Multidimensional assessment of the eTranslation output for English–Slovene

Mateja Arnejšek
European Commission
Directorate-General for Translation
Slovene Language Department
mateja.arnejsek@ec.europa.eu

Alenka Unk
European Commission
Directorate-General for Translation
Slovene Language Department
alenka.unk@ec.europa.eu

Abstract

The Slovene language department of the European Commission Directorate-General for Translation has always been an early adopter of new developments in the area of machine translation. In 2018, the department started using neural machine translation produced by the eTranslation in-house engines. In 2019, a multidimensional assessment of the eTranslation output for the language combination English–Slovene was carried out. It was based on two user satisfaction surveys, an analysis of reported errors and an ex post analysis of a sample. As part of the assessment effort, a categorisation of errors was devised in order to raise awareness among translators of the potential pitfalls of neural machine translation.

1 Machine translation in DGT

eTranslation[1] is one of the Building Blocks for a Digital Connected Europe in the framework of the Connecting Europe Facility (CEF).[2] It was launched in November 2017 with the progressive addition of engines for different language combinations. eTranslation took over from MT@EC, which had been fully operational since June 2013. MT@EC was a statistical machine translation (SMT) system based on the MOSES open-source translation toolkit.[3] The Directorate-General for Translation (DGT) of the European Commission had developed MT@EC under the Interoperability Solutions for European Public Administrations (ISA) programme with co-funding from EU research and innovation programmes. CEF eTranslation followed the field's move into neural machine translation (NMT).

DGT is organised into language departments (LDs), one for each official language of the EU.[4] Right from the launch of the NMT engines, LDs were provided with practical guidelines that aim to ensure that machine translation is used consistently and effectively within DGT, encouraging translators to at the very least try using machine translation, but still allowing for different approaches to cater to specific needs. Training has also been organised to present the new technology and its known general pitfalls. Based on the guidelines and the training, the LDs adopted different approaches to the uptake of NMT and used it in different ways and to differing extents.

In autumn 2018, after the initial period of introduction, uptake and testing, DGT decided to assess NMT output in the LDs, gathering general opinions on how useful neural engines are for the individual LDs and on the kind of impact these engines can have on the efficiency and quality of translation. The objectives of the exercise were to check which of the two engines, NMT or SMT, was preferred as the default engine in the automated pre-processing of translation requests and what the translators should be aware of when using NMT. It also aimed to promote machine translation among users. Since the quality of machine translation output varies depending on the target language, each LD had to carry out the assessment individually, following broad pre-set guidelines.

[1] https://webgate.ec.europa.eu/etranslation.

[2] CEF is a key EU funding instrument to promote growth, jobs and competitiveness through targeted infrastructure investment at European level.

[3] Koehn et al. (2007).

[4] Irish is an exception since it is not yet a fully-fledged department in terms of the number of translators.

2 Machine translation in SL LD

The Slovene Language Department (SL LD) of DGT has always been an early adopter of new technologies, including machine translation. In an online survey conducted in November 2012, presented by Leal Fontes (2013), the majority of users of the English–Slovene machine translation of the time (SMT from MT@EC) already responded that they used machine translation for around 50 % of their translation jobs.

By the time eTranslation was launched, MT@EC had been regularly and extensively used. Following the announcement of the launch of the English–Slovene NMT engine on 4 April 2018, interested members of the department started testing NMT output. After the initial DGT training on NMT on 25 April 2018, the SL LD on 29 April 2018 invited all members of the department to test NMT. Three months afterwards, a survey was carried out in the SL LD that showed that the use of NMT was widespread and preferred to the use of SMT. Consequently, the decision was taken to switch to NMT as the default output prepared automatically for every translation request, as of 5 July 2018 for a trial period of three months. After those three months, a new user survey was carried out that confirmed that users were satisfied with NMT and the use of NMT as the default engine was confirmed on a permanent basis.

The method and extent of machine translation use has always been left to the discretion of the translators. They can include machine translation as one of the reference memories in their CAT tool (with a 25 % penalty) in their translation projects. They can use this machine translation as a typing aid (based on an autocomplete functionality), look up the machine translation results in concordance searches, decide to use individual segments and post-edit them, or opt for a combination of these methods. It is also possible to pre-translate the whole document using machine translation and post-edit the result, but such use has not yet been recorded in the department.[5]

The typing aid approach and concordance use are sub-segment-based types of MT use. The translators using NMT in this way use only limited phrases from the machine translation output at a time. The typing aid approach follows the push principle, as the autocomplete suggestions are automatically shown while the translator is typing. The concordance use, on the other hand, applies the pull principle, as the translator needs to select a phrase and launch a search. In neither of these types of use does the translation file contain any record of the machine translation origin of the used phrases.

If, however, the translator uses entire NMT segments and post-edits them, either by recalling them from NMT output individually or in the hypothetical case of pre-translating the document with NMT output, the metadata of the segment in the translation file registers machine translation as the starting point of the translation, regardless of whether the NMT segment has been edited or to what extent.

Since the first user survey, department members were invited to report any examples of very noticeable or repetitive errors. Their contributions were gathered into a list, along with possible causes and explanations that could serve as useful tips to users when working with NMT. To ensure a more objective and comprehensive insight into the usefulness of EN–SL NMT to feed into the DGT-wide assessment of NMT, the department also carried out an ex post analysis to check the quality of NMT in April 2019. The three-step exercise that included gauging user satisfaction (see section 3), an analysis of reported errors (see section 4) and an analysis of a sample (see section 5) allowed for a broad and thorough assessment of the EN–SL eTranslation NMT.

3 User satisfaction

3.1 Summer 2018 survey

The first user satisfaction survey was carried out at the meetings of the two SL LD units (29 June 2018 in SL.1 and 2 July 2018 in SL.2). At that moment, both units combined had 51 active translators (and 2 trainees), 39 of whom attended the two meetings and 35 responded to the survey.

It transpired that the use of NMT was already widespread in the department at that point (only 3 respondents out of 35 did not use NMT and 1 respondent reported that their use of NMT depended on the type of document). Nearly all respondents also preferred NMT to SMT; in fact, all translators who used NMT liked it better than

[5] Lesznyák (2019) reports that also in the Hungarian LD the translators apply divergent practices to integrate NMT into their workflow.

SMT except for 1 respondent, who answered that they did not notice any difference between SMT and NMT.

I use NMT …		I don't use NMT
and I prefer it to SMT	and I prefer SMT	
SL.1: 14* SL.2: 18** Total: 32	SL.1: / SL.2: / Total: /	SL.1: 3 SL.2: / Total: 3

*1 uses NMT depending on the type of document
**1 uses NMT, but doesn't see the difference

***Table 1**: Summer 2018 survey in SL LD*

The only purpose of the survey was to determine if it would make sense to switch to NMT as the default machine translation product in the automated pre-processing of translation requests. It merely gauged the uptake of NMT at the time and the first impression of how it compares to SMT. There were no questions about the usefulness or quality of machine translation.

The results of the survey showed that at least 32 members of the SL LD were already using NMT, even though it was not part of the automated workflow (at that time SMT was provided automatically with every translation request). This meant that they were regularly requesting the NMT output manually via the eTranslation portal. Therefore, the decision was made to switch from SMT to NMT in the automated pre-processing for a trial period of three months.

3.2 Autumn 2018 survey

A new user survey was carried out in November 2018, this time online, among all 55 active translators in the department. Its purpose was to check the state of play and whether the department would continue using NMT as the default machine translation product.

Out of the 43 respondents, a majority (29 respondents) used machine translation with every or almost every translation and another 7 respondents used it often. Only 2 respondents rarely used machine translation and just 1 replied that they never used it. All 43 respondents (even those who rarely or never used machine translation) expressed their preference for NMT over SMT in pre-processing. The use of NMT as the default machine translation product was confirmed. Again, there were no questions regarding the quality of machine translation as such.[6]

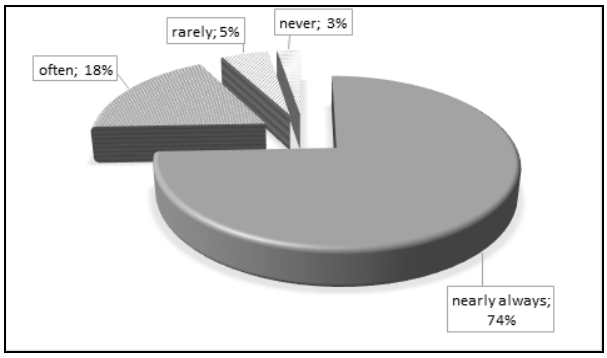

***Chart 1:** Autumn 2018 survey in SL LD*

When asked to compare SMT with NMT, translators said that SMT might be more (terminologically) consistent, but had more incorrect inflections and the sentence structure followed the source language more closely rather than adapting to the structure of the target language, which is consistent with the findings of Toral and Sánchez-Cartagena (2017). In some instances, following the structure of the source language too closely resulted in awkward wording, wrong theme–rheme structure and, in the worst cases, mistranslation. This meant that translators had to edit the text heavily. Furthermore, although the mistakes were more obvious and therefore might be easier to spot and correct, the translators might still miss some mistakes simply due to their frequency and introduce new errors in the process of correcting the text. This is consistent with the findings of Lacruz et al. (2014) that transfer errors (which are those that require the MT user to review the source text to understand the meaning) generate a greater cognitive demand than mechanical errors (which are those that can routinely be fixed without reference to the source text), translators noted that NMT errors might require more attention and concentration as they are not easy to spot. The trust placed in NMT is also generally higher than that accorded to SMT. Both of these factors might lead to oversights.

[6] Lesznyák (2019) reports a different approach in the Hungarian LD: structured interviews with the majority of the translators were carried out from June 2018 to January 2019 (which is approximately the same period as when the two surveys in the SL LD took place) to gauge the translators' views on NMT in more detail than the SL LD surveys did, and the interviews also served to elicit typical errors and quality issues.

4 Analysis of reported errors

Since the introduction of eTranslation NMT in the department, translators were invited on several occasions to report striking or repetitive errors in the NMT output. We gathered them in a list with each assigned a possible cause or plausible explanation as to its origin. The purpose of the exercise was to provide translators with useful tips for working with NMT. Over time, the list has been split into sections for different categories of NMT issues. These share many similarities with the findings of Van Brussel et al. (2018) for English–Dutch NMT:

- NMT includes polysemic misinterpretations (e.g. "swings" translated as "gugalnice", so in the meaning of playground swings instead of statistical fluctuations, which would translate as "nihanja"). This phenomenon invariably produces terminological errors in NMT output.

- NMT includes complete semantic blunders. Some of these erroneous translations seem to come from mix-ups of similar-looking words. The similarity might exist either in the source language (e.g. "skill", which should have been "spretnost" or "veščina", translated rather as "ubijanje", which means "kill(ing)"; or "Slavery" (capitalised) translated not as "suženjstvo" but as "slovanski", meaning "Slavic"). A similar type of error is presented by Van Brussel et al. (2018), who analysed NMT for English–Dutch, with the example of "grace" ~ "graze". The similarity might also exist in the target language (e.g. "moving away" translated as "odmiranje", which back-translates as "dying off", instead of "odmikanje", meaning "distancing"). Some semantic mistakes at least remain in the same field (e.g. "Cypriot" translated as "evropski", meaning "European"). Some of these errors, however, are utterly perplexing to the human user as their origin remains unclear (e.g. "hosting" in technological context translated as "počitek", which back-translates to "rest").

- NMT output includes "neural neologisms", a new type of lexical mistake in NMT output, evidenced also by Macken (2019), consisting of words that do not exist in the target language, invented by the NMT engine. It seems that when the engine encounters words not included in the data sets used in its training, the machine starts inventing translations based on statistically probable patterns, creating nonsensical words. "Reformed" Protestant Church, for example, was not translated as "Reformirana", but as "retvorna". Apparently the elements of the word in the source text were identified ("re" + "formed"), the second element was translated into "tvoren" (which could make sense as a Slovene translation) and the feminine gender was applied ("retvorna"). "Becoming a Maltese citizen", on the other hand, was translated as "popolanje malteškega državljana", where "popolanje" is a non-existent word with no apparent semantic associations with (any meaning of) becoming, but still with the expected gerundial ending. A document with religious vocabulary was especially rife with such neural neologisms (e.g. "Chief Rabbi" as "Chiebi Rabi" instead of "glavni rabin"), probably because such vocabulary is not very common in the material on which NMT engines were trained.

- NMT does not handle proper nouns well as they tend not to be repeated in the data on which NMT is trained. The engine changes them or tries to translate them (e.g. "KRIEGER" becomes "KRIMEGER" and the city of "Christchurch" is translated as "božična cerkev", which back-translates as "Christmas church"). Van Brussel et al. (2018) also note this feature of NMT of translating proper names.

- NMT has trouble translating abbreviations (e.g. "PM" translated not as "predsednica" – feminine form is needed, because it referred to Theresa May – but as "podpredsednica", correctly in feminine form, but meaning "Madame Vice President").

- NMT omits some words (e.g. elements in enumerations), sometimes in such a way as to produce the opposite meaning in the translation (e.g. omitting the negative particle, i.e. "not" or similar). Sometimes it omits also whole parts of sentences, which reflects the finding of Van Brussel et al. (2018) that in omission errors NMT omits more words than SMT.

- NMT sometimes adds words or elements in a sentence – or repeats some of the elements (e.g. "the Portuguese-Spanish border" translated as "portugalsko portugalsko-španska meja").

- NMT has trouble with structures with implicit relationships between words, such as long compound noun phrases. "Short-term travel possibilities", for example, was translated as "kratkoročne možnosti potovanja" – "short-term possibilities for travel" instead of "možnosti za kratkoročno potovanje" – "possibilities for short-term travel". A similar problem occurs with

structures involving explicit relationships between words, misrendering syntactical relations, sometimes even producing the opposite meaning (e.g. "requirements on SMEs" translated not as "zahteve za MSP" but as "zahteve MSP", which back-translates as "requirements of SMEs").

• NMT can have problems with bulleted content. This might be caused by the structure in the bullet points deviating from the structure of normal text. The bulleted items are text fragments and not syntactically complete sentences. Furthermore, the engine might be attempting to translate the bullet symbol (e.g. "State of the Union Address 2016 …" translated as "barbara, govor o stanju v Uniji 2016 …" – translation of "–" ("bar"?) as "Barbara"?).

• NMT has trouble translating misspelt words, which does not come as a surprise (e.g. "Strenghtening" translated with a neural neologism "Strengnetenje", creating a Slovene-sounding gerund with a non-existent stem). What is surprising is that it sometimes manages to translate incorrectly spelt words correctly (e.g. "TheDirective" translated as "Direktiva") – this is something that SMT was not able to do.

• NMT sometimes reproduces spelling mistakes. An error was reported, and we discovered that it originated in a spelling mistake in the data set used to train the NMT engine ("audiovisual" translated as "avdiovizulani" instead of "avdiovizualni"). There was only one spelling mistake in the data, but the error was reproduced several times in the NMT output. This highlights the importance of the quality of data that is used to train NMT engines.

5 Analysis of a sample in April 2019

5.1 Sampling

The sample on which the eTranslation English–Slovene NMT output was to be checked was chosen from among the English-to-Slovene translations of the SL LD. Documents were selected to include legislative, non-legislative and general public documents from different domains.

The documents were chosen from the finalised translation requests received in the department between 1 September 2018 and 31 March 2019. Finalised requests were needed so that a final human translation to which NMT could be compared would be available. The selected period corresponded to the period when NMT was automatically available to translators, in order to facilitate the process of the ex post review (comparing the NMT output to the final translation).

Although the NMT output used for the analysis was actually the one available to the translators at the time of translation, we did not attempt to find out if or how translators had used NMT. Without disproportionate further efforts, it would be difficult to ascertain whether NMT was indeed used in the translation of these documents. Moreover, it would be almost if not completely impossible with the means and tools available to determine in which parts of the text it was used. Checking the segment attributes in xliffs would reveal if machine translation was used, but only if it was used as the starting point of translation. However, machine translation can also be used as a typing aid (based on the autocomplete functionality of the CAT tool) or in concordance searches, and the xliff does not register this.

We only chose documents translated in-house and with a low match rate. The quality of NMT is most relevant in translating documents with a low match rate. If there are no other sources, there is a greater need for machine translation and a greater probability that translators will use it. In addition, this minimises interference from other sources. However, as Lesznyák (2019) points out, NMT for documents with a low match rate might be less useful, as the low match rate indicates that there might have been less material in the databases on which the NMT engines were trained than would be the case of documents with a higher match rate.

Type	Document title
Legislative	Commission Implementing Regulation entering the name 'Havarti'(PGI) in the register of protected designations of origin and protected geographical indications
	Proposal for a Council Decision on the position to be taken on behalf of the European Union in the World Customs Organization in relation to the Harmonised System
	Annex to the Commission Delegated Regulation supplementing ITS Directive 2010/40/EU of the European Parliament and of the Council with regard to the provision of cooperative intelligent transport systems

Non-legislative	Communication from the Commission to the European Parliament, the European Council, the European Central Bank, the European Economic and Social Committee and the Committee of the Regions – Capital Markets Union: progress on building a single market for capital for a strong Economic and Monetary Union (CMU report)
	Report from the Commission to the European Parliament and the Council – EU and the Paris Climate Agreement: Taking stock of progress at Katowice COP (Climate action progress report)
	Replies of the Commission to the Special Report of the European Court of Auditors "The control system for organic products has improved, but some challenges remain"
General public	Citizens' Dialogues and Citizens' Consultations Progress report
	Questions and answers on the EU list of non-cooperative tax jurisdictions
	Health-EU Newsletter 223 - link group
	MEMORY GAME - Match the flags!

Table 2: *Sampled documents*

We selected two consecutive pages in each of the ten selected documents, which produced a sample of 20 pages.[7]

5.2 Methodology

Each document section in the sample was pre-translated using the automatically provided eTranslation NMT and the result was compared (using the Microsoft Word compare function) to the final, human translation, which is considered the gold standard also to Maučec and Donaj (2020) in all types of evaluations of MT quality. The SL LD quality officer[8] checked the comparison file and inserted comments for the differences that amounted to errors according to the DGT and SL LD standards. The annotated differences were labelled with the error categories found in the error grid used in DGT for the evaluation of freelance translations, so as mistranslations, omissions, or errors relating to terminology, reference documents, clarity, grammar, punctuation or spelling; all of them further classified as minor or major errors.[9, 10] Further observations were added to the labels to make it easier to draw conclusions. Additionally, to gain a better understanding of the gravity of errors in NMT, the categorised errors were fed into the internal quality assessment tool to see where NMT ranks compared to human translations.

As the analysis was based on the differences between NMT and the final translation, potential mistakes in NMT that were present in the final translation as well were not detected (false negatives). Differences between NMT and the final translation where there was no mistake in NMT were disregarded (false positives). As the quality of the final translation itself fell outside the scope of this exercise, any errors in the final translation were also disregarded, whether these were caused by NMT or not.

5.3 The result

More examples for already identified error categories

The analysis confirmed the occurrence of errors that had already been reported by users (see section 4). For certain categories of errors in the eTranslation NMT output, the examples found during the ex post review offered a greater overview:

- NMT output contains terminological errors, with generic words instead of terms. For example, in a document on Protected Geographical Indication, "opponents" were translated as the generic "nasprotniki", which back-translates as "adversaries", "antagonists", instead of the domain specific "vložniki ugovorov", meaning persons who oppose the registration of a designation.

- NMT output contains wrong terms. For example, in an act on the service of judicial and extrajudicial documents, "service" was translated in the economic sense "storitev" instead of as "vročanje" – "delivery".

- NMT can be highly terminologically inconsistent. For example, "operators" within one document translated as different valid terms, but

[7] Pages were counted as standard DGT pages, defined as 1500 characters without spaces.

[8] For the role of quality officers, see Drugan et al. (2018).

[9] For more information on the error grid and other elements of evaluation of freelance translations, see Strandvik (2017).

[10] At the time of this writing, DGT is preparing for new outsourcing contracts with a new error categorisation based on multidimensional quality metrics (MQM).

stemming from other areas of regulation; first as "nosilci dejavnosti", which verbatim back-translates as "activity holders", and in the next sentence as "izvajalci" – "implementers", when they should both have been "gospodarski subjekti" – "economic subjects").

• The problems of NMT with abbreviations are exacerbated by inconsistency. "CA", which in a sampled document stood for "certification authority", was translated in six different ways on the two consecutive pages, including as "pristojni organ" – "competent authority" and "organ za konkurenco" – "competition authority". This was a technical annex, therefore the English abbreviation should have been kept, otherwise it could have been translated as "overitelj potrdil".

• NMT has many problems with structures which (if the context is disregarded) allow for different interpretations (e.g. "awareness on the benefits of earlier hepatitis and HIV testing" translated as "ozaveščenost o koristih prejšnjega hepatitisa in testiranju HIV", which back-translates as "awareness about benefits of earlier hepatitis and (awareness) about testing for HIV").

New error categories

The ex post analysis also produced examples of other types of errors in the eTranslation NMT output that had not been reported before:

• NMT seems to have problems with or around punctuation. Full stops went missing after numbers in numbered paragraphs or points (e.g. "70." translated as "70"). The wrong type of quotation marks was used and spaces around them were added (e.g. „ xyz „ instead of „xyz"). There were redundant spaces around formatted text.

Although easy to correct, these errors are as time consuming to correct as semantic mistakes. Some had originated from the pre-processing of the text before being sent to the engine (formatting converted to tags, subsequently replaced by spaces). Pre-processing has improved and such errors now occur less frequently.

NMT and different document types

The analysis also provided an insight into the usefulness of eTranslation NMT for different document types:

• In legislative documents, in the acts the biggest problem with NMT seems to be terminological errors and inconsistency, but as a tool it is generally useful. In terminology-heavy Annexes, the terminological errors and inconsistency might make NMT useless, especially if there are tables with fragmented text and many abbreviations.

• In non-legislative documents, the NMT output is in general at least somewhat useful. Terminology is still problematic, but as these documents are less terminology-heavy as a rule, machine translation produces fewer errors. However, domain-specific vocabulary still causes errors in the NMT output.

• NMT seems to be least useful for documents for the general public. Although terminology is mostly unproblematic, due to the less standardised vocabulary, complex structures and metaphors, NMT suggestions are mostly useless. In segments with problematic elements in the source, NMT output was poor not only due to handling them badly, but also did worse in the aspects where it is usually superior to SMT (e.g. incorrect inflections or awkward word order).

Legislative documents are still the bulk of the SL LD source texts. Even in the case of low match rate documents, there is more appropriate material in NMT engines' training. Consequently, NMT might produce better results for these than for other types of documents. The nature of these texts makes any errors critical, however. The document sample for the ex post analysis was small (three or four (partial) documents per category) and possibly not representative of the categories. Other documents might demonstrate NMT as more (or less) useful.

Application of the quality assessment tool

DGT uses an internally developed quality assessment calculator to evaluate outsourced translations. Errors are categorised according to the above-mentioned error grid and entered in the calculator, which assigns to them language-specific weights based on the type of document and the length of the sample. The final grade is displayed after deducting points from the initial 100.[11] The tool produced devastating marks for all sample documents of eTranslation NMT output. All received the lowest grade (<21/100 points), in each case reaching negative values. The table below includes the number of points awarded, the number of errors and remarks.

[11] As already stated, at the time of this writing, DGT is preparing for new outsourcing contracts, with a new error categorisation and a new evaluation procedure.

Implementing Regulation regarding 'Havarti' (PGI)							*final mark: –79*
sens[12] = 1	om = 0	term = 1	rd = 1	cl = 5	gr = 1	pt = 3	sp = 0
SENS = 5	OM = 1	TERM = 4	RD = 0	CL = 0	GR = 2	PT = 0	SP = 1
• wrong and inconsistent terminology, also inconsistencies in wording							
• problem with proper nouns (missing capitalisation, misspelling)							
Council Decision regarding HS (WCO)							*final mark: –34*
sens = 0	om = 0	term = 1	rd = 3	cl = 1	gr = 2	pt = 3	sp = 0
SENS = 0	OM = 0	TERM = 5	RD = 1	CL = 4	GR = 1	PT = 0	SP = 0
• incorrect standard phrases and terminology							
Annex to the Delegated Regulation supplementing ITS Directive							*final mark: –690*
sens = 0	om = 1	term = 4	rd = 0	cl = 6	gr = 5	pt = 2	sp = 0
SENS = 37	OM = 4	TERM = 9	RD = 0	CL = 7	GR = 5	PT = 0	SP = 0
• incorrect or nonsense translations due to the table format with text fragments							
• the many repeated abbreviations translated incorrectly and inconsistently							
• wrong and inconsistent terminology							
CMU report							*final mark: –89*
sens = 5	om = 0	term = 1	rd = 0	cl = 6	gr = 2	pt = 2	sp = 0
SENS = 3	OM = 2	TERM = 4	RD = 0	CL = 1	GR = 1	PT = 0	SP = 0
• several mistranslations of the domain specific financial vocabulary							
• some wrong and inconsistent terminology							
Climate action progress report							*final mark: –215*
sens = 6	om = 2	term = 1	rd = 1	cl = 18	gr = 4	pt = 4	sp = 0
SENS = 10	OM = 0	TERM = 0	RD = 0	CL = 3	GR = 2	PT = 1	SP = 0
• many mistranslations of the domain specific vocabulary							
• quite some reader unfriendly translations that needed rewording							
Replies of the Commission to an ECA Special Report							*final mark: –50*
sens = 1	om = 2	term = 4	rd = 0	cl = 12	gr = 6	pt = 2	sp = 0
SENS = 3	OM = 0	TERM = 1	RD = 0	CL = 2	GR = 0	PT = 0	SP = 0
• many minor mistranslations of the domain specific vocabulary							
Citizens' Dialogues and Consultations Progress report							*final mark: –335*
sens = 13	om = 1	term = 1	rd = 0	cl = 44	gr = 6	pt = 5	sp = 1
SENS = 10	OM = 2	TERM = 0	RD = 0	CL = 4	GR = 2	PT = 0	SP = 0
• many mistranslations and difficult to understand translations due to metaphorical language and vocabulary otherwise not frequently used in the translated documents							
Q&A on the EU list of non-cooperative tax jurisdictions							*final mark: –145*
sens = 6	om = 0	term = 3	rd = 0	cl = 12	gr = 3	pt = 1	sp = 0
SENS = 9	OM = 2	TERM = 0	RD = 0	CL = 3	GR = 0	PT = 0	SP = 0
• many mistranslations and difficult to understand translations due to rare vocabulary							
Health-EU Newsletter 223							*final mark: –140*
sens = 2	om = 0	term = 3	rd = 0	cl = 11	gr = 3	pt = 2	sp = 0
SENS = 9	OM = 0	TERM = 0	RD = 0	CL = 7	GR = 0	PT = 0	SP = 0
• mistranslations due to wrong deciphering of complex structures							
• non-translation of titles (considered proper nouns?)							
• difficult to understand translations of text fragments							
Memory game - Match the flags!							*final mark: –620*
sens = 24	om = 1	term = 0	rd = 0	cl = 23	gr = 9	pt = 4	sp = 1
SENS = 34	OM = 1	TERM = 0	RD = 0	CL = 8	GR = 1	PT = 0	SP = 0
• an extraordinary number of mistranslations and neural neologisms due to rare vocabulary and unusual structures							
• problems with proper nouns							

Table 3: *Results per document*

[12] The abbreviations in this table reflect the ones in the error grid and the quality assessment calculator: sens/SENS – mistranslation, om/OM – omission, term/TERM – terminology, rd/RD – reference documents, cl/CL – clarity, gr/GR – grammar, pt/PT – punctuation, sp/SP – spelling. Lowercase is used for minor errors and uppercase for major errors.

These results cannot, however, be taken as the definitive quality marks for the eTranslation NMT, as the methodology for the ex post review did not focus on the revision of the NMT, but on the analysis of the differences between the NMT and the final translation. Therefore, there is a possibility of false negatives (mistakes occurring in the NMT output and in the final translation). Moreover, in some documents with many errors (the document with the fewest comments contained 25 marked errors on the two pages and the document with the most contained a whopping 100 comments), some errors might not have been counted. Therefore, it is highly likely that a revision of NMT would have given even lower marks.

However, it should be born in mind that NMT was assessed against the criteria for human translation and that a specialised set of criteria might have given a different result. Consequently, a low mark earned by the output of NMT does not mean that NMT as a tool is not useful, but that the NMT output cannot be used as a final product. Maučec and Donaj (2020) also indicate a difference between evaluating the quality of MT output as a final product and evaluating its usability to human translators. Numerical results were assigned relatively low importance in the ex post analysis due to the methodological issues with using the quality assessment calculator on NMT and the small size of the sample that cannot produce representative results. The main focus was on consolidating and expanding the error categorisation started with the reported errors, as such indications can direct the translators' attention to the problem areas, and improve and speed up working with NMT.

6 Conclusion and follow-up

eTranslation NMT is widely used in the SL LD and is highly appreciated by the translators. They have assessed it to be a better tool than the SMT previously used. This corresponds to the finding of Burchardt et al. (2017) that turning from a phrase-based to a neural engine produced a striking improvement. Nevertheless, it has been clearly demonstrated that NMT is just a tool and not the final product (against the high standards required for Commission translations).

The reported errors revealed a variety of problems in the NMT output (polysemic misinterpretations, complete semantic blunders, terminological mistakes and inconsistencies, neural neologisms, omissions and additions; problems with proper nouns and abbreviations, complex structures and text fragments, text containing spelling errors and with (and/or next to) punctuation).

The ex post analysis confirmed that NMT output cannot be used as is. None of the sampled documents when simply pre-translated with the NMT output is a fit-for-purpose translation (fit for publication). A contributing factor for the poor result might also have been the fact that the sampled documents had a low match rate.

How much if any time is saved by machine translation remains unknown.[13] Furthermore, we gathered no definitive data to support the claim that the use of NMT saves time in comparison with the use of SMT.[14] Even though translators prefer working with NMT to working with SMT, we cannot claim that NMT use is necessarily easier and that it accelerates the translation process when compared to SMT. Maučec and Donaj (2020) identify three levels of post-editing effort: temporal, cognitive and technical. The use of NMT involves different types of issues that need to be dealt with and might be considered more mentally taxing than those inherent in the use of SMT (problems at higher levels of grammar and beyond grammar).

Clearly a skilled human is needed to guarantee a high quality of translation. This means NMT as a tool needs to be understood better in order to be used better, which requires educating and training its users, which is in line with Maučec and Donaj (2020), who emphasize the need of research on and teaching of skills specific to post-editing. Therefore, we extended the list of categorised reported errors with additional examples and added to the list new categories of recurring problems discovered during the ex post analysis. The list now paints as comprehensive a picture as possible of all potential English–Slovene NMT pitfalls. A department-level training was held to familiarise translators with the list and with the

[13] At least for the users of the EN–SL eTranslation NMT output. Macken et al. (2020) have worked with the French and Finnish LDs of DGT to assess how much time translators gain (or lose) in real-world conditions when they use machine translation, and observed the average speed gain of 14 % for English–Finnish NMT (and 12 % for English–French phrase-based SMT).

[14] In their analyses, Bentivogli et al. (2016), and Toral and Sánchez-Cartagena (2017), do arrive to the conclusion that NMT decreases post-editing effort compared to SMT. Klubička et al. (2017) also arrive to the same conclusion, and for a Slavic language close to Slovene.

results of the ex post analysis. The training raised awareness of expected error types among translators and informed our reflection on how to best use NMT. The request for translators to report examples remains open – the reporting is not systematic and the analysed sample was small, so it is possible that some errors elude categorisation due to a low number of occurrences and that some error types have not been detected yet.

Acknowledgement

The authors would like to thank the management of DGT for the opportunity to present our experience. We give our thanks to Jan Bednarich, the Head of the Slovene Language Department, for his continued support and to our colleagues Mojca Šauperl, Maksimiljan Gulič and Sarah Butcher for their indispensable help in improving the text, as well as to Markus Foti, the eTranslation Project Manager, for his insights.

References

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

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.

Drugan, Joanna, Ingemar Strandvik, and Erkka Vuorinen. 2018. Translation quality, quality management and agency: Principles and practice in the European Union institutions. In Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, editors, *Translation Quality Assessment: From Principles to Practice*, Springer, Cham, pages 39–68.

Klubička, Filip, Antonio Toral, and Víctor M. Sánchez-Cartagena. 2017. Fine-grained human evaluation of neural versus phrase-based machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108.1: 121–132.

Koehn, Philip, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 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 Interactive and Demonstration Sessions*, pages 177–180, Prague.

Lacruz, Isabel, Michael Denkowski, and Alon Lavie. 2014. Cognitive demand and cognitive effort in post-editing. In *Proceedings of the Third Workshop on Post-Editing Technology and Practice*. Association for Machine Translation in the Americas, pages 73–84, Vancouver.

Leal Fontes, Hilário. 2013. Evaluating machine translation: Preliminary findings from the first DGT-wide translators' survey. *Languages and Translation*, 6: 10–11.

Lesznyák, Ágnes. 2019. Hungarian translators' perceptions of neural machine translation in the European Commission. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 16–22, Dublin.

Macken, Lieve, Laura Van Brussel, and Joke Daems. 2019. NMT's wonderland where people turn into rabbits. A study on the comprehensibility of newly invented words in NMT output. *Computational Linguistics in the Netherlands Journal*, 9: 67–80.

Macken, Lieve, Daniel Prou, and Arda Tezcan. 2020. Quantifying the effect of machine translation in a high-quality human translation production process. *Informatics*, 7(2):12.

Maučec, Mirjam Sepesy, and Gregor Donaj. 2019. Machine translation and the evaluation of its quality. [Online First], IntechOpen, DOI: 10.5772/intechopen.89063. Available from: https://www.intechopen.com/online-first/machine-translation-and-the-evaluation-of-its-quality.

Strandvik, Ingemar. 2017. Evaluation of outsourced translations. State of play in the European Commission's Directorate-General for Translation (DGT). In Tomáš Svoboda, Łucja Biel and Krzysztof Łoboda, editors, *Quality aspects in institutional translation*, 8, pages 123–137, Berlin.

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

Van Brussel, Laura, Arda Tezcan, and Lieve Macken. 2018. A fine-grained error analysis of NMT, PBMT and RBMT output for English-to-Dutch. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3799–3804, Miyazaki.