Evaluation of NMT and SMT systems: A study on uses and perceptions

Emmanuelle Esperança-Rodier, Caroline Rossi, Alexandre Bérard, Laurent Besacier

Translating and the Computer 39

16-17 November 2017
One Birdcage Walk, London

Proceedings

Evaluation of NMT and SMT Systems: A Study on Uses and Perceptions

Emmanuelle Esperança-Rodier
Univ. Grenoble Alpes, CNRS,
Grenoble INP[*], LIG, 38000 Grenoble,
France
Emmanuelle.Esperanca-Rodier@univ-grenoble-alpes.fr

Caroline Rossi
Univ. Grenoble Alpes, ILCEA4,
38000 Grenoble, France
Caroline.Rossi@univ-grenoble-alpes.fr

Alexandre Bérard [†]
Univ Lille, CNRS, Centrale Lille,
UMR 9189 CRIStAL – Lille France
Alexandre.Berard@ed.univ-lille1.fr

Laurent Besacier
Univ. Grenoble Alpes, CNRS,
Grenoble INP[*], LIG, 38000 Grenoble,
France
Laurent.Besacier@univ-grenoble-alpes.fr

Abstract

Statistical and neural approaches have permitted fast improvement in the quality of machine translation, but we are yet to discover how those technologies can best “serve translators and end users of translations” (Kenny, 2017). To address human issues in machine translation, we propose an interdisciplinary approach linking Translation Studies, Natural Language Processing and Philosophy of Cognition. Our collaborative project is a first step in connecting sound knowledge of Machine Translation (MT) systems to a reflection on their implications for the translator. It focuses on the most recent Statistical MT (SMT) and Neural MT (NMT) systems, and their impact on the translator's activity. BTEC-corpus machine translations, from in-house SMT and NMT systems, are subjected to a comparative quantitative analysis, based on BLEU, TER (Translation Edit Rate) and Meteor. Then, we qualitatively analyse translation errors from linguistic criteria (Vilar, 2006) using LIG tools, to determine for each MT system, which syntactic patterns imply translation errors and which error type is mainly made. We then assess translators’ interactions with the main error types in a short evaluation task, completed by participants in the Master's degree in Multilingual Specialized Translation of Grenoble Alps University.

1 Introduction

In a context where statistical and neural approaches have allowed an extremely rapid improvement in the quality of machine translation (MT), we propose an interdisciplinary approach linking Translation Studies, Natural Language Processing (NLP) and Philosophy of Cognitive science, which has three objectives:

- Identify the uses and perceptions of Statistical/Neuronal MT (SMT/NMT) systems in professional translators and trainee translators;

[*] Institute of Engineering Univ. Grenoble Alpes

[†] also at LIG, Univ. Grenoble Alpes

Proceedings of the 39th Conference Translating and the Computer, pages 11–24,
London, UK, November 16-17, 2017. ©2017 AsLing

- Compare these uses and perceptions with the architecture, functioning and effective potentialities of the systems;
- Put these comparisons into perspective with the conceptions of human action and the conceptions of cognition underlying SMT/NMT.

Access to the site is guaranteed because of the involvement of one of the project's members in the Master's degree in Multilingual Specialized Translation at Grenoble Alps University (UGA).

Current research on MT is for the most part carried out in a single disciplinary field, that of Natural Language Processing (NLP). However, some aspects are also covered in Translation Studies, in particular the cognitive ergonomics of post-editing, (inter alia, O'Brien, 2012, Martikainen and Kübler, 2016). Research that articulates good knowledge of the functioning of MT systems and a reflection on their implications for the translator is still very rare. The efforts of P. Koehn (2013 and 2016) or A. Way (2010), to facilitate understanding of current developments and encourage interactions between linguists and computer scientists are remarkable in this respect but they remain exceptional, just like the book by Ehrensberger-Dow et al. (2015), which brings together ten multidisciplinary contributions to advance our understanding of translation processes. Furthermore, to our knowledge, no attempt has been made to anchor these interactions between Translation Studies and NLP in a broader epistemological reflection on the conceptions of language, cognition and action underlying the empirical turn of MT.

Our collaborative project is a first step in filling these gaps. We are interested in the most recent MT systems, based on statistical and then neural models, and the impact of these systems on the translator's activity. The project combines three disciplines. The role of Translation Studies is to identify the uses and perceptions of SMT/NMT in professional translators and trainee translators (i.e. students of the Master's degree in Multilingual Specialized Translation at UGA). The role of Natural Language Processing is to provide thorough knowledge of the internal functioning of the MT systems which will be compared with the representations and uses of the translators. The object of this comparison is to know whether the translators have a vision of the systems that is faithful to their internal functioning and to examine the relation between this vision and their capacity to exploit the potentials of the systems and to know their limits. The role of Philosophy of Cognitive science is to include these questions in broader conceptions. First, we seek to put into perspective the representations of translators and the functioning of systems with the conceptions of human cognition underlying SMT/NMT. Secondly, it will be necessary to articulate the question of uses with a more general conception of human action and its relation to mental states. This broadening of perspective to a more general reflection on human cognition and action is all the more necessary as the deep learning algorithms implemented in recent MT systems (Bahdanau et al., 2014, Cho et al, 2014, Jean et al., 2014) have emerged as a promising conceptual tool for modelling some aspects of linguistic cognition (Dupoux, 2016; Becerra-Bonache & Jimenez Lopez 2016).

The present paper is a case study which puts into perspective the differences between SMT and NMT, combining NLP metrics and error coding with surveys to document perceptions. We have used in-house SMT and NMT systems, to translate documents from the Basic Travel Expression Corpus (BTEC) from French to English. The SMT and NMT used as well as our corpus are described in the second section of this article.

The data collected will be the subject of a comparative quantitative analysis, based on BLEU, Meteor and its empowered version from the LIG (Servan & al, 2016), and TER (Translation Error Rate), and the most often corrected errors will then be analysed more deeply, using LIG tools to perform the analysis of translation errors according to linguistic criteria such as those proposed by Vilar (2006) to determine a set of implemented strategies. The results of those comparisons and analyses are given in the first part of the article's third section.

The second part of the third section is dedicated to our analysis of perceptions of MT in trainee translators (Master's students). We distinguish two stages in the analysis of perceptions. The first deals with students' overall perceptions of MT, based on questionnaires that were answered before and after a 12-hour MT class, as well as on focus group data. Second, we seek to assess students' perception of the differences between an SMT and an NMT system. Metrics are used to convey an objective evaluation of the systems before we discuss students' assessment, in the last subpart.

2 Tools and Corpus

To achieve this study, we needed to perform a detailed comparison of an SMT to an NMT system. While SMT systems are yet quite well known, NMT models are not so obvious to seize, even if we can find a lot of available tools to construct one's own. This is why we are going to describe in greater detail the NMT system we have developed.

Our NMT model is an attention-based encoder-decoder neural network (Sutskever et al., 2014; Bahdanau et al., 2015). LIG implementation, described in Bérard et al. (2016) is based on the seq2seq model implemented by TensorFlow (Abadi et al., 2015). It reuses some of its components, while adding a number of features, like a bidirectional encoder (Bahdanau et al., 2015), a beam-search decoder, a convolutional attention model and a hierarchical encoder (Chorowski et al., 2015). The NMT model uses a decoder with 2 layers of 256 LSTM units, with word embeddings of size 256. Encoder is a 2-layer bidirectional LSTMs, with 256 units. We use a standard attention model. For training, we use the Adam algorithm with an initial learning rate of 0.001 (Kingma and Ba, 2014), and a mini-batch size of 64. We apply dropout (with a rate of 0.5) during training on the connections between LSTM layers in the encoder and decoder (Zaremba et al., 2014).

Turning now to the SMT baseline we use, it is a phrase-based model using Moses Toolkit (Koehn et al., 2007), trained on BTEC train, that represents a 201k words for French, and a 189k words for English), without any monolingual data added, and tuned on BTEC dev of 12.2k words for French, and 11,5k for English.

As a corpus, we have chosen to work on the BTEC (Basic Travel Expression Corpus) which, as described in the BTEC Task of the IWSLT 2010 evaluation campaign1, "[…] is a multilingual speech corpus containing tourism-related sentences similar to those that are usually found in phrasebooks for tourists going abroad". We thought that as the BTEC contains short sentences (10 words/sentence on average), it would be easier and quicker for our students to work on it. We have worked on the translations of BTEC Test 1, which represents 3,9k words for French and 3,6k for English, from our SMT system and our NMT one. We have first proceeded to a trivial empirical evaluation of the output quality of both MT systems based on fluency and adequacy. From the source text, we have given a score to the corresponding output translation, i.e. 1 when the translation was bad (not fluent and/or not adequate), 2 when the translation was average which means that it was adequate and/or fluent, and finally 3, when the translation was good (fluent and adequate). Table 1 below shows an extract of this first manual human evaluation.

1 http://iwslt2010.fbk.eu/

Source Text	SMT output	Evaluation	NMT output	Evaluation
au secours !	help something like	1	help .	2
pouvez-vous nettoyer ma chamber ? …	can you clean my room ? …	2	could you clean my room ? …	3

Table 1: First Evaluation

Once this first evaluation was done, among the BTEC Test 1 corpus, we have selected a total of 50 source sentences along with their corresponding translation using the SMT system and their corresponding translation issued from the NMT system, thus building the so-called BTEC-50 to be evaluated by students in the Master's degree in Multilingual Specialized Translation of Grenoble Alps University.

The selection has been conducted as follows. We have selected the sentences according to the first evaluation results. When the SMT output and the NMT output received contrasted scores, that is to say 1 vs. 3, the source sentence and the SMT and NMT outputs were added to the BTEC-50. Also some of the source sentences for which the system outputs received less contrasted scores were added to the score in order to see which average or bad scores were better accepted by students according to the systems. Finally some sentences for which the SMT and NMT outputs received both a good score, i.e. 3, were added. An overview of the selection done for creating BTEC-50 appears in Table 2.

Source Text	SMT output	Eval.	NMT output	Eval.	BTEC-50
au secours !	help something like	1	help.	2	Yes
pouvez-vous nettoyer ma chambre ?	can you clean my room?	2	could you clean my room?	3	Yes
…	…		…		
c'est trop brillant	it is too brilliant	1	it is too flashy	3	Yes
pouvez-vous me conseiller une bonne boite de nuit ?	can you recommend a good night club?	3	can you recommend a good night for me?	1	Yes
	…		…		
…	I am nauseus		I am nauseus		
j'ai la nausée	…	3	…	3	No
…	where is the service		where is the lost		
où se trouve le service des objets trouvés ?	charge and found?	2	and found?	2	No
	…		…		
	lie down over here and		please lie down		
…	déboutonnez your shirt	2	here and your shirt	2	Yes
allongez-vous ici et déboutonner votre chemise.	…		…		
…					

Table 2: Selection of sentences for BTEC-50 - Examples

Having completed the BTEC-50, we created an Excel sheet, (reproduced in Appendix A), to be given to the students in the Master's degree in Multilingual Specialized Translation in order for them to rank the translated output from 1 very bad to 4 very good. We have decided not to show from which MT system the output were coming, so that the participants could approach the evaluation without prejudice. Nevertheless, we did present the two systems side by side for comparison, so that preferences may appear: the outputs found in the column labelled "EN translation 1" from the Excel sheet come from the SMT system defined previously, whereas the outputs found in the column labelled "EN translation 2" come from the above-mentioned NMT system.

3 Experiment

3.1 Linguistic error analysis

As we said previously, we have performed a first evaluation of the overall quality of SMT and NMT systems. In table 3, we show the results obtained for BTEC-50.

NMT			**SMT**		
1 (bad)	2 (average)	3 (good)	1 (bad)	2 (average)	3 (good)
16	17	17	14	19	17

Table 3: First evaluation

If we look at table 3, we cannot find a real distinction between the results leading us to conclude that both systems give equivalent results.

Looking at scores more in depth, and focusing on the common results, we found out that the NMT and SMT systems obtained 6 times a bad score (1) on the same source sentences, while they got 7 times an average score (2) on the same source sentences and 6 times a good score (3). For any scores given, 1, 2 or 3, when the NMT output obtains the same score as the SMT output, it can be because they provide two outputs that have the same mistakes, see example 1.

Example 1
French source: de rien
SMT translation: * 'anything'
NMT translation: * 'anything'

It also can be that the two outputs provide the same correct translation, as in example 2.

Example 2
French source: je vais prendre la même chose, s'il vous plaît.
SMT translation: ' i'll have the same, please . '
NMT translation: ' i'll have the same, please. '

Or, the two outputs can be two distinct correct translations, as in example 3.

Example 3
French source: est-ce que je dois réserver ?
SMT translation: 'shall I book? '
NMT translation: 'do I have to make a reservation?'

But it can also be two outputs that are different and not corresponding to the source, as shown in example 4 below.

Example 4
French source: avez-vous de la sauce de salade au bleu ?
SMT translation: * 'do you have sauce of salad in blue?'
NMT translation: * 'do you have any chicken salad?'

Having completed the first manual human evaluation, we proceeded to linguistic error analysis, using the error type from the Vilar's (2006) typology. Table 4 below shows the error types, and sub-error types encountered for each system, as well as the number of times that they occur.

	NMT	SMT
Missing Words/Content Words	13	7
Word Order/Word Level/Local Range	0	3
Word Order/Word Level/Long Range	0	1
Incorrect Words/Sense/Wrong Lexical Choice	17	13
Incorrect Words/Incorrect Forms	6	15
Incorrect Words/Extra Words	11	0
Incorrect Words/Style	1	2
Incorrect Words/Idiom	4	5
Unknown words/Unknown Stem	0	12
Punctuation	1	0

Table 4: linguistic error analysis

Again, we cannot find a huge discrepancy between the results of the linguistic error analysis, concluding again that both MT systems are equivalent. Nevertheless we could spot four error types, out of the ten errors encountered, for which there is a significant difference.

A first error type for which we find a difference is the Missing Words with sub-type Content words. This error type is used to label the non translation of a word that appears in the source sentence. The sub-type indicates that the missing word is a word without which the translation cannot be understood. That is to say that the translation of the meaning of a content word, as opposed to filler word, from the source sentence, does not appear in the target sentence. The Content Words error sub-type happens 13 times for the NMT system and only occurs 7 times for the SMT system. Example 5 shows one of those occurrences.

Example 5
French source: c' est le contrat d' achat de mes chèques de voyage.
SMT translation: *' it's the purchase agreement of my checks. '
NMT translation: *' it's the seniority wage system.'

The analysis of this error sub-type can be dealt at the same time as the Unknown Word error type, and especially the sub-type Unknown Stem. This sub-type is used to tag when a source occurrence is not translated and is put as it stands in the translation. The NMT system occurrences of such an error never happen while for the SMT system it occurs 12 times. It can be easily explained by the fact that SMT systems are more likely to reproduce as a translation a word from the source sentence when the system does not recognize the stem as shown in Example 6. At the same time, the core functioning of NMT systems entails a bias among NMT system toward hallucinated translations as there is no linguistic link between “jeux videos” and its translation provided by the NMT system "in fashion". Such things cannot happen with SMT system as it only considers the source.

Example 6
French source: les adolescents japonais aiment les jeux vidéos .
SMT translation: *' the adolescents japanese love electronic vidéos . '
NMT translation: *' japanese teenagers are interested in fashion .'

A second sub-type Incorrect Forms that falls into the Incorrect Words error type. This error type is used to tag mistranslations. The NMT system provided 6 occurrences of this type of error while the SMT system gave 15 occurrences of this type of error. One of the errors, as shown in example 7, is due to the tense use when asking questions. This gap could be explained by the core functioning of the NMT system which is better at lexical diversity.

Example 7
French source: pouvez-vous nettoyer ma chambre ?
SMT translation: *' can you clean my room? '
NMT translation: ' could you clean my room?'

The last error sub-type is also part of the Incorrect Word error type, labelled Extra Words. This error sub-type is used when a word appears in the translation while it does not exist This time, it is the NMT system occurrences of this error that are more numerous, eleven errors, than the ones from the SMT system which do not ever happen! This also can be explained by the core functioning of NMT systems, which use a beam search to enlarge the space of translations in which the system can find more appropriate solutions. Sometimes when the best solution cannot be find, the NMT system goes on and produces a wrong translation of a word from the source or a kind of stuttering of the last word translated, as shown respectively in examples 8 and 9 below.

Example 8
French source: avez-vous un menu ?
SMT translation: ' do you have a menu? '
NMT translation: *' do you have a fixed menu?'

Example 9
French source: je voudrais manger de la vraie nourriture indienne
SMT translation: *' I'd like to have true food indienne'
NMT translation: *' I'd like to eat some food food'

3.2 Assessing students' perceptions

During the course of the Master's degree in Multilingual Specialized Translation, the students were trained on SDL Trados Studio, but they did not integrate MT to their computer-aided translation environment. It is known that students with less experience in working with MT systems are the ones who have the most sceptical perceptions of such systems (see e.g. Koskinen and Ruokonen, 2017: 18). Students from this Master's degree had little experience in working with MT systems. Fourteen out of nineteen had already used an MT system, but when it came to using MT in a professional environment, only one had had this experience in the course of an internship.

The task-based assessment consisted of two timed tasks. The first one consisted in manually correcting MT outputs from two different systems (Google’s NMT versus MT@EC, the MOSES-based SMT engine provided by the European Commission). As for the second task, the whole group had to alternate tasks of translation and post-editing: this was done using a simple word processor and tracking changes. After each task, the students were asked to give their feelings, and what they wrote was collected as a small corpus for perception analysis (Rossi, submitted). It was clear from the corpus that although students figured out that the MT system helped them and speeded them up; this realisation did not significantly impact their primary perceptions.

In order to better assess these perceptions, a series of two 20-question surveys were used to get a contrastive assessment of students' perceptions before and after the course. From those surveys, negative perceptions and fears of MT appeared to have been slightly reinforced by the course, and a positive correlation was evidenced suggesting that fear accounted for lower self- efficacy scores (Rossi, ibid). We concluded that the students' fears needed to be addressed in order to make sure they received proper training with MT and were well-equipped to deal with contemporary translation environments.

However, the perception of loss of control and authorship voiced by students is likely to increase with the current improvement of MT systems. If indeed NMT brings about unprecedented change in the quality of MT outputs, it remains to be seen how students will react. In order to gain insight on the impact of such differences, we started by measuring the differences in our NMT versus SMT corpora, using three distinct evaluation metrics, before asking students to produce broad, comparative judgments on the quality of the translated sentences.

3.3 Evaluation Metrics

We have evaluated the two systems (Bérard et al., 2016) using BLEU, TER and Meteor 1.4 metrics which results are shown in Table 5 hereafter.

Corpus	**NMT**			**SMT**		
	BLEU	**TER**	**Meteor 1.4**	**BLEU**	**TER**	**Meteor 1.4**
Dev	51.56	30.75	40.58	54.35	28.66	43.40
Test1	47.07	33.16	39.73	49.44	32.20	42.07

Table 5: BLEU/TER/Meteor 1.4 mono-reference scores

Results concerning Test1 corpus show that the NMT system gives similar results to the SMT system as regards to the three metrics, which is quite promising as we now know that NMT systems need time to get better. It also confirms the results from the Linguistic error analysis on a smaller set, i.e. BTEC-50.

3.4 Evaluation results from the participants in Master's degree in Multilingual Specialized Translation

At the end of our study, as we have mentioned earlier, the participants from the Master's degree in Multilingual Specialized Translation were sent an Excel file in which they had to evaluate from 1 (very bad) to 4 (very good) the SMT output and the NMT output for the BTEC-50, without knowing which output was provided by which system, thus making sure we were not introducing a bias. However, the students did know they were dealing with MT outputs, and this might have had an impact on their choice of scores. A first set of 16 answers gave us the following results.

From the scores given by each participant for each sentence, we have computed a mean per sentence as well as the related standard deviation. Then we have calculated the mean of all the scores per sentences, per participants. Results are shown in table 6.

	NMT		SMT	
	Mean	**Standard deviation**	**Mean**	**Standard deviation**
BTEC-50	2.310	0.1633	2.166	0.1625
Only most contrasted	2.893	0.7656	1.836	0.1904

Table 6: Participants' evaluation from 1 very bad to 4 very good

Participants have equally judged both system with a mean of 2.166 for the SMT system and one of 2,310 for the NMT system. This means that the participants have evaluated both systems as bad, which is equivalent to a score of 2. This confirms the assumption as well as the results of the perception assessment presented in section 3.2 that students have a negative or low perception of MT systems. Nevertheless, when focusing only on the most contrasted translations, the NMT system is evaluated as almost good, thus increasing its mean, while there is a slight decrease for the SMT mean. On the whole, the experiment returns negative perceptions, regardless of the type of MT, statistical or neuronal, even if the NMT system slightly outpaces the SMT system.

Furthermore, if we look at the standard deviation obtained for each MT system, we can notice than there is almost the same agreement between participants for the SMT outputs (standard deviation of 0.1625) as for the NMT outputs (standard deviation of 0.1633).

Once again, the different evaluation and assessment tend to prove that the NMT and SMT systems are equivalent.

If we now put together the above evaluation results with the linguistic error analysis described in section 3.1, we obtain table 7 below, in which we concatenated the main error types found during the linguistic evaluation along with the evaluation by participants.

Once again, as regards the evaluation from the students, the NMT and SMT systems we have worked on seem to be equivalent. We can notice only for two examples, i.e. example 5 and example 8 a difference going from very bad (1) to bad (2) and from bad to good. From example 5, we can deduce that the hallucinated translation issued from the NMT system was less appreciated by the participants than the missing translation in the SMT output, which still makes sense. Looking at example 8, it is the other way round; it seems that the students were more indulgent with the NMT system than with the SMT system.

4 Conclusion and Perspectives

Even if the study has to be performed on a larger dataset, we can already see that all the experiments have proved that our two systems were equivalent, with only a slight advantage for the NMT system. Nevertheless, we have to take into account the fact that our in-house NMT system at the time of the experiment was at its very beginning and that we now should try with its improved version to see if the promising results we have found here are confirmed or even more conclusive.

Finally, it is worth noting that the human evaluation seems to correlate with the metrics obtained. Performing similar tests on richer data would enable us to see whether there is more to this result than a mere coincidence.

	NMT			SMT		
SOURCE	**Translation**	**Score mean**	**Standard deviation**	**Translation**	**Score mean**	**Standard deviation**
Example 1 - De rien	* 'anything'	1	0	* 'anything'	1	0
Example 2 - je vais prendre la même chose, s'il vous plaît.	'I'll have the same, please. '	3.647	0.5398	'I'll have the same, please. '	3.75	0.4062
Example 3 French source: est-ce que je dois réserver ?	: ' do I have to make a reservation?'	3.533	0.4977	: ' shall I book? '	3.133	0.577
Example 4 French source: avez-vous de la sauce de salade au bleu ?	* 'do you have any chicken salad?'	1.47	0.5536	* 'do you have sauce of salad in blue?'	1.375	0.4687
Example 5 French source: c'est le contrat d'achat de mes chèques de voyage.	*'it's the seniority wage system.'	1.3529	0.4962	*'it's the purchase agreement of my checks. '	2.5625	0.4922
Example 6 French source: les adolescents japonais aiment les jeux vidéos .	*'japanese teenagers are interested in fashion.'	1.187	0.3046	*'the adolescents japanese love electronic vidéos. '	1.6875	0.5156
Example 7 French source: pouvez-vous nettoyer ma chambre ?	'could you clean my room ?'	3.5294	0.4982	*' can you clean my room? '	2.93	0.4680
Example 8 French source: avez-vous un menu ?	*' do you have a fixed menu?'	2.5882	0.6228	'do you have a menu? '	3.375	0.5468
Example 9 French source: je voudrais manger de la vraie nourriture indienne	*' I'd like to eat some food food'	1.4117	0.5328	*' I'd like to have true food indienne'	1.5	0.5625

Table 7: Student evaluation from 1 very bad to 4 very good

Acknowledgements

We would like to thank the "Pôle Grenoble Cognition" for the grant we have received to conduct our interdisciplinary project, and the students who agreed to participate in this research.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, Ł., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org.

Becerra-Bonache, L. & Jiménez López, M.D (2016). Could Machine Learning Shed Light on Natural Language Complexity? Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity, pages 1–11, Osaka, Japan, December 11-17.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR 2015*, pages 3104–3112, San Diego, California, USA.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. CoRR, abs/1409.0473. Retrieved from http://arxiv.org/abs/1409.0473

Bérard, A., Pietquin, O., Besacier, L., Servan, C. Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation. *NIPS Workshop on end-to-end learning for speech and audio processing*, Dec 2016, Barcelona, Spain. 2016. 〈hal-01408086〉

Cho, K., Merrienboer, B. van, Gülçehre, Ç., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. CoRR, abs/1406.1078. Retrieved from http://arxiv.org/abs/1406.1078

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y. (2015). Attention-Based Models for Speech Recognition. In *Advances in Neural Information Processing Systems (NIPS 2015)*, pages 577–585, Montréal, Canada.

Dupoux, E. (2016). Cognitive Science in the era of Artificial Intelligence: A roadmap for reverse-engineering the infant language-learner. CoRR, abs/1607.08723. Retrieved from http://arxiv.org/abs/1607.08723

Ehrensberger-Dow, M., Göpferich, S., & O'Brien, S. (2015). *Interdisciplinarity in Translation and Interpreting Process Research*. John Benjamins Publishing Company.

Jean, S., Cho, K., Memisevic, R., & Bengio, Y. (2014). On Using Very Large Target Vocabulary for Neural Machine Translation. CoRR, abs/1412.2007. Retrieved from http://arxiv.org/abs/1412.2007

Kingma, D. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Koskinen, Kaisa et Minna Ruokonen, Love letters or hate mail? Translators' technology acceptance in the light of their emotional narratives. In D. Kenny (Ed.), *Human issues in translation technology*, Londres et New York, Routledge, 2017, p. 8–24.

Knowles, R., & Koehn, P. (2016). Neural interactive translation prediction. *AMTA 2016, Vol.*, 107.

Philipp Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin,A., and Herbst, E., 2007. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Koehn, P. (2017). *Introduction to Neural Machine Translation*, webinar du 24 janvier 2017, Webinar series by Omniscien Technologies.

Koehn, P. et al. (2013). *Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation* (CASMACAT), Final Public Report. http://www.casmacat.eu/uploads/Deliverables/final-public-report.pdf [consulté le 26 janvier 2017].

Kublička, F., Toral, A., Sanchez-Cartagena, V. (2017) *Fine-grained human evaluation of neural versus phrase-based machine translation*. The Prague Bulletin of Mathematical Linguistics. Available from: https://www.researchgate.net/publication/317304955_Fine-Grained_Human_Evaluation_of_Neural_Versus_Phrase-Based_Machine_Translation

Martikainen, H., & Kübler, N. (2016). Ergonomie cognitive de la post-édition de traduction automatique : enjeux pour la qualité des traductions. *ILCEA. Revue de l'Institut des langues et cultures d'Europe, Amérique, Afrique, Asie et Australie*, (27). Consulté à l'adresse https://ilcea.revues.org/3863

O'Brien, S. (2012) Translation as human–computer interaction. Translation Spaces, Vol. 1(1), pages 101-122.

Rossi, C. (submitted) 'Introducing statistical machine translation in translator training: from uses and perceptions to course design, and back again' *Tradumatica*.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems (NIPS 2014)*, pages 3104–3112, Montréal, Canada.

Vilar, D. Xu, J., D'Haro L. F., et al., 2006. Error analysis of statistical machine translation output. In : Proceedings of LREC. 2006. p. 697-702.

Zaremba, W., Sutskever, I., and Vinyals, O. (2014). Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.