, taken from the News WMT 2010 corpus (40 000 words), by automatically translating them into Polish using DeepL and Google Translate. The French corpus has already been manually annotated with MWE, according to the typology developed at LIDILEM (Tutin and Esperança-Rodier, 2019). Then, we evaluated the quality of the MWE translations of the FR-PL parallel corpus by annotating the translation errors and the linguistic features of MWE translations using the ACCOLÉ annotation platform (Brunet-Manquat and Esperança-Rodier, 2018).

In what follows, we are presenting the corpus creation, the MWE, the MT Error typologies used, and ACCOLÉ.

3.1. Corpus

As a start, we have selected the first 154 sentences, representing 4 332 French words, from the French Document and translated them into Polish using Google Translate and then DeepL, thus obtaining two translated documents of respectively 3 519 and 3 599 Polish words, as illustrated in figure 1.

We used the News corpus from WMT 2010, as it had already been annotated according to the MWE typology explained hereafter.

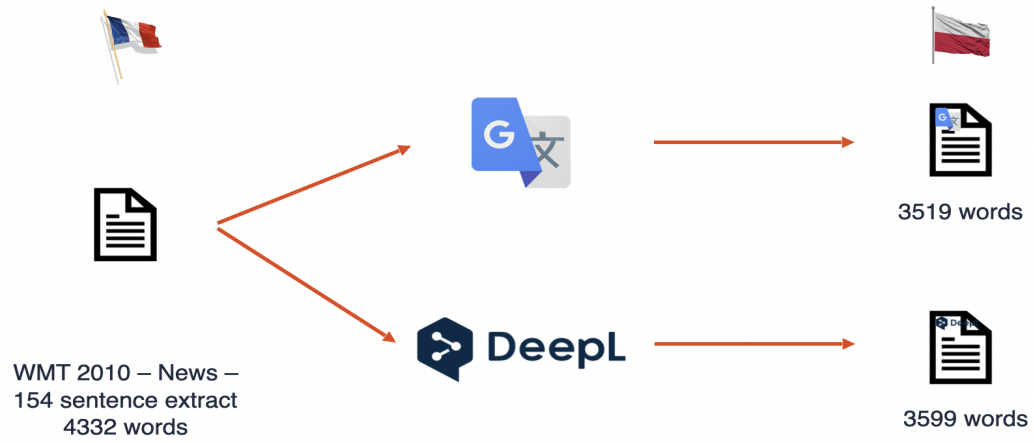

Figure 1: Corpus Creation Scheme

3.2. MWE Typology

We have selected Tutin's typology (Tutin et al., 2019), as described in Table 1, as it allows to annotate non-continuous expressions, and covers a large range of MWEs.
Nevertheless, we can note that the typology is made of eight types, including routine formulae, such as "it must be noted" and Pragmatic MWEs "you're welcome".

Multiword expressions	**Descriptions**	**Examples**
Idioms	frozen multiword expressions	*cul de sac* (fr) 'dead end'; *prendre en compte* (fr) 'take into account'
Collocations	preferred binary association, including light verb constructions	*gros fumeur* (fr) 'heavy smoker'; *faire une promenade* (fr) 'to take a walk'
Functional Multiword Expressions	functional adverbs, prepositions, conjunctions, determiners, pronouns.	*c'est pourquoi* (fr) 'that is why'; *d'autre part* (fr) 'on the other hand'; insofar as
Pragmatic MWEs	multiword expressions related to specific speech situations.	*de rien* (fr) 'You're welcome'; *à plus tard* (fr) 'see you later'.
Proverbs		*Pierre qui roule n'amasse pas mousse* (fr) 'A rolling stone gathers no moss'
Complex terms		natural language processing
Multiword Named entities		*Université Grenoble Alpes*; the European Union;
Routine formulae	routines generally associated to rhetorical functions	*force est de constater* (fr) 'it must be noted'.

Table 1: Tutin et al (2019) MWE typology

In our typology, it is noteworthy that idioms also include compound nouns; besides, only multiword named entities are considered as MWEs. On top of annotating the type of the MWE, the Part of Speech (POS) of the MWE was also annotated.
Once translated, we proceeded to the qualitative evaluation of the two Polish documents using the collaborative Platform ACCOLÉ, described below.

3.3. Annotation Platform: ACCOLÉ

ACCOLÉ is a collaborative error annotation platform described in Brunet-Manquat et al (2018). To annotate the errors made by the two MT systems, we have created two projects, one for each MT system.

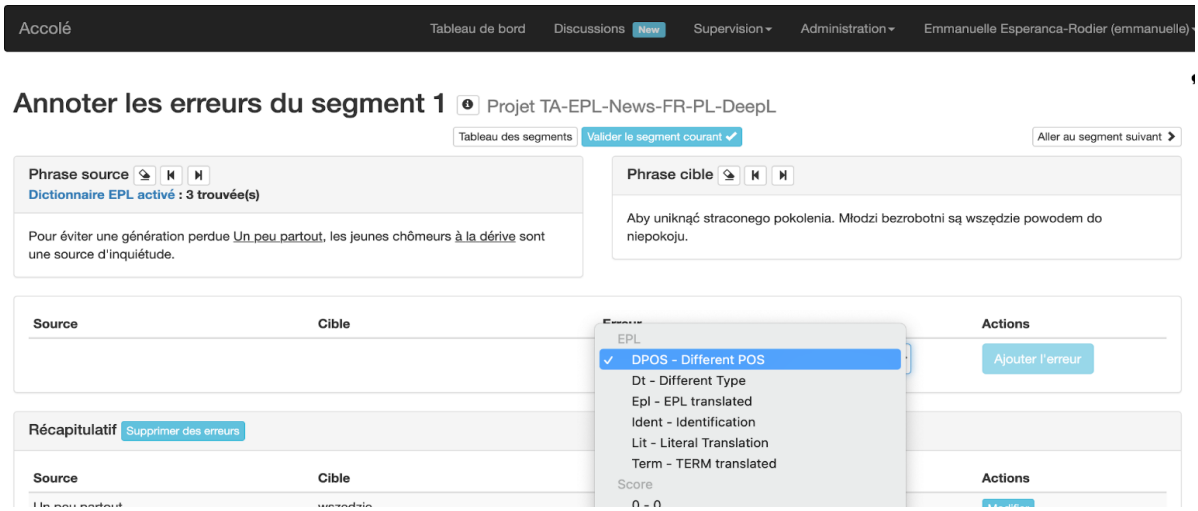

Figure 2: ACCOLÉ screenshot of MWE and MT error annotation

As shown in Figure 2, ACCOLÉ displays the French source and the Polish translation. The platform indicates a potential MWE by underlying several words in the French source. The annotator, a native Polish linguist graduated of English studies at Jagiellonian University and French studies (translation and interpretation) at Grenoble Alps University, selected the French MWE as well as its Polish translation and annotated them according to the MT error typology that we explain henceforth.

3.4. MT Error Typology

We have elaborated our own typology from MQM-DQF (Lommel et al, 2018a) addressing MT

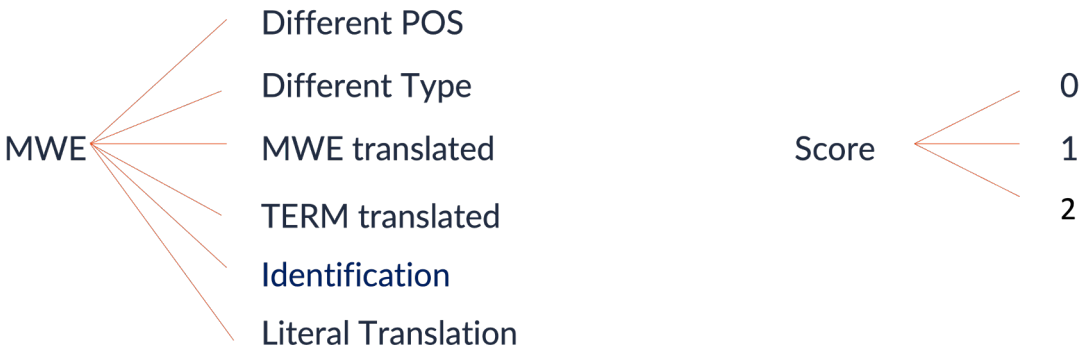

Figure 3: MWE annotation

errors, as MQM-DQF aimed at being a harmonized typology, and that offers a wide range of error types best suited to the Polish language; but also, the way MWEs were translated, as shown in Figure 3. “Different POS” refers to a MWE whose part of speech was different from the original French one. “Different Type” means that the MWE translated into Polish was of a different type than the French one. “MWE translated” indicates the case when the French MWE was translated into Polish as a MWE as well. “Term translated” denotes the translation of a French MWE into a Polish one, as a single term. We added “identification” because of an issue

with the internal identification. “Literal” is the label which we gave to the literal translation of MWE into Polish.

Inspired by MQM-DQF metrics, we have assigned a score to each MWE translation. Yet, we have not computed any metrics. Score 0 is assigned when the MWE is translated incorrectly with a mistake; Score 1 when the MWE translation is comprehensible without a mistake, however, we can correct it to make the translation more appropriate to the context; and finally for translations that are well translated and without a mistake, a score of 2 is attributed.

Turning to the MT errors, we have selected the items provided by MQM-DQF typology (Lommel et al, 2018a), listed in Figure 4.

Figure 4: MT Error Typology

We have selected four types for Accuracy. When a MWE is added into the Polish translation, the type is “Addition”, and “Omission” when a MWE is missing in the translation. When the French MWE is not translated into Polish, the type is “Untranslated”, and “Mistranslated” when the French MWE is not translated as it should.

As far as “Fluency” is concerned, we have selected six types, of which three, “Spelling”, “Style” and “Typo” were divided into subtypes.

As said, we have tested our typology on fifty sentences to check that all the encountered cases could be annotated using our typology.

Having described the MWE typology, and the MT error typology used to annotate the translation quality of Google Translate and DeepL within ACCOLÉ, we now introduce the evaluation results in what follows.

4. Evaluation Results

In order to provide a better picture of our data, we have performed a short analysis of the MWEs contained in the French document,

4.1. French-Source MWE-distribution

As shown previously in Figure 1, our French source corpus is made of 4 332 words. We have annotated 1 546 words as being MWE, which represents 35% of the total words. This figure complies with Cartier (2008) who showed that MWEs represented 30% of the unit of a given document. Among those 1 546 words, we have mainly annotated “Collocations” (32%), and “Function Word” (“Functional Multiword expressions” in the MWE typology) (24%), as shown in Figure 5. We had only two “Pragmatemes” (“Pragmatic MWEs”) representing 0%

in our data and no proverbs.

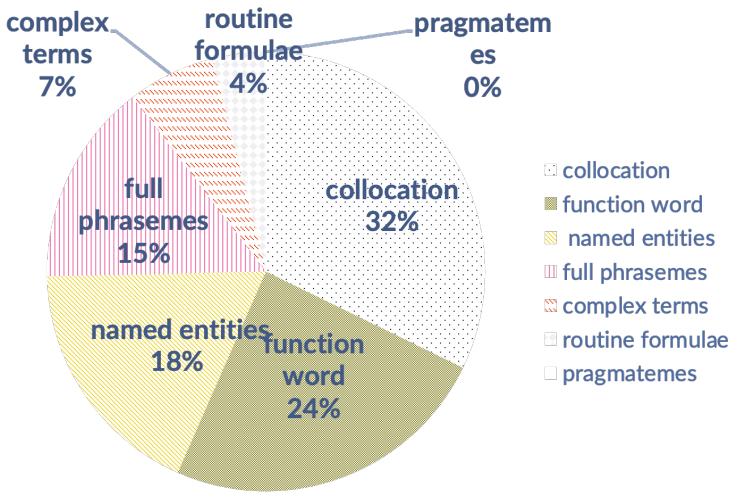

Figure 5: MWE distribution in French Source Document

But we did have some multiword "Named Entities" (18%), and some "Full Phrasemes" ("idioms "- 15%). Then, to a lesser extent, we had a few "Complex Terms" (7%), and "Routine Formulae" (4%).

Keeping this in mind, we provide the quality evaluation results according to Google Translate and DeepL.

4.2. Google Translate hypothesis evaluation results

Figure 6 shows that most of the translation errors were made on "Collocations" (C - 34%), then in order of most occurrences, on "Function words" (F - 20,5%), on "Full Phrasemes" (PH - 16%), on "Complex Terms" (T - 13%), on "Routine Formulae" (R - 8%), on "Named Entities" (NE - 8%) and a negligible number on "Pragmatemes" (P - 0,5%). Regarding the distribution of MWEs in the source document, it is logical that most of the translation errors are found in the collocation type as it is the most numerous. The same works for "Function Words". We cannot conclude anything with "Pragmatemes" as they are underrepresented in our corpus, so we will not analyze this type in what follows. Nevertheless, it should be noted that Google Translate made two mistakes on this MWE Type.

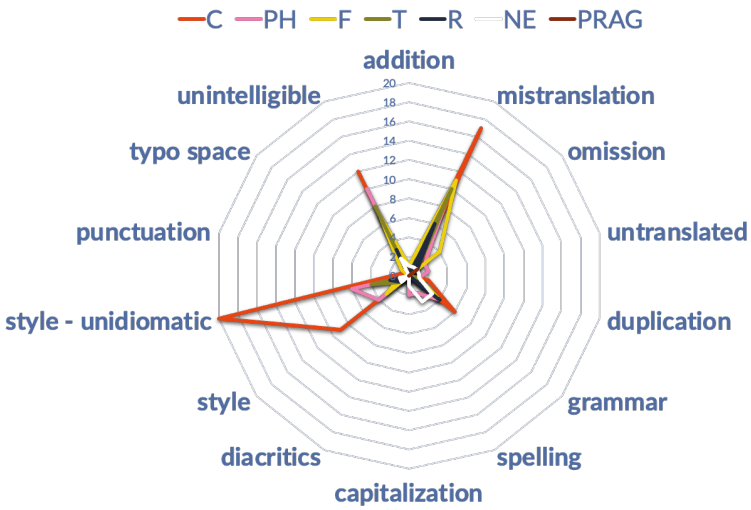

Figure 6: Google Translate error type per MWE type

As shown in Appendix A, the very few errors made on “Named Entities” appear in all the “Accuracy” error types while they occur on seven of the ten error types of “Fluency”. “Mistranslation” is the main error type for “Routine Formulae”, while for the “Fluency” error types, “Grammar” is the main error followed by “Unintelligible” and the two “Style” types. “Complex Terms” behave in the same way for “Accuracy” errors, but when examining the “Fluency” errors, “Unintelligible” comes second, followed respectively by “Style-unidiomatic” and “Grammar”. Looking at “Function Words, most of the “Accuracy” errors were of the “Mistranslation” type, with the second rank of errors being the “Omission” type. Still on “Function Words”, most of the “Fluency” errors are from the “Unintelligible” error type, the first type of errors, followed by “Style-unidiomatic”, “Style”, “Spelling” and “Capitalization”. To finish with MWE types, “Collocations” being the most represented MWE type in the source document, got most of the “Accuracy” error type, with the most of “Mistranslation” errors. “Collocations” also happened to have the most of “Fluency” errors, in the “Style-unidiomatic” type, followed by “Unintelligible”, “Style”, “Grammar”, “Duplication” and “Capitalization”.

Translation errors were mainly, almost 66%, “Fluency” errors, with “Unintelligible” ranking first at 20,5% of the total errors and roughly 31% of “Fluency” errors, happening in the whole range of MWE type but principally on “Collocations” and “Full Phrasemes”. “Unintelligible” type is followed by a 17% of “Style-unidiomatic” type, representing a quarter of “Fluency” errors and ranging over all the MWE types but mainly on “Collocations”. The third error type is the “Style” one, representing 9% of the total errors, and 14% of the “Fluency” errors, happening on the whole range of MWE types but for a majority on “Collocation”, then “Full Phraseme” and “Function terms”. Finally, “Grammar” type occurred to a lesser extent (11% of total errors but 17% of “Fluency” errors) and almost equally among the whole MWE types, with a little rise on “Collocations”.

For “Accuracy” errors, 34% of total errors, “Mistranslation” type was the most spread over the MWE types (26% of total errors, and 77% of “Accuracy” errors), occurring on the whole range of MWE types, as well as the “Omission” type (16% of “Accuracy” errors and merely 5,5% of total errors), that was also spread over all the MWE types but to a much lesser extent. Looking at the “Addition” type (merely 3% of “Accuracy” errors), it only occurred on “Function terms” and “Named Entities”. As regard to the “Untranslated” type (4% of “Accuracy” errors), it happened for “Full phrasemes” and “Named Entities.

However, we can see in Table 2, that the second rank of errors corresponds to “Full Phrasemes” that only represents 15% of the whole MWE types. “Full Phrasemes” represent

					Google Translate			DeepL		
	#Words/ types	% #total Words	% #MWE Words	MWE type Rank	#TotalErrors/MWE	%	Error Rank	#Total Errors/ MWE	%	Error Rank
collocation	498	2,30	32,21	1	70,00	33,98	1	20	48,78	1
function word	377	1,74	24,39	2	33,00	16,02	3	10	24,39	2
named entity	279	1,29	18,05	3	16,00	7,77	6	6	14,63	3
full phraseme	231	1,07	14,94	4	42,00	20,39	2	10	24,39	2
complex term	100	0,46	6,47	5	27,00	13,11	4	0	0,00	-
routine formulae	59	0,27	3,82	6	17,00	0,08	5	0	0,00	-
pragmateme	2	0,01	0,13	7	1,00	0,49	7	0	0,00	-
TOTAL	1546	-	-	-	206	-	-	41	-	-

Table 2: Error ranking according to MWE types

the fourth types to occur in the distribution of MWE on the French source document. Unlike "Named Entities", which ranked third of the MWE types, representing 18%, while ranking sixth in the error types.

We can thus draw the conclusion that Google Translate had difficulties translating "Full Phrasemes" as it is the fourth most represented MWE type in the French text and the second with the most errors, while it performed quite well on "Named Entities", which were the third MWE type in the French document, but only the sixth represented with errors. Google Translate also had more difficulties on "Fluency" than on "Accuracy". And when making errors, it was, respectively, mostly "Unintelligible" and "Mistranslations".

We are now examining the examples given in Appendix B, taken from the research we have conducted. We will first discuss the examples of the "Mistranslation" type taken from Google Translate.

In [1], the French term *Gouverneur* 'Governor' has been translated by Google Translate into Polish as *gubernator* which is a literal translation. It should rather be 'president of the national bank of Poland' and in Polish we say *prezes*. As far as the word *gubernator* is concerned, it is applied in the Polish language to name the governor of the state in the USA, Connecticut or Illinois for example.

In [2], the expression *faire la guerre* 'to wage war' was not translated correctly into Polish. The exact translation should be *prowadzić wojnę* meaning in French *faire la guerre*. Google Translate rendered the French expression *faire la guerre* as *a wyruszyć na wojnę* 'to set off', 'to go to war'.

The following example [3] seems to be a little bit amusing and improbable. *À la seule condition que* 'with the only condition that' was translated into Polish as *podeszwy* 'soles' which means the bottom of the foot and has nothing to do with the context. The correct version in Polish is *pod jednym warunkiem, że*.

Turning to [4], the phrase *pour que* is used to express purpose in French. Nevertheless, Google Translate has rendered it as a conjunction *że* 'that'. The mistake may result from the fact that in Polish to express purpose we use a very similar word *żeby*.

This leads us to the "Unintelligible" translations. Our fifth example is also absurd and amusing.

Observing [6], we can see the word *itonia*. The correct translation is the country *Estonia* 'Estonia', therefore the letters e and s were omitted.

Going over [7], we can see that the words *l'adoption de l'euro* 'the euro adoption' were repeated several times and additionally there is an apostrophe which is not required in this context. This is strange as stammering happens only with MT systems which are not well trained. And Google Translate should be well trained.

Ending this subsection, we are now going to focus on Deepl Quality Evaluation of the MWE translations.

4.3. DeepL hypothesis evaluation results

We are now analyzing the results obtained while performing the Quality Evaluation of the DeepL MWE translations.

Figure 7 shows that most of the translation errors were made on "Collocations" (C-49% of the total errors) followed then by "Function Terms" (T-24%), "Named Entities" (NE-15%) and finally "Full Phrasemes" (PH-12%). DeepL did not make any mistakes on "Complex Terms", "Routine Formulae" nor "Pragmatemes". The distribution of MWE types over the source document does not provide sufficient occurrences of "Pragmatemes" to conclude anything. However, over the two words annotated as "Pragmateme" MWE type, DeepL did not make any mistakes.

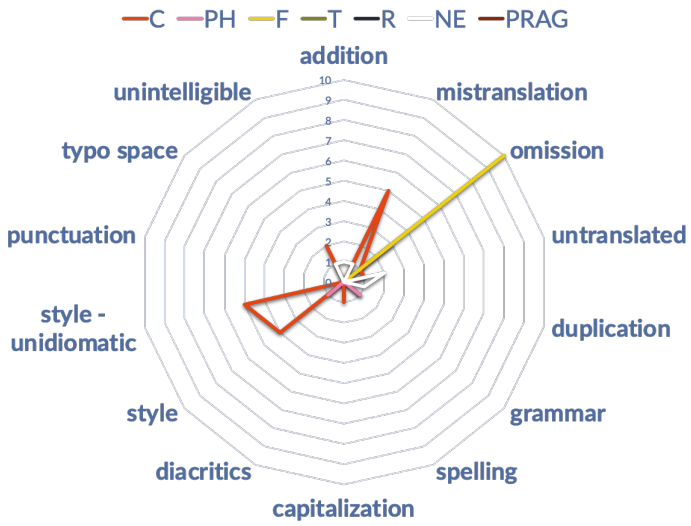

Figure 7**:** DeepL error type per MWE type

Referring to Appendix C, “Collocations” addressed most of the Translation errors, especially from the “Fluency” type, with mainly the two “Style” error types, followed by the “Unintelligible” type and “Capitalization” and “Grammar”; while only being the second MWE type for the “Accuracy” errors. The first rank for “Accuracy” is taken by the “Function Term” MWE type, with all the “Accuracy” mistakes labeled with the “Omission” type, while no error was labeled under “Fluency”. The third MWE type, with the most errors, is the “Named Entity” type, with four errors in “Accuracy”, mostly of the “Untranslated” type, and only two “Fluency” errors i.e., “Unintelligible” and “Duplication”.

Looking at the “Accuracy” errors which represent most of the errors made by DeepL with 58,5% of the total amount of errors, most of those errors were on the “Omission” type with 34% of the total errors and 58% of the “Accuracy” errors. “Omission” type occurred most on “Function” Terms, then to a much lesser extent on “Full Phrasemes”. “Mistranslations” ranked second representing 15% of the total errors and a quarter of the “Accuracy” errors, occurring only for “Collocations” and “Named entities” (one occurrence). The “Untranslated” error type comes third representing 7,5% of the total errors and 12,5% of the “Accuracy” errors, happening mainly on “Named Entities”. Addition was the last error type to be made by DeepL (2,5% of total errors and 4% of “Accuracy” errors) and with only one error on “Named Entities”.

Nevertheless, we can notice in Table 2, leaving “Collocations” in first place, that “Full Phrasemes” rank second at the translation error type scale, equally with “Function” Word type, while representing 15% of the MWE types over the French source document.

This leads us to the conclusion that DeepL has difficulties in translating “Full Phrasemes”, while it behaves pretty well on “Complex Terms” and “Routine Formulae”. When DeepL makes errors, they are mainly of the “Accuracy” type with the habit of omitting to translate. As far as the “Fluency” type is concerned, DeepL struggles with the “Style”.

Addressing, in Appendix B; example [8], DeepL has translated the whole sentence into Polish yet the expression *le Projet d'assistance à l'école secondaire féminine* was left out and translated into English. This may be evidence of the use of English as a pivot language by DeepL.

In order to illustrate the “Omission” error type, [9] shows that the omission of *de plus en plus* ‘more and more’ implied a Score 0. Although the phrase *de plus en plus* wasn't translated into Polish, the meaning of the sentence wasn't affected and the accuracy as well the fluency was

preserved. If we wish to translate the phrase de *plus en plus*, the translation would be *coraz to bardziej*.

Having investigated the MT errors in both Google Translate and DeepL translations, we are now analyzing the way MWEs have been translated.

4.4. MWE evaluation

Both DeepL and Google Translate translated correctly French MWEs into equivalent Polish MWEs. Google Translate translated a bit more French MWE into terms than DeepL but did the double of literal translations. Finally, DeepL used more frequently a Polish MWE with a different POS and a different MWE type than Google Translate.

	Google Translate	DeepL
EPL translated	406	410
Term translated	163	150
Literal translation	20	9
Different POS	54	77
Different type	4	8

Table 4: Number of MWEs translated by Google Translate and DeepL according to the different MWE types.

Keeping this in mind, we are going to comment on the Score evaluation.

4.5. Score evaluation

Now looking at the scores ("Score 0" for incorrect translations, "Score 1" for approximated translations which could be improved, and "Score 2" for correct translations) we see that Google Translate makes correct translations but, makes more incorrect and approximated translations than DeepL. Nevertheless, DeepL makes much less mistakes and provides more correct translations.

	Google Translate	DeepL
Score 0	70	17
Score 1	86	30
Score 2	434	540

Table 5: Number of MWE translations being scored 0, 1 or 2 according to Google Translate and DeepL

5. Google Translate vs DeepL

Here, we compare the evaluation results of Google Translate and DeepL by going over the MT error, the MWE and the Score analyses.

DeepL makes much less errors than Google Translate (cf. Table 2), respectively 41 error annotations versus 206 ones, which is confirmed by the scores. Google Translate makes errors over the whole range of MWE types, while DeepL only makes mistakes on "Collocations", "Function Words", "Named Entities" and "Full Phraseme". DeepL makes more "Accuracy" errors than Google Translate (respectively 58,5% and 34%) which makes more "Fluency" errors (respectively 41,5% and 66%). Google Translate makes more "Mistranslation" and "Unintelligible" translations than DeepL which makes more "Omission" and "Style" errors.

DeepL uses more MWEs with a different POS or MWE type than Google Translate which

translates more literally and uses more terms than equivalent Polish MWEs (cf. Table 3).

Back to Appendix B, we can notice in [10] that both Google Translate and DeepL translate exactly in the same way and make a "Style-unidiomatic" error. In the phrase *offrir une deuxième chance*, the verb *offrir* 'to offer' was translated into Polish literally as *zaoferować*. Nevertheless, the expression *zaoferować drugą szansę* is unidiomatic and we rather *say dać* which means *offrir*.

We can notice in [11] that the noun *l'offre* 'the offer' derived from the verb *offrir* which we saw in [10]. In this case, the noun was translated correctly by Google Translate but DeepL failed to correctly translate the noun.

In [12], the collocation *dure realité* 'harsh reality' has been mistranslated in each case. The Polish translation is correct however, it is inappropriate in this context. Instead of using the adjective *ostrej* 'sharp' in Google Translate hypothesis or *twardej* 'hard' in DeepL hypothesis, in Polish we more frequently say *trudna* 'difficult' which seems more natural.

In the subsequent example [13] of "Style" error type below, both expressions have been translated in a way that does not accurately represent the MWE in the source sentence.

The sentence rendered by Google Translate is comprehensible and grammatically correct. Nonetheless the Polish sentence has a stylistic error. Studying the MWE *ułatwić dostęp do edukacji na poziomie średnim* we believe that this translation needs to be rephrased. As such, it would be better to say *ułatwić dostęp do edukacji w szkole średniej* 'facilitate access to education at secondary school'.

In the example [14], *premier cycle d'éducation* should also be corrected to sound more natural and less artificial in Polish. Therefore, instead of *szkołę średnią I stopnia* it would be better to say *pierwszy etap edukacji* 'the first stage of education'.

This last example concludes this section, we are now going to sum up our findings.

6. Conclusion

Having considered MT error, MWE translation, and Score evaluations, as well as the different examples, we can now easily answer our question, namely "DeepL vs Google Translate: who's the best at translating MWEs from French into Polish?": DeepL best translates French MWEs into Polish MWEs, with much less errors compared to Google Translate, even if it tends to "omissions" and struggles with the "style" error type. DeepL also uses more MWEs with a different part of speech or MWE type, than the ones in the French source. This, as translation studies have already demonstrated, denotes that it does not impact the translation quality. Both MT systems struggled to translate Full Phrasemes. We also found out that DeepL used English as a pivot between French and Polish. Finally, as the length of our sentences was homogeneous, we could not see if the length of a sentence did imply any specific mistakes.

Further to this work, we will finish the FR-PL annotated translation corpus. We will also investigate the explanation of the MT errors. We plan to study from this corpus what are the different factors that imply MWEs to be MWE translated, Term Translated or Literally translated, and which factors entail that MWEs are more likely to be translated with different POS or MWE type.

Acknowledgements

We would like to thank the Pôle Grenoble Cognition who funded the internship during which this work has been performed.

References

Cartier, E. (2008). Repérage automatique des expressions figées : état des lieux, perspectives. In Blumenthal, P., & Mejri, S. Les séquences figées : entre langue et discours. (p. 55-70). Stuttgart: F. Steiner

Candito, M., Constant M., Ramisch C., Savary A., Parmentier Y., Pasquer C. & Antoine J.-Y. (2017). Annotation

d'expressions polylexicales verbales en français. In TALN 2017, Actes de TALN 2017, Orléans, France : Association pour le Traitement Automatique des Langues.
Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J. and Way, A. (2017) 'Is Neural Machine Translation the New State of the Art?', The Prague Bulletin of Mathematical Linguistics. 108, 109-120
Colson, J.P. (2020). Computational phraseology and translation studies. Dans G. Corpas Pasto, & J.P. Colson. (édité). *Computational Phraseology* (p.65-81). John Benjamins Publishing Company.
Constant, M., Eryiğit, G., Monti, J., Van Der Plas, L., Ramisch, C., Rosner, M. and Todirascu, A. (2017) Multiword Expression Processing: A Survey. Computational Linguistics, 43(4), 837-892.
Brunet-Manquat, F., Esperança-Rodier, E. (2018). ACCOLÉ : Annotation Collaborative d'erreurs de traduction pour COrpus aLignÉs, Démonstration TALN18, Rennes.
Firth, J.R. (1957). A Synopsis of Linguistic Theory, 1930-1955 Studies in Linguistic Analysis Special Volume, Philological Society. 1-32.
Koehn, P. and Knowles, R. (2017) 'Six Challenges for Neural Machine Translation', Proceedings of the First Workshop on Neural Machine Translation. 28-39. Vancouver, Canada, 4 August.
Laporte, E., Nakamura, T., & Voyatzi, S. (2008a). A French corpus annotated for multiword nouns. In Language Resources and Evaluation Conference. Workshop Towards a Shared Task on Multiword Expressions. pp. 27-30.
Laporte, E., Nakamura, T., & Voyatzi, S. (2008b). A french corpus annotated for multiword expressions with adverbial function. In Language Resources and Evaluation Conference (LREC). Linguistic Annotation Workshop. pp. 48-51.
Lommel, Arle, and Alan K. Melby (2018a) Tutorial: MQM-DQF: A Good Marriage (Translation Quality for the 21st Century). 13th Conference of the Association for Machine Translation in the Americas (Volume 2: User Papers). Vol. 2.
Lommel, A. (2018b) Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies. In Moorkens, J., Castilho, Sh., Gaspari, F. and Doherty, S. (Eds.) Translation Quality Assessment: From Principles to Practice. Dublin: Springer International Publishing. 109-127.
MWE, http://multiword.sourceforge.net/PHITE.php?sitesig=CONF&page=CONF_03_MWE-WN_2019___lb__ACL__rb__, dernier accès le 26/10/2019
Popović, M. (2018) Error Classification and Analysis for Machine Translation Quality Assessment. In Moorkens, J., Castilho, Sh., Gaspari, F. and Doherty, S. (Eds.) Translation Quality Assessment: From Principles to Practice. Dublin: Springer International Publishing. 129-158.
Rikters, M., and Bojar, O. (2017) Paying Attention to Multi-Word Expressions in Neural Machine Translation. Preprint.
Tutin, A. & Esperança-Rodier, E. (2019). The difficult identification of multiword expressions: from decision criteria to annotated corpora; European Society of Phraseology Conference (EUROPHRAS 2019), Sept. 2019, Malaga, Spain.
Sag, I. A., Baldwin, T., Bond, F., Copestake, A. and Flickinger, D. (2002) Multiword Expressions: A Pain in the Neck for NLP. International Conference on Intelligent Text Processing and Computational Linguistics. 1-15. Springer, Berlin, Heidelberg.
Savary, A., Piskorski, J. (2001). Language Resources for Named Entity Annotation in the National Corpus of Polish. Control and Cybernetics, Polish Academy of Sciences,, 40 (2), pp.361-391.
Vilar, D., D'Haro, L. F., Xu, X. and Ney, H. (2006) 'Error Analysis of Machine Translation Output', Proceedings of the International Conference on Language Resources and Evaluation (LREC). 697-702. Genoa, Italy, 24-26 May.
Zaninello, A. and Birch, A. (2020) 'Multiword Expression Aware Neural Machine Translation', Proceedings of the 12th Language Resources and Evaluation Conference. 3816-3825. Marseille, France, 11-16 May.

Appendix A: Google Translate - Number of Translation errors per MWE types. A same MWE can be annotated with several translation error types.

Error Types/MWE	C	PH	F	T	R	NE	PRAG	TOTAL	TOTAL (%)
ACCURACY (TOTAL)	18	13	16	11	7	4	1	70	33,98
addition	0	0	1	0	0	1	0	2	0,97
mistranslation	17	9	11	10	6	1	0	54	26,21
omission	1	2	4	1	1	1	1	11	5,34
untranslated	0	2	0	0	0	1	0	3	1,46

FLUENCY (TOTAL)	52	29	17	16	10	12	0	136	66,02
duplication	2	1	1	1	0	1	0	6	2,91
grammar	6	4	4	2	4	3	0	23	11,17
spelling	0	2	0	0	0	3	0	5	2,43
capitalization	2	2	0	0	0	1	0	5	2,43
diacritics	0	0	0	0	0	0	0	0	0,00
style	9	4	3	1	1	1	0	19	9,22
style - unidiomatic	20	6	1	4	2	1	0	34	16,50
punctuation	1	0	0	0	0	0	0	1	0,49
typo space	0	0	1	0	0	0	0	1	0,49
unintelligible	12	10	7	8	3	2	0	42	20,39
TOTAL	70	42	33	27	17	16	1	206	-
TOTAL (%)	33,98	20,39	16,02	13,11	8,25	7,77	0,49	-	-

Appendix B: Translation Error Examples

MISTRANSLATION

Google Translate

[1] FR: Comme le fait remarquer Leszek Balcerowicz, le Gouverneur de la Banque nationale de Pologne, l'adoption de l'euro par les pays candidats [...]
*PL Hyp.: Ponieważ latka Balcerowicz wskazuje, że gubernator Narodowego Banku Polskiego, przyjęcie euro przez kraje kandydujące [...]

[2] FR: [...] plus personne n'estimera utile de faire la guerre pour déterminer quelle langue doit être parlée en Alsace-Lorraine.
*PL Hyp.: [...] aby nikt nie uznał za pożyteczne wyruszenia na wojnę w celu ustalenia, jakim językiem należy się posługiwać w Alzacji i Lotaryngii.
PL: [...] nikt nie uznał za pożyteczne prowadzić wojnę celu ustalenia, jakim językiem należy się posługiwać w Alzacji i Lotaryngii

[3] FR: Mais contre toute attente, il a adopté une position éthiquement avantageuse en proposant de renoncer à une partie de son rabais budgétaire, à la seule condition que le reste de l'Union européenne (et la France, en particulier) s'accorde sur les réformes fondamentales en matière de politique agricole européenne.
*PL Hyp.: Jednak wbrew wszelkim szansom, przyjęło etycznie korzystne stanowisko, proponując rezygnację z jej zniżki budżetowej, podeszwy, że reszta Unii Europejskiej (iw szczególności w szczególności) zgadza się na reformy fundamentalne w obszarze Europejska polityka rolna.
PL: ednak wbrew wszelkim szansom, przyjęło etycznie korzystne stanowisko, proponując rezygnację z jej zniżki budżetowej, pod jednym warunkiem, że Unii Europejskiej (i w szczególności w szczególności) zgadza się na reformy fundamentalne w obszarze Europejska polityka rolna

[4] FR: Il a fallu du temps pour que les Britanniques comprennent qu'étant donné que le Président Jacques Chirac [...]
*PL Hyp.: Brytyjczycy zajęło to, że Brytyjscy zrozumieli, że ponieważ prezydent Jacques Chirac [...]
PL: Brytyjczycy zajęło to, żeby Brytyjscy zrozumieli, że ponieważ prezydent Jacques Chirac [...]

UNINTELLIGIBLE

Google Translate

[5] FR: Ils pourraient commencer par considérer l'adoption précoce de l'euro par les pays candidats d'un œil plus favorable, aussi bien pour les pays disposant d'un système de caisse d'émission que pour ceux utilisant un taux de change flottant
*PL Hyp.: Mogliby zacząć od rozważenia wczesnego przyjęcia euro przez kandydujące kraje bardziej korzystnego oka, zarówno dla krajów z systemem ciasta, jak i dla osób korzystających z pływającego

kursu.

[6] FR: Trois d'entre eux --l'Estonie, la Lettonie et la Lituanie-- possèdent un système de caisse d'émission ou un taux de change fixe avec l'euro, tout comme la Bulgarie, qui doit rejoindre l'Union en 2007
*PL Hyp.: Trzy z nich –itonia , Łotwa i Litwa mają system funduszy emisji lub stałego kursu z euro, podobnie jak Bułgaria, aby dołączyć do Unii w 2007 roku.
PL: Trzy z nich –Estonia , Łotwa i Litwa mają system funduszy emisji lub stałego kursu z euro, podobnie jak Bułgaria, aby dołączyć do Unii w 2007 roku.

[7] FR: Malheureusement, l'entêtement de la BCE et de la Commission apporte maintenant de l'eau au moulin des gouvernements populistes dans les pays accédants qui comptent sur la variation des cours de change pour échapper à la discipline fiscale qu'implique l'adoption de l'euro
*PL Hyp.: Niestety, upór EBC i Komisja przynoszą teraz wodę do młynów rządów populistycznych w krajach przystępujących, które polegają na zmianie kursów walutowych, aby uciec przed dyscypliną podatkową dorozumianą przez przyjęcie przyjęcia przyjęcia Przyjęcie przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia "euro

DeepL

[8] FR: Au Bangladesh, le Projet d'assistance à l'école secondaire féminine a aidé financièrement les jeunes filles âgées de 11 à 14 ans à poursuivre leurs études et à retarder le moment du mariage.
*PL hyp.: W Bangladeszu projekt Women's Secondary School Assistance Project zapewniał pomoc finansową dziewczętom w wieku od 11 do 14 lat, aby mogły kontynuować naukę i opóźnić zawarcie małżeństwa

OMISSION

DeepL

[9] FR: Ils constituent notamment un problème dans les pays en développement qui connaissent une explosion démographique de plus en plus forte, l'équivalent du 'baby boom' en Occident à la fin de la Deuxième Guerre mondiale.
*PL Hyp.: W szczególności stanowią one problem w krajach rozwijających się, które doświadczają rosnącej eksplozji populacji, odpowiednika "wyżu demograficznego" na Zachodzie pod koniec drugiej wojny światowej
PL: W szczególności stanowią one problem w krajach rozwijających się, które doświadczają coraz to bardziej rosnącej eksplozji populacji, odpowiednika "wyżu demograficznego" na Zachodzie pod koniec drugiej wojny światowej.

STYLE - UNIDIOMATIC

[10] FR: Les gouvernements doivent aussi offrir une 'deuxième chance' aux jeunes qui ont échoué en raison des circonstances ou de mauvais choix.

Google Translate

*PL Hyp.: Rządy muszą również zaoferować „drugą szansę” młodym ludziom, którym zawiodły okoliczności lub złe wybory.

DeepL

*PL Hyp.: Rządy muszą również zaoferować "drugą szansę" młodym ludziom, którzy zawiedli z powodu okoliczności lub złych wyborów.
PL: Rządy muszą również dać „drugą szansę” młodym ludziom, którym zawiodły okoliczności lub złe wybory.

[11] FR: Avec la volonté politique, des mesures reposant sur les efforts des jeunes eux-mêmes […], et l'offre d'une deuxième chance à ceux qui en ont besoin, […].

Google Translate

*PL Hyp.: Dzięki woli politycznej, działaniom opartym na wysiłkach samych młodych ludzi […] oraz zapewnieniu drugiej szansy potrzebującym, […].

DeepL

*PL Hyp.: Dzięki woli politycznej, działaniom opartym na wysiłkach samych młodych ludzi w celu zwiększenia ich szans na sukces i umiejętności oraz oferowaniu drugiej szansy tym, którzy jej potrzebują […] .

[12] FR: Une fois que les Britanniques ont pris conscience de cette dure réalité, ils ont joué leur lamentable dernière carte dans la gestion des négociations.

Google Translate

*PL Hyp.: Gdy Brytyjczycy stali się świadomy tej ostrej rzeczywistości, grali ich żałowała ostatnią kartę w zarządzaniu negocjacjami.

DeepL

*PL Hyp.: Kiedy Brytyjczycy zdali sobie sprawę z tej twardej rzeczywistości, zagrali swoją żałosną ostatnią kartę w zarządzaniu negocjacjami.
PL: Gdy Brytyjczycy stali się świadomy tej trudnej rzeczywistości, grali ich żałowała ostatnią kartę w zarządzaniu negocjacjami

STYLE

Google Translate

[13] FR: Plusieurs pays essayent maintenant de faciliter l'accès à l'enseignement secondaire, notamment grâce à des programmes conditionnels de transferts' […]
*PL Hyp.: Obecnie kilka krajów próbuje ułatwić dostęp do edukacji na poziomie średnim, w szkole średniej w szczególności poprzez programy warunkowego przeniesienia, „[…]
PL: Obecnie kilka krajów próbuje ułatwić dostęp do edukacji, w szkole średniej w szczególności poprzez programy warunkowego przeniesienia, „[…]

DeepL

[14] FR: Au Maroc par exemple, plus de 80% des enfants accomplissent un premier cycle d'éducation, mais moins de 20% atteignent le niveau requis.
*PL Hyp.: W Maroku, na przykład, ponad 80 procent dzieci kończy szkołę średnią I stopnia, ale mniej niż 20 procent osiąga wymagany poziom.
PL: W Maroku, na przykład, ponad 80 procent dzieci kończy pierwszy etap edukacji, ale mniej niż 20 procent osiąga wymagany poziom.

Appendix C: DeepL - Number of Translation errors per MWE types. A same MWE can be annotated with several translation error types.

Error Types/MWE	C	PH	F	T	R	NE	PRAG	TOTAL	TOTAL (%)
ACCURACY	7	3	10	0	0	4	0	24	58,54
addition	0	0	0	0	0	1	0	1	2,44
mistranslation	5	0	0	0	0	1	0	6	14,63
omission	1	3	10	0	0	0	0	14	34,15
untranslated	1	0	0	0	0	2	0	3	7,32
FLUENCY	13	2	0	0	0	2	0	17	41,46
duplication	0	0	0	0	0	1	0	1	2,44
grammar	1	1	0	0	0	0	0	2	4,88
spelling	0	0	0	0	0	0	0	0	0,00
capitalization	1	0	0	0	0	0	0	1	2,44
diacritics	0	0	0	0	0	0	0	0	0,00
style	4	1	0	0	0	0	0	5	12,20
style - unidiomatic	5	0	0	0	0	0	0	5	12,20
punctuation	0	0	0	0	0	0	0	0	0,00
typo space	0	0	0	0	0	0	0	0	0,00
unintelligible	2	0	0	0	0	1	0	3	7,32
TOTAL	20	5	10	0	0	6	0	41	-
TOTAL (%)	48,78	12,20	24,39	0,00	0,00	14,63	0,00	-	-