Comparative Analysis of Errors in MT Output and Computer-assisted Translation: Effect of the Human Factor

Irina Ovchinnikova
Sechenov First Moscow State Medical University, Moscow, Russia /
Trubetskaya str., 8, building 2,
119992 Moscow, Russia

Lichi Translations LTD /
P.O. Box 18,
Be'er Ya'akov, 70300 Israel
Ira.ovchi@gmail.com

Daria Morozova
HEC Paris /
1 rue de la Libération,
78351 Jouy-en-Josas, France
daria.morozova@hec.edu

Abstract

The paper presents a comparative analysis of errors in outputs of MT and Computer-assisted Translation (CAT) platforms in translation from Hebrew into Russian. A MT system, shared translation memory (TM), and dictionaries are available on CAT platforms. The platforms allow for editing and improving any MT output as well as performing manual translation. Evaluation of the efficiency of the platforms in comparison with the MT systems shows advantages of the CAT platforms in the translation industry. The comparison reveals the impact of the human factor on the CAT output providing developers with the feedback from translation industry. The research was conducted on documents translated from Hebrew into Russian (approximately 35,000 words, 3118 segments) on Smartcat. Errors in MT output for Russian as a target language show almost equal shares of fluency and accuracy errors in PBSTM and prevalence of the accuracy errors in NMT. Errors on the Smartcat platform reveal difficulties in mastering semantic and stylistic coherence of the whole document. In general, however, the translation is accurate and readable. The influence of English as lingua franca appears in peculiar orthographic and punctuation errors. The errors in translation on Smartcat performed by professional translators uncover insufficiency of CAT tools for the language pair as well as peculiar problems in applying CAT tools while translating from Hebrew into Russian.

1 Introduction

The objective of the study is to analyze the peculiar errors in translation from Hebrew into Russian on a CAT platform as compared to the errors in the MT output and reveal their sources to provide developers with the translators' feedback. The comparison is efficient from the practical point of view since a target text, being translated by a MT system or human translator, must deliver the source message and has to be relevant to the target culture. In the translation industry, revised MT outputs compete with human translations, including those performed on CAT platforms. Therefore, awareness of the peculiar errors in translation on the platforms will provide basis for improving Hebrew-Russian MT and for choosing the way to translate the particular project applying MT or hiring a human translator who has access to a CAT platform.

The research was conducted on the material of translation projects on Smartcat.[1] In the paper, we discuss Hebrew-Russian translation of a tourist guide (9 files in Smartcat; 35,000[2] word forms approximately; 3118 segments; on average, 10 word forms in a segment) by a team of professional translators. The errors in the MT

[1] Smartcat Platform Inc. 2019 https://www.smartcat.ai/

[2] It is hard to determine the exact number of word forms because the author amended the text in Hebrew owing to the necessity to provide the accurate data and information. Nevertheless, the number of segments was constant.

output are considered as a baseline for analysis of errors on Smartcat.

To the best of our knowledge, translation from Hebrew into Russian on a CAT platform was not analyzed in the aspect of the errors type as compared to the MT output. Meanwhile, the comparison allows for developers and users to revise tools on the platforms. Translators will get the insight into specific advantages of MT systems depending on the language pair. Being aware of the advantages, translators can improve the quality of target texts combining MT and human translation.

All tools of computer-assisted translation in one place (CAT platforms) provide translators with an opportunity to quickly deliver a readable and accurate output. The platforms support the cycle of translation projects: selecting translators, teamwork, project management, delivery of the final product, and payment transfer. A CAT platform includes a MT system, access to shared TM, dictionaries, thesauruses and other necessary resources. Collaborators have the opportunity to discuss options, comment on the source text, and share information required to understand the content. Nevertheless, translators and revisers cannot avoid errors while using all of the advantages.

Classification of translators' errors varies according to domains. In the industry, the classification is very simple and pragmatically oriented. In academia, the errors are classified with respect to the target text functioning in the target culture, mental mechanisms of bilingualism and code switching. Since a human bilingual translator operates the platform, the output reveals particular errors. Thus, we take into consideration the classification of translation errors in both domains (See: Hansen, 2009: 316).

2 Related Work: Classification of Typical Errors in MT Output

On Smartcat, a translator can use different MT systems evaluating and editing their output. Thus, we consider the errors distribution for phrase-based statistical and neural MT systems (PBSMT and NMT, respectively).

Distribution of the errors in MT outputs is usually described in the aspect of quality difference between statistical and neural MT systems (Bentivogli, et al., 2016). Human and automatic quality evaluations of outputs of MT systems show different results; however, NMT quality substantially surpasses that of PBSMT (Shterionov et al., 2018).

Researchers differentiate between errors in fluency and accuracy of translation. Fluency errors reduce the readability of the target text, while accuracy errors distort its content. According to the data from different evaluation systems and different languages, fluency errors are more prevalent than accuracy errors (Aranberri et al., 2016). The most typical fluency errors are grammatical errors (close to 80%: Aranberri et al., 2016: 1880). They include morphological, word order and syntax errors. In general, NMT systems outperform PBSMT in fluency (Bentivogli et al., 2016).

The target language affects the kind of morphological information learned by the NMT system. Words of the source text are better represented in a morphologically poorer target language, while a morphologically rich language (e.g., Hebrew and Russian) needs character-based representation of less frequent words in the NMT to enhance the quality of translation (Belinkov et al., 2017). Bilingual post-editors handle the errors in the MT output.

2.1 Errors in Hebrew in MT Output

Hebrew as a source or target language of MT has undeservingly received very sparse researchers' attention. As a morphologically rich language, Hebrew features grammatical affixes, endings and cliticization. The inflections and addition of the subordinate elements to the main word evokes difficulties in processing morphology that were overcome in SMT thanks to pre-processing techniques based on morphological analysis and disambiguation (Singh and Habash, 2012).

In general, NMT outperform PBSMT in Hebrew-Arabic / Arabic-Hebrew translation (Belinkov and Glass, 2016). For better results, Hebrew needs a character-based encoding / decoding model that improves identification of word structure for less frequent words, while words that are more frequent are possible to be identified in the word-based model (Belinkov et al., 2017; Richardson et al., 2016). Nowadays, the most suitable solution for MT translation from Modern Hebrew is Google's Multilingual NMT that involves English as an interlanguage (Johnson et al., 2017). In translation from Hebrew (Modern and Archaic) into English, the omissions and additions occur due to high degree of compression in Hebrew (Cheesman and Roos, 2017: 11).

2.2 Errors in Russian in MT Output

Errors in Hebrew-Russian MT output have not been described or explained in publications. In the translation industry, translators often prefer to apply a Hebrew-English-Russian MT, as in the case of Google's Multilingual NMT. Due to this practice, we need to consider errors in English-Russian MT output. According to human evaluation, NMT English-Russian output received marks "Near native or Native" for 75% segments, whilst PBSMT got the same marks for 60% of segments in the output (Castilho et al., 2017b: 121). The most frequent errors are morphological (42% for PBSMT, 38% for NMT), wrong word order occurs in 12% and 9% of the segments for PBSMT and NMT, respectively (Castilho et al., 2017b: 124).

The distribution of the accuracy errors varies in different domains and genres (Castilho et al., 2017a). The category of accuracy errors includes additions, omissions, mistranslations, and terminology (Burchardt et al., 2017). The class of terminology errors contains wrong choice in terminology, while mistranslations concern general lexicon (Lommel, 2014). In English-Russian output, omissions occur in 12% of segments, equally for NMT and PBSMT; almost the same frequency describes the additions (11% equally for the both) (Castilho et al., 2017b: 124). Meanwhile, mistranslations cover 23% in PBSMT and 30% in NMT (Castilho et al., 2017b: 125). In Russian-English output, PBSMT also outperforms NMT in accuracy of lexical choice (Toral and Sánchez-Cartagena 2017). In translations into Russian as a language with rich morphology, NMT systems lead to less accurate output as compared to the best of PBSMT; the PBSMT contained fewer mistranslations (Castilho et al., 2017b: 125).

2.3 Classification of Errors

In the MT output evaluation, the category of fluency errors includes grammatical (morphological, word order, syntax), orthographic and punctuation errors. The category of accuracy errors contains omissions, additions, mistranslations, and wrong terminology choice. The classification does not account for discourse and pragmatic errors because to detect and prevent these errors, additional tools are needed (Khadivi et al., 2017). A reviser of the MT output evaluates semantic correlation between two segments (the source and the target) and adequacy of the target segment in the aspect of the target language norms and usage.

In general, the target text delivers its message and performs the adequate function in the target culture thanks to the accuracy of its discourse and pragmatic features, and their correspondence to those of the source text. For different target languages, peculiar MT systems were developed to translate English texts of various domains (Specia et al., 2017). Since every source text is semantically coherent and has contiguity, pragmatic purposes, and discourse peculiarities, application of a relevant MT system affects the corresponding quality of the MT output. Meanwhile, the peculiar MT systems do not exist for Hebrew-Russian or Hebrew-English-Russian. Therefore, every Hebrew-Russian MT output needs post-editing in the aspect of its discourse and pragmatic peculiarities.

The discourse and pragmatic characteristics describe the whole document, while the object of the MT output evaluation is a text segment. Thus, the evaluation of the MT output does not consider discourse-pragmatic errors. Eliminating these errors, the evaluation of MT output considers the segment of the target text but skips the evaluation of the correspondence between the source and target messages. Rules for software localization envisage consideration of the discourse and pragmatic issues in the MT output (Specia et al., 2017: 61). The CAT platforms acquire tools for localization of the target text. Therefore, in the evaluation of Smartcat output, we take into consideration all types of errors described in (Hansen 2009: 316). We apply the data of the errors distribution in the MT output as the baseline to consider whether a human translator offers a better option than a raw or even post-edited output of MT systems. Errors and mistakes in translation on Smartcat disclose the value of the human factor as a contributor to the quality of the final product.

3 Results: Description of Errors in Translation on Smartcat

3.1 Working on Smartcat

CAT platforms transform the translators' environment into computer-mediated communication (CMC) with colleagues and customers. In CMC and in the translation industry, English functions as lingua franca. CMC restricts the feedback to comments in a chat window on the platform. Facilitating decoding and encoding, working on a CAT platform exposes a translator / editor / reviser to the effect of text formatting in the working window with segments of the source text. Under the effect, even a competent translator experiences interference of different languages in CMC. Smartcat provides tools for monitoring task performance, navigating in the document, tracking revisions of target segments, and quality assurance. The CAT platform enhances the efficacy of the translator's work, on the one hand; on the other hand, it makes possible the mixed influence of the human factor on the final product: a post-editor revises the output enhancing the target text, although it is an opportunity to miss errors. Translators and post-editors rarely use a particular post editor's tool or environment to identify the errors (Blagodarna 2018: 16). In addition, they often neglect the MT and shared TM in the process of translation (Zaretskaya, Pastor, Seghiri 2015).

3.2 Distribution of the Errors: Comparison between MT and Smatrcat

We analyze the completed translation of a tourist guide that was accepted by the customer as the first draft of the book to be edited by a professional writer. Three professional revisers performed the manual error evaluation. An expert, the professional linguist,[3] annotated the errors. Such expert evaluation of the final product appears to be a common practice in the industry. In the revised Hebrew-Russian Smartcat translation of the tourist guide, 11 segments with various errors include approximately 1080 word forms (3% of the word forms of the source text). The distribution of the errors reflects particular characteristics of the translation and target text revision on Smartcat (See Table 1).

Type	Discourse – pragmatic	Orthographic	Punctuation	Terminology / lexical choice	Grammatical	Omission / Addition
%	40	18	18	14	9	1

Table 1. Errors distribution in the Hebrew-Russian translation on Smartcat (percentage to all errors in the draft).

The distribution differs from that in the MT output for Russian.

1) The most typical of the Smartcat translation failures are the discourse-pragmatic errors. We are not able to compare our data with the volume of the discourse errors in the MT output due to the difference in the errors classification between the industry and academia. Some of the discourse-pragmatic errors are considered as mistranslations in the MT output.

2) In Smartcat, omissions, additions and wrong lexical choice account for 15% of all errors, while in the MT output the accuracy errors occur in 46% of segments for PBSMT and 53% for NMT.[4]

3) Style-shifting usually manifests in a wrong choice of a word from the synset. The errors are represented on Smartcat as a 10% share included in the category of discourse-pragmatic errors. In the MT output, the style-shifting is probably identified as mistranslations. Therefore, the difference in the distributions of accuracy errors between Smartcat and MT could appear less essential.

4) Smartcat output is almost error-free from grammatical errors. Nevertheless, errors in orthography and punctuation diminish the fluency of the target text.

4 Discussion of the Errors in Hebrew-Russian Translation on Smartcat

4.1 Reasons for Errors of Different Types

Even after revisions, **the discourse-pragmatic errors** (unnecessary style-shifting and provocative intertextual associations) occur regularly. The stylistic errors (included in the category of discourse-pragmatic errors) reflect the well-known peculiarities in Hebrew-Russian translation caused by the rich network of synonyms in the Russian vocabulary in comparison with the Hebrew lexicon, and usage of the distinctive syntactic constructions in Russian texts according to the particular style. For example, in the following sentence, official and high literary styles are mixed: *Шахматная держава, национальная и международная, прославившаяся достижениями как в юношеской, таки и во взрослой категориях* (literal translation: *Chess empire, national and international, famous for achievements in both youth and adult categories*). Besides that, the meaning of the lexeme *держава* (*empire*) semantically contradicts the attribute *международная* (*international*). However, the content of the sentence is most seriously damaged by the association generated by *Chess empire*: the phrase associates with *Ostap Bender*, a popular

[3] The expert is Professor, PhD in Russian Linguistics from Saint-Petersburg State University.

[4] See the data in 2.2.

adventurer from Russian satirical novels. The association adds an ironic estimation to the city described as *Chess empire*. The irony ruins the pragmatic purpose of the city guide translation.

The percent of **the orthographic and punctuation errors** is surprisingly high as Smartcat presupposes automatic spelling and grammar checking to prevent the errors. In table 2, we provide examples of the orthographic errors (marked by bold).

Description of error	%	Example
Skipping spaces between words	27	в **концешестидесятых** годов
Overuse of capitalisation	59	Война за **Н**езависимость Израиля
Wrong spelling and misprint	11	На территории центра действуют городская консерватория "Акадма", балетная школа и студия танца, местные ансамбли исполнителей и городские оркестры, а **так же** великолепный **музеей** искусств, известный по всей стране и за рубежом.
Erratum in compound	3	Ультр**а-о**ртодоксальный

Table 2. Description of the orthographic errors in the target text

The **orthographic errors** reveal interference with the English language norms and gaps in technological competence and fluency in the target language. The effect of text formating in the working field of Smartcat appears because of the use of signs preserving the formatting of the source text. The signs mask the space between words, so what is displayed on the work screen is not what will be transferred into the final output in the target text.

The misuse of capitalization shows the effect of the English language norm on the Russian output. In Hebrew, capitalization is not in use. In Russian, the norms usually prescribe to capitalize the first word in compound names of organizations and events. The translator made mistakes under the influence of English as lingua franca.

The two reasons – display of the translated text and the influence of English as lingua franca – explain 86% of the orthographic errors on Smartcat. Another 11% of orthographic errors are caused by gaps in the translator's target language competence.

Similar reasons cause the **punctuation errors**. Under the influence of the English language, translators overused commas (,) after complements in the beginning of the sentence and often use a colon (:) instead of an em dash (–). Due to the signs of text formatting on the platform, translators miss marks in compound sentences. Almost 30% of the errors show insufficient competence in Russian punctuation norms.

Terminology and lexical errors in Smartcat are similar to those in MT; they reveal misunderstanding of terminology and wrong lexical choices. For example, instead of *блуждающие пески* (*wondering sand*) the translator used *зыбучие пески* (*quicksand*). The most typical of the lexical choice errors concerns wrong selection within the synset ignoring collocations and semantic restrictions. MT systems outperform human translators in the lexical choice associated with peculiar semantic restriction. For example, to refer to people or other entities in Russian, speakers need to choose between two different words; *имя* (*name*) is appropriate only for people, while objects are referred to by their *название* (*name*). NMT systems are able to process the semantic difference offering the relevant Russian word in Hebrew-English-Russian translation.

The **grammatical errors** are akin to those in Russian colloquial speech. Translators and a post-editor recognized the specific errors similar to those in the MT output, but they sometimes failed to identify word forms and idioms that belong to official style, which is irrelevant for the tourist guide. The most typical of the grammatical errors belong to the morphological class when a wrong inflection generates wrong syntactic dependencies in long clauses. In addition, adverbial participles regularly occur in impersonal sentences that is prohibited in Russian: *Проведя* (Adv. Participle-past-perf.) *время в парке, рекомендуется* (Verb-pres.-imper.-impersonal) *продолжить прогулку в южном направлении по прекрасной прогулочной дороге* (literal translation: *After spending time in the park, it is recommended to continue walking in the south direction along the beautiful walking road*).

Translating into Russian, professional translators attempt to shorten target segments and sometimes this leads to omissions (Kunilovskaya, Morgoun, Pariy 2018). Meanwhile, omissions and additions in the MT output from Hebrew appear due to a concise character of the language (Cheesman, Roos 2017: 11). The omission, as well as the addition, can be useful for semantic coherence of the whole document as means to avoid repetition in contact segments and establish cohesion for distant segments. Thus, an omission of information in the process of Hebrew-Russian translation represents an error in accuracy in the segment, but can be purposeful in the whole text perspective. Nevertheless, in the human-revised Hebrew-Russian translation on Smartcat, some of the omissions and additions lead to the distortion of information in the target segment: *Исследователи приводят два возможных варианта жителей крепости, руины которой находятся на холме* (literal translation: *The researchers raised two possible options of the inhabitants of the fortress whose remains were found on the hill*). In the source segment in Hebrew, the author mentioned two different theories explaining the origin of the fortress inhabitants.

4.2 The Human Factor as the Ground for Errors on Smartcat

In summary, orthographic and punctuation errors reveal insufficient command of the CAT tools and gaps in the target language competence of translators. On the one hand, it is necessary to train skills to master CAT beforehand; on the other hand, due to the errors, CAT platform developers can foresee particular problems of implementing text-formatting instruments in the platform. The orthographic and punctuation errors uncover interlanguage interference and impact of English in Hebrew-Russian translation as English strongly affects CMC (Jiménez-Crespo 2010). The discourse-pragmatic errors are caused by neglecting the target language usage, the target cultural context and the purpose of the text (the message itself). Alongside with lexical errors, they break the contiguity of the target text and its semantic coherence.

Compared to the errors in the MT outputs, the translation errors on the CAT platform disclose a skillful mastery of the target language grammar and more accurate lexical choice. However, the MT provides post-editors with the translation that is almost free of orthographic errors. Smartcat improves the technological environment for translators and overcomes disadvantages of MT thanks to the opportunity to use different tools according to the particular source segment.

4.3 Errors Associated with Design of CAT Platform

The source text segmentation and working window formatting on the CAT platform provoke difficulties in expressing the coherence and the anaphora resolution in distant semantically coherent segments. The problems are similar to those that occur in the MT output. Incorrect use of pronouns can be recognized in the process of post-editing the target text.

Peculiar errors reveal the problems associated with the source text segmentation into sentences. This can trigger a translator to preserve the sentence boundaries and use a complicated Russian compound sentence leading to punctuation errors.

5 Conclusion

Our study of the set of errors in Hebrew-Russian translation on the CAT platform found that CAT platforms provide users with good translation quality. The quality is better than the MT output for this pair of languages. The negative impact of the human factor is associated with the mismatch of the capabilities of the CAT tools and the degree of their use by translators. Our analysis found that the particular errors are caused by the effect of English as lingua franca in the translation industry and CMC. These errors diminish the fluency, while the discourse-pragmatic errors decrease the accuracy of the target text. In this aspect, the translation on the Smartcat is similar to the NMT output for Russian in that the fluency of the target text is better than the accuracy. The discourse-pragmatic errors are not recognised in the MT output evaluation because the contiguity of the whole text does not appear as an object of the MT quality evaluation. By combining human competence and computer tools, translation on the CAT platforms enables acceptable translation quality to be quickly generated.

The comparison of errors in MT and on the CAT platform for the Hebrew-Russian language pair provides a basis for training MT systems to achieve the acceptable quality. The distribution of the errors in translation on Smartcat shows the direction for translator and post-editor training. These results are also of importance to developers of CAT platforms as enhancement of user interfaces considering the human factor-triggered errors might contribute to greater accuracy and efficiency of translations.

References

Aranberri, Nora, Eleftherios Avramidis, Aljoscha Burchardt, Ondřej Klejch, Martin Popel, and Maja Popovic. 2016. Tools and Guidelines for Principled Machine Translation Development. *Proceedings of the 10th Conference on Language Resources and Evaluation* (*LREC*). Portorož, Slovenia 1877–1882.

Belinkov, Yonatan, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL).* Vancouver, Canada 861–872.

Belinkov, Yonatan, and James Glass. 2016. Large-scale machine translation between Arabic and Hebrew: Available corpora and initial results. *Proceedings of the Workshop on Semitic Machine Translation*, Austin, Texas, USA 7–12.

Bentivogli, Luisa, Arianna Bisazza, Mario Cettolo and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, USA 257–267.

Blagodarna, Olena. 2018. Insights into post-editors' profiles and post-editing practices. *Revista Tradumàtica*. *Tecnologies de la Traducció*, 16: 35–51.

Burchardt, Aljoscha, Vivien Macketanz, Jon Dehdari, Georg Heigold, Jan-Thorsten Peter, and Philip Williams. 2017. A linguistic evaluation of rule-based, phrase-based, and neural MT engines. *The Prague Bulletin of Mathematical Linguistics*, 108(1), 159–170.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017a. Is neural machine translation the new state of the art?. *The Prague Bulletin of Mathematical Linguistics*, 108(1), 109–120.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Panayota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Barone, and Maria Gialama, M. 2017b. A comparative quality evaluation of PBSMT and NMT using professional translators. *Proceedings of Machine Translation Summit XVI, vol. 1: Research Track*. Nagoya, Japan 116–132.

Cheesman, Tom, and Avraham Roos. 2017. Version Variation Visualization (VVV): Case Studies on the Hebrew Haggadah in English. *Journal of Data Mining and Digital Humanities, July, 5, 2017, Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages*. 1–12.

Hansen, Gyde. 2009. A classification of errors in translation and revision. *CIUTI-Forum 2008: Enhancing Translation Quality: Ways, Means, Methods.* Peter Lang, Bern, Switzerland. 313–326.

Jiménez-Crespo, Miguel A. 2010. Localization and writing for a new medium: a review of digital style guides. *Tradumàtica: traducció i tecnologies de la informació i la comunicació*, 8, 1–9.

Johnson, Melvin, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Deanet. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, vol. 5, 339–351.

Khadivi, Shahram, Patrick Wilken, Leonard Dahlmann, and Evgeny Matusov. 2017. Neural and Statistical Methods for Leveraging Meta-information in Machine Translation. *Proceedings of Machine Translation Summit XVI, vol. 1: Research Track*. Nagoya, Japan 41–54.

Kunilovskaya, Maria, Natalia Morgoun, and Alexey Pariy. 2018. Learner vs. professional translations into Russian: Lexical profiles. *Translation & Interpreting*, 10(1), 33–52

Lommel, Arle R., Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Revista tradumàtica: traducció i tecnologies de la informació i la comunicació*, 12, 455–463.

Richardson, John, Taku Kudo, Hideto Kazawa, and Sadao Kurohashi. 2016. A generalized dependency tree language model for SMT. *Information and Media Technologies*, 11, 213–235.

Shterionov, Dimitar, Riccardo Superbo, Pat Nagle, Laura Casanellas, Tony O'Dowd, and Andy Way. 2018. Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation*, 32(3), 217–235.

Singh, Nimesh, and Nadir Habash. 2012. Hebrew morphological preprocessing for statistical machine translation. *Proceedings of The European Association for Machine Translation (EAMT12)*, Trento, Italy 43–50.

Specia, Lucia, Kim Harris, Frederic Blain, Vivien Macketanz, Aljoscha Burchardt, Inguna Skadina, Matteo Negri, and Marko Turchi. 2017. Translation Quality and Productivity: A Study on Rich Morphology Languages. *Proceedings of Machine Translation Summit XVI, vol. 1: Research Track*. Nagoya, Japan 55-71.

Toral, Antonio, and Victor M. Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, vol. 1: Long Papers*, Valencia, Spain 1063–1073.