DeepL and Google Translate Translating Portuguese Multi-Word Units into French: Progress, Decline and Remaining Challenges (2019-2023)

Françoise Bacquelaine[1]

[1]Centre of Linguistics of the University of Porto
franba@letras.up.pt/shirleybac@gmail.com

Abstract

The transition from statistical machine translation trained with machine learning to neural machine translation (NMT) using deep machine learning has proved successful for high-resourced languages. Researchers are exploring new avenues such as zero-shot NMT models for less-resourced languages or the use of English as a pivot language to improve NMT performance. A comparative study conducted in 2019 and 2021 on DeepL (DL) and Google Translate (GT) raw NMT output shows that the performance of GT deteriorated significantly in 2021, mainly because it seemed to use English as a pivot language between two romance languages. In 2023, the same sample of 167 instances of Portuguese multi-word units (MWU) expressing progression and proportion was translated into French by DL and GT. The output in 2019, 2021 and 2023 NMT is analyzed in terms of potential error factors in the Portuguese sample and actual error types in NMT output. The progress of DL from 2019 to 2023 is insignificant while GT exceeds its 2019 score after the 2021 decline. Stronger error factors are unusual structures, combination of potential error factors, and longer MWUs. Phraseology, calque and nonsense are the most frequent error types in this study on NMT progress, decline and remaining challenges.

Keywords: neural machine translation; pivot language; corpora; phraseology; error

1. Introduction

The rather soft transition from statistical machine translation (SMT), trained with machine learning, to neural machine translation (NMT), using ever-evolving deep machine learning, has proved successful. Apart from using the same data as SMT since the neural adventure began in the mid-2010s, researchers are exploring new avenues such as zero-shot NMT models for less-resourced languages (Zhang et al., 2020) or the use of English as a pivot language (Soler Uguet et al., 2022). After a remarkable improvement over SMT, progress has slowed down and successive attempts to improve the model may or may not be successful. A study conducted in 2021 (Bacquelaine, 2022a; idem 2022b) suggests that Google Translate (GT) sometimes uses English (EN) as a pivot language[1] to translate multi-word units (MWU) from Portuguese (PT) into French (FR). Consequently, its score drops dramatically compared to 2019 and other NMT systems (DeepL, eTranslation).

MWU translation is a challenge both for human translators and machines, mainly because ambiguity can raise "problems" at phrase, syntactic and semantic level (Koehn 2020). This study focuses on three PT MWUs. The first (*cada vez* COMP[2]) expresses quantitative or qualitative progression (PROG), the second (typically: NUM *em cada* QP[3]) indicates proportion between a set and a subset (P3S), and the third (typically: QP *por cada* QP) a proportion between two sets (P2S), as shown in examples (1) to (4) taken from the aligned corpus Europarl v7 (Tiedermann, 2012):

(1) Quantitative PROG
- *... cada vez mais mercados...*
- *... more and more markets ...*

(2) Qualitative PROG
- *... artes da pesca cada vez mais selectivas.*
- *... increasingly selective fishing gear.*

(3) P3S
- *... um em cada dois homens ...*
- *... one in two men ...*

(4) P2S
- *... uma embarcação por cada 70 cidadãos.*
- *... one boat for every 70 citizens.*

If the EN universal quantifier *every* is possible to translate P3S and P2S and the FR universal quantifier *toujours* to translate PROG, *each* and *chaque* are not, according to a human translation model obtained from several good quality aligned corpora (Bacquelaine, 2020). Hence the first criterium to assess NMT is literality or word-for-word translation. Any translation of these three MWUs in FR with *chaque* is considered as literal and therefore wrong. The second criterium is acceptability. Any translation by one of the model's solutions is acceptable. Typically, PROG translates in FR as *de* COMP *en* COMP, P3S as QP *sur* NUM, and P2S as QP *pour* QP.

In this narrow scope, this paper aims to examine the evolution of the literality and acceptability performance of DeepL (DL) and GT between 2019 and 2023, to determine a possible link between error factors in the PT sample and error types in NMT, and to identify, system by system, the progress, decline and remaining challenges according to error types detected in the output.

First, methodology, tools, corpora and analysis criteria are described in section 2. Then, results are presented and discussed in three subsections: the global evolution of DL and GT performance between August 2019 and January 2023, the assessment of the causal link between potential

[1] Markus Foti (DGT), personal communication (2021): EN is used as a pivot language in eTranslation.

[2] Comparative adjective or adverb: *mais, menos, melhor, pior, menor, maior*.

[3] QP: quantifier phrase consisting of a cardinal numeral adjective (NUM) and a noun (N), such as *três Portugueses - three Portuguese*.

error factors in the PT MWU instances and actual error types in the FR NMT output, and the evaluation of progress, decline and remaining challenges in 2023.

2. Materials and Methods

We adopt first a diachronic approach. The period covered is very short, but NMT has shown more progress in less than ten years than any other (hybrid) model before. The global evolution of GT and DL is evaluated in terms of acceptable MWU translations in FR. Then, potential error factors in the PT sample and actual errors in the NMT output by GT and DL (2019-2023) are analyzed to determine whether there is a causal link between them. Finally, the respective progress and decline of GT and DL are observed year by year, and remaining challenges in 2023 are identified.

DL and GT are two well-known NMT systems that can be used in daily life, mostly to translate general language. GT developed from statistical to neural MT and DL emerged as NMT from Linguee (dictionary and search engine for aligned bilingual segments).

For the first part of this study, a sample of 102 instances of PROG, 41 of P3S, and 24 of P2S was selected from *CETEMPúblico* (CTP), a Portuguese journalistic corpus from the end of the 20th century explored with AC/DC (Santos and Bick, 2000). PROG is much more frequent in general language than P3S and P2S, and some instances were selected because of specific translation challenges. So, the sample does not reflect general use, but we must presume the PT instances are correct. It was translated into FR by GT and DL in August 2019, September 2021, and January 2023. The raw NMT output is analyzed in terms of literality and acceptability.

For the second and third parts, PT instances that had been well translated (non-literal and acceptable FR MWUs) by GT and DL in 2019, 2021 and 2023 were excluded. The remaining corpus consists of 110 PT instances and 660 raw (mis)translations in FR by GT and DL in 2019, 2021, and 2023: 64 PT instances of PROG, 30 of P3S, and 16 of P2S.

Potential error factors in the PT sample result from the selection of specific instances to challenge the machine. Some challenges are common to PROG and proportion MWUs. They are classified into eight categories: (1) *cada vez mais/menos* as a sentence adverb of frequency quantification (AQF, Leal 2012); (2) splitting (*scission*); (3) inversion; (4) split inversion; (5) coordinated instance; (6) ellipsis in coordinated instance; (7) non-compositional sense (idiom in the broadest sense, including puns); (8) atypical preposition (PREP) in PT, i.e. other than *em* for P3S and *por* for P2S; (9) atypical structure of P3S (2 N instead of one); (10) long QP. Two or three factors may combine in a single instance.

Actual NMT errors detected in FR fall into eight types: (1) calque from PT; (2) calque from EN; (3) omission; (4) addition; (5) nonsense; (6) wrong meaning; (7) opposite meaning; (8) phraseological inadequacy. Apart from typical equivalents of PROG, P3S and P2S, other FR equivalents are attested in the model and accepted as possible translations of the PT MWUs. Omission and addition of part of the PT segment in the NMT output in FR usually result in semantic errors (5, 6 and 7) while phraseological errors at lexical or syntactic level may lead to lack of fluency (wrong collocation, wrong PREP, unusual inversion or splitting, …), agrammaticality (omission of internal argument, …) or semantic issues (5, 6, 7). So, a combination of errors is also possible.

To evaluate the causal link between potential error factors and actual errors, the approach is global, and results are presented in two steps: average number of errors per potential error factor then number of actual errors by type. So, the aim in the second subsection is not to examine the evolution of systems, but to try to establish a causal link between the error factors and the actual errors.

To conclude the analysis, the progress and decline of each system are analyzed diachronically, comparing the number of actual errors by type, to identify some remaining challenges in 2023 GT and DL output.

3. Results and Discussion

The results are divided into three subsections. The global performance of GT and DL from August 2019 to January 2023 is presented in 3.1.; the possible causal link between potential error factors and actual errors is discussed in 3.2.; the progress, decline and remaining challenges are addressed in the last subsection.

3.1. DL and GT from 2019 to 2023

Figure 1 presents the evolution of the global performance of Gt and DL in translating the 167 PT instances in French.

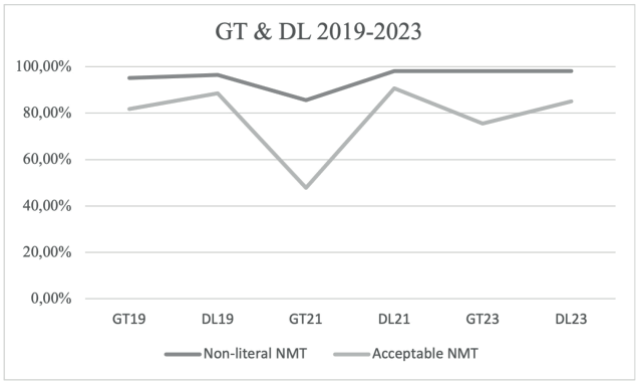

Fig. 1: Global evolution of GT and DL in terms of literality and acceptability performance

In terms of literality, DL improves less than 2% from 2019 to 2021 and stagnates after that. From 2019 to 2023, GT improves nearly 3%, after declining almost 10% from 2019 to 2021. As to acceptability, all scores are lower. The decline of GT in 2021 is much more obvious. Surprisingly, the 2019 scores (GT: 81,74% and DL: 88,62%) are slightly higher than in 2023 (GT: 75,45% and DL: 85,03%). Here are some examples illustrating these results:

(5) Literality

- *Cada vez com maior frequência …* (CTP)
- *De plus en plus souvent, …* (GT23, DL21-23)
- *?... de plus en plus …* (GT19, DL19)
- ** Chaque fois avec une plus grande fréquence …* (GT21)
- *More and more often* / *Increasingly* /... (EN)

In (5), the PT MWU is split. GT23, DL21 and DL23 produce the best output. In 2019, both systems give an acceptable output, i.e. attested in the human translation model, but GT21 proposes an unacceptable calque from PT.

(6) Acceptability

- *uma em cada 5000 gravidezes*.
- *une grossesse sur 5 000.* (GT19, GT23, DL21-23)
- **une femme sur 5 000 grossesses.* (DL19)
- **un dans tous les 5000 grossesses.* (GT21)
- *one in every 5,000 pregnancies.* (EN)

The PT instance in (6) does not present any significant challenge, and none of the proposals contains *chaque*. Nevertheless, DL19 added a second N (*femme*), which results in an unacceptable agrammatical MWU, while GT21 produces a calque from EN (NUM *in every* QP).

The next example illustrates the consistent performance of DL, the decline of GT in 2021, and the better output of GT in 2019 than 2023.

(7) Evolution of acceptability scores

- *em cada dez segundos que passam*
- *toutes les dix secondes* (GT19, DL19-23)
- ?*toutes les dix secondes qui passent* (GT23)
- **dans chaque seconde dix qui passent* (GT21)
- *every ten seconds* (EN)

In (7), GT21 produces a hybrid calque from PT and EN that results in nonsense. Surprisingly, GT NMT is more literal in 2023 than in 2019.

Literality is not a significant challenge any more for GT and DL, but acceptability still is. It is therefore necessary to go deeper into the analysis to identify some obstacles to the production of quality translations of these MWUs.

3.2. Potential error factors in PT and actual errors in FR

Potential error factors are distributed among the 110 instances of the corpus as shown in Table 2:

Potential error factor	Nr
Split instance	22
Idiom in the broadest sense	21
No particular challenge	17
Combination of 2 to 3 factors	16
Atypical PREP (P3S and P2S)	8
AQF (PROG)	7
P3S with two N instead of one	6
Coordinated instance	4
Inversion (P3S and P2S)	3
Long QP	3
Split inversion (P3S and P2S)	2
Ellipsis in coordinated instances	1

Table 1. Distribution of error factors among instances.

As to hybrid error factors, splitting combines with idiom (4 instances), coordination (1), ellipsis (2); inversion combines with atypical PREP (1); split inversion combines with long QP (5) or with two N in P3S and long QP (1); and two N in P3S combines with long QP (2).

Globally, 386 errors were identified in the output of GT19-23 and DL19-23: 155 phraseology issues, 53 calques from EN, 50 calques from PT, 45 nonsenses, 42 wrong meanings, 27 omissions, 12 additions, and only 2 opposite meanings in this corpus.

The average number of any error type per factor was calculated as the number of actual errors divided by the number of instances of each factor. The results are presented in Figure 2:

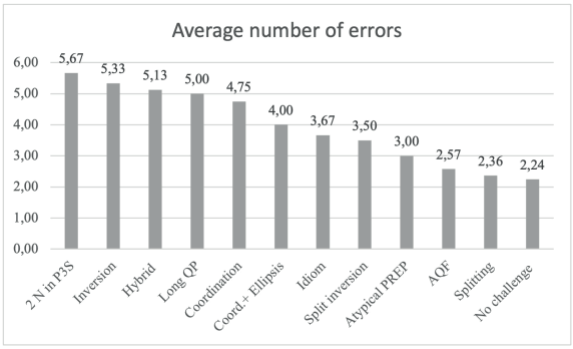

Fig. 2: Average number of errors per error factor instance

According to this chart, GT and DL's performance is higher with more usual structures (no challenge, splitting, AQF), as could be expected. The four strongest error factors provoke mostly phraseological errors, such as agrammatical use of two N in the French P3S output, but calques from PT and EN, nonsense, wrong meanings and omissions are other frequent errors in the case of factor combination (*Hybrid*). Example (8) illustrates the causal link between combination of three potential error factors and actual errors:

(8) Split inversion, 2N in P3S and long QP

- *... em cada mil habitantes há respectivamente 592 e 582 pessoas que compram pelo menos um jornal por dia.* (CTP)
- ?*...avec respectivement 592 et 582 personnes achetant au moins un journal par jour pour mille habitants.* (DL19)
- ?*... avec respectivement 592 et 582 personnes pour mille habitants achetant au moins un journal par jour.* (DL21-23)
- ?*... 592 et 582 personnes achètent respectivement au moins un journal par jour pour 1 000 habitants.* (GT23)
- **... dans tous les mille habitants il y a respectivement 592 et 582 personnes qui achètent au moins un journal par jour.* (GT19)
- **... hors de mille habitants, il y a respectivement 592 et 582 personnes qui achètent au moins un journal par jour.* (GT21)
- *... with 592 and 582 out of every thousand inhabitants who respectively buy at least one newspaper a day.* (EN)

In PT, split inversion of P3S combines with two N (*habitantes*, *pessoas*) and a long QP (*592 e 582 pessoas que compram pelo menos um jornal por dia*) resulting in an unusual structure (*em cada* QP […] QP). The two nearly synonymous N (*personnes* and *habitants*) are systematically translated in FR. GT19 and GT21 give calques from EN (*in every QP* and *out of QP*) that are not attested in the human translation model.

The coordination and ellipsis scores are in the middle. Example (9) illustrates a challenging ellipsis of the ordering element *cada vez* in coordinated instances:

(9) Ellipsis in coordination

- *cada vez mais conflitos e mais violentos", ...* (CTP)
- *des conflits toujours plus nombreux et plus violents", ...* (GT21)
- ?*des conflits de plus en plus violents", ...* (GT19)
- ?*de plus en plus de conflits et de violence", ...* (DL19-23 and GT23)
- *ever more numerous and more violent conflicts...* (EN)

In (9), the PT instance coordinates a N (*conflitos*) with an ADJ (*violentos*), which is unusual in FR. Exceptionally, GT21 gives the best output using a calque from EN attested in the human translation model (*toujours plus*) and coordinating two ADJ (*nombreux* and *violents*) Like *cada vez*, *toujours* can operate on both coordinated elements while *de plus en plus* should be repeated. GT19 keeps an N and an ADJ. It expresses the qualitative progression correctly (*de plus en plus violents*) but omits the quantitative progression of N (*conflits*). The others select two N and produce a wrong meaning.

Idioms are particularly challenging when they combine with humour, as in example (10):

(10) Idiom and splitting

- *Para os noruegueses, isto está cada vez com menos espinhas...* (CTP)
- ?*Pour les Norvégiens, cela devient de moins en moins acnéique ....* (GT19)
- ?*Pour les Norvégiens, cela fait de moins en moins de boutons...* (GT23)
- ?*Pour les Norvégiens, cela devient de moins en moins osseux...* (DL19)
- ?*Pour les Norvégiens, il s'agit de devenir de moins en moins boutonneux...* (DL21-23)
- **Pour les Norvégiens, c'est devient moins et moins de boutons ...* (GT21)
- *For Norwegians, it's getting easier and easier...* (EN)

In its literal sense, the idiom *estar com espinhas* means "to suffer from acne". In informal usage, *sem espinhas* means *easily*, *without problems*. In (10), ambiguity arises from the pun based on these two idioms. All translations are considered as nonsense. Besides, GT21 produces an agrammatical (**est devient moins et moins*) calque from EN (*is becoming less and less*).

The number of actual errors is presented by type in Figure 3:

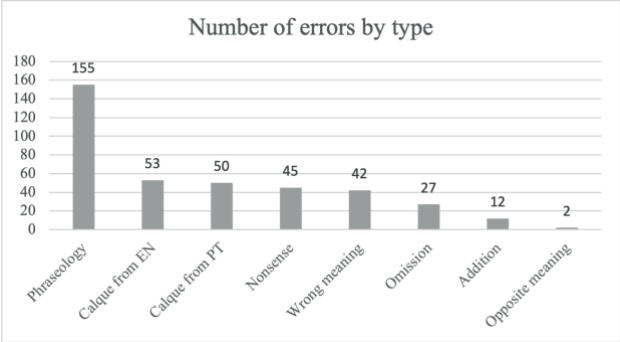

Fig. 3: Number of actual errors by type

Phraseology in the broadest sense includes agrammaticality as in example (8) in the case of DL19-23 and GT23 or GT21's output in example (10). This error type also includes lack of fluency, such as GT23 in (7) or wrong collocations illustrated in example (11):

(11) Phraseology: wrong collocation

- *... quem tem trabalho trabalha cada vez mais.* (CTP)
- ?*ceux qui ont un emploi travaillent de plus en plus fort.* (GT19)
- *... those who have jobs are working more and more.* (EN)

GT19 and DL19 added an ADJ in FR, but the collocation *travailler fort* is very odd whereas *travailler dur* proposed by DL19 is the best output, since GT21-23 and DL21-23 choose the easy correct solution without ADJ (*travailler de plus en plus*) and with a certain meaning loss.

The diversity of phraseology explains its high score. Calques from EN and from PT are on a par (53 and 50). They are illustrated in most of the above examples (5, 6, 7, 8, 10). At the semantic level, nonsense like in (7) and (10) and wrong meaning as in (9) follow with 45 and 42 errors. Omission, addition shown in (6), and opposite meaning are not very significant in this corpus. Here is the only example of opposite meaning in a segment containing two PT MWU instances:

(12) Opposite meaning and omission

- *Há quem, com um certo humor, defina como especialista aquele que sabe cada vez mais de cada vez menos.* (CTP)
- *Il y a ceux qui, avec un certain humour, définissent comme experts ceux qui en savent de moins en moins.* (GT19)
- *There are those who, with a certain humour, define a specialist as one who knows more and more about less and less...* (EN)

In (12), the PT instance is classified as an idiom due to its idiomatic structure (*saber muito de pouco*) and its explicit humorous nature. As with all the others, GT19 supplies the necessary pronoun *en* in FR, but it omits the first MWU (*cada vez mais*), whose co-occurrence with its antonym is unusual. It results in the opposite meaning since the adverbial translated MWU modifies the V *savoir* and does not have any internal argument. DL21-23 produce a perfect solution (*celui qui en sait de plus en plus sur de moins en moins de choses*). DL19 and GT23 propose the same solution for the first MWU but with an adverb instead of the necessary internal argument (NP) in FR (?*celui qui en sait de plus en plus de moins en moins*). GT21 output is nonsense: *Certaines personnes avec une certaine humeur, définies comme un expert qui en sait toujours plus sur un temps moins.*

3.3. Progress, decline, remaining challenges

Progress, decline, and remaining challenges are identified system by system according to error types in Table 2:

	GT19	GT21	GT23	DL19	DL21	DL23
Phraseology	31	42	25	22	19	16
Calque from EN	3	44	2	0	2	2
Calque from PT	8	22	4	6	5	5
Nonsense	4	28	3	4	3	3
Wrong meaning	15	3	4	12	4	4
Omission	9	0	6	4	4	4
Addition	3	5	2	2	0	0
Opposite meaning	2	0	0	0	0	0

Table 2. Progress, decline, remaining challenges.

The phraseology is improving, but it remains a challenge for GT and DL. GT21 can be seen as an unfortunate attempt to improve NMT performance, possibly using EN as a pivot language or poor quality data, so only the evolution between GT19 and GT23 is relevant here. GT23 has fewer errors of

any type than GT19, but fewer acceptable instances (Fig. 1) since error types can combine. Calques from EN are the only slight setback in the progress of DL. As to semantic issues and calques from EN, GT23 and DL23 are very similar, but DL23 outperforms GT23 as far as other error types are concerned, except for calques from PT as in example (13):

(13) Phraseology and calque from PT

- *Cada vez fico mais esclarecido com a instituição com que lido.* (CTP)
- ?*Je suis de plus en plus éclairé sur l'institution avec laquelle je traite.* (GT23)
- **Chaque fois, je deviens plus éclairé sur l'institution avec laquelle je traite.* (DL23)
- *I understand increasingly well the institution I am dealing with.* (EN)

None of the systems proposes an acceptable translation of the idiom *ficar esclarecido com*, which is considered as a phraseological error. Besides, DL23 produces a calque from PT.

(14) Addition

- *Saramago [...] afirma que existe uma analfabetização lenta, que vai minando a área dos alfabetizados, que sabem cada vez menos ler, escrever e «sobretudo pensar».* (CTP)
- *... qui savent de moins en moins lire, écrire et surtout penser.* (DL23)
- ?*... qui savent de moins en moins comment lire, écrire et « surtout penser »*. (GT23)
- *... who increasingly know less about how to read, write and, "above all, think".* (EN)

In this last example, there isn't any specific challenge in PT. DL23 gives a correct output, but GT23 adds *comment*, a calque from EN.

4. Conclusion

The analysis of the small corpus provides some insight into the evolution of DL and GT from 2019 to 2023, it confirms the causal link between atypicality and actual errors, and it identifies some remaining challenges facing machines and humans translating PT MWU expressing PROG, P3S and P2S into FR. Apart from the decline of GT in 2021, 2023 results are encouraging as to literality. Nevertheless, calques from EN are still present in 2023 and calques from Portuguese seem hard to avoid completely, even though the number of word-for-word translations decreases. Phraseology in the broadest sense remains a major challenge in the case of these three MWUs including variables (COMP and QP). These variables are usually short, but challenging, atypical instances were selected on purpose. It is therefore only natural that syntactic, semantic and phraseological errors are more numerous with longer combinations and unusual complex structures that are under-represented in the data since the machine generalizes from the data, without considering unusual MWU structures.

This study only confirms some well-known weaknesses of NMT. As a linguist, I leave it to the engineers to find solutions to the linguistic problems raised here.

References

Bacquelaine, F. (2022a). DeepL et Google Translate face à l'ambiguïté phraséologique. *Journal of Data Mining and Digital* Humanities, 2022, *Towards robotic translation?*. https://doi.org/10.46298/jdmdh.9118.

Bacquelaine, F. (2022b). Traduction d'unités polylexicales du portugais en français par MT@EC et eTranslation. *Revue Traduction et Langues* 21(1), pp. 56-76.

Bacquelaine, F. (2020). Traduction humaine et traduction automatique du quantificateur universel portugais en français et en anglais [unpublished doctoral dissertation]. Faculty of Arts of University of Porto. https://catalogo.up.pt/exlibris/aleph/a23_1/apache_media/FJ5KML8897M4P7HDKXC8RMN37XLAB8.pdf

Koehn, P. (2020). *Neural Machine Translation*. Cambridge: Cambridge University Press.

Leal, A. (2012). Cada vez mais/menos: comparative construction or quantification over eventualities?. In: Schnedecker C., Armbrecht C. (Eds.) *La quantification et ses domaines : actes du colloque de Strasbourg 19-21 octobre 2006*, pp. 355-366. Paris: Honoré Champion.

Santos, D. and Bick, E. (2000). Providing Internet access to Portuguese corpora: the AC/DC project. In: Gavrilidou, M., Carayannis, G., Markantonatou, S. Piperidis, S., Stainhauer, G. (Eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000) 31 May - June 2, 2000, Athens, Greece*, pp. 205-210. European Language Resources Association (ELRA). Online at: http://www.lrec-conf.org/proceedings/lrec2000/pdf/85.pdf. Access date: February 19, 2023.

Soler Uguet, C., Bane, F., Anna Zaretskaya, A. and Tània Blanch Miró, T. (2022). Comparing Multilingual NMT Models and Pivoting. In: Moniz, H., Macken, L., Rufener, A., Barrault,, L., Costa-jussà, M. R., Declercq, C., Koponen, M., Kemp, E., Pilos, S., Forcada, M. L., Scarton, C., Van den Bogaert, J., Daems, J., Tezcan, A., Vanroy, B., Fonteyne, M. (Eds.) *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pp. 231–239. Online. European Association for Machine Translation. https://aclanthology.org/2022.eamt-1.26. Access date: February 19, 2023.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In: Calzolari N., Choukri K., Declerck T., Dogan M. U., Maegaard B., Mariani J., Moreno A., Odijk J., Piperidis S. (Eds.) *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*. Retrieved from: http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf. Access date: February 19, 2023.

Zhang, B., Williams, P., Titov, I. and Sennrich, R. (2020). Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (Eds.) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1628–1639. Online. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.