Beyond Philology No. 17/4, 2020
ISSN 1732-1220, eISSN 2451-1498

https://doi.org/10.26881/bp.2020.4.02

Beyond MT metrics in specialised translation: Automated and manual evaluation of machine translation output for freelance translators and small LSPs in the context of EU documents

KRZYSZTOF ŁOBODA

Received 22.09.2020,
received in revised form 27.12.2020,
accepted 28.12.2020.

Abstract

This paper discusses simplified methods of translation evaluation in two seemingly disparate areas: machine translation (MT) technology and translation for EU institutions. It provides a brief overview of methods for evaluating MT output and proposes simplified solutions for small LSPs and freelancers dealing with specialised translation of this kind. After discussing the context of the study and the process of machine translation, an analysis of fragments of the selected specialist text (an EU regulation) is carried out. The official English and Polish versions of this document provide the basis for a comparative evaluation of raw machine translation output obtained with selected commercially available (paid) neural machine translation engines (NMT). Quantitative analysis, including the Damerau-Levenshstein edit distance parameters and the number of erroneous segments in the text, combined with a manual qualitative analysis of errors and terminology

can be a serviceable method for small LSPs and freelance translators to evaluate the usefulness of neural machine translation engines.

Keywords

machine translation, neural MT, institutional translation, MT evaluation, specialised translation

Miary jakości tłumaczenia maszynowego a przekład specjalistyczny. Metody automatycznej i manualnej oceny tłumaczenia maszynowego możliwe do zastosowania przez niezależnych tłumaczy i małe biura tłumaczeń w kontekście przekładu dokumentów UE

Abstrakt

Niniejszy artykuł przedstawia przyjęte i proponuje uproszczone metody oceny silników tłumaczenia maszynowego z myślą o małych biurach tłumaczeń i niezależnych tłumaczach zajmujących się przekładem specjalistycznym. Po omówieniu kontekstu badania oraz procesu tłumaczenia maszynowego przeprowadzona zostaje analiza fragmentów jednego tekstu specjalistycznego, którym jest wybrany akt prawny UE. Oficjalne wersje angielska i polska zestawione zostały z surowym tłumaczeniem maszynowym uzyskanym za pomocą 2 komercyjnych silników neuronowego tłumaczenia maszynowego (NMT): Microsoft Translator oraz Amazon Translate. Analiza ilościowa (m.in. parametrów odległości edycyjnej Damerau-Levenshteina i liczby błędnych segmentów w tekście) w połączeniu z manualną analizą jakościową błędów w tłumaczeniach może być przydatną metodą oceny przydatności silników neuronowego tłumaczenia maszynowego dla niezależnych tłumaczy.

Słowa kluczowe

tłumaczenie maszynowe, tłumaczenie neuronowe, przekład instytucjonalny, ocena tłumaczenia maszynowego, przekład specjalistyczny

1. The translation industry and machine transprocessing of texts

As the use of computer-aided translation tools and machine translation (MT) technology in the translation industry is gradually becoming the norm rather than an exception, we can observe an industry-wide tendency to seek synergy in incorporating these tools in the translation process (Moorkens and O'Brien 2017). Machine translation engines enable an automated[1] processing of the language code whereby a document in the source language is the basis for an almost instantaneous generation of another text in the target language. However, what is time and cost saving for translation agencies can be a source of trouble for freelance translators since raw MT output is often of mixed quality and the results of the MT process might seem unpredictable. The recently introduced translation industry standard ISO 18587:2017 "Translation services – Post-editing of machine translation output – Requirements", which has been in use since February 2018, defines the workflow of full post-editing. It is implemented mostly by larger language service providers (LSPs) who strive to achieve "human parity", i.e. to make a MT post-editing indistinguishable from a human translation. In order to compete with the Goliaths in the industry, many smaller LSPs and experienced freelancers who work for their direct clients are also increasingly turning to machine translation as an efficiency-boosting technology.

Over the last 70 years various machine translation solutions have been proposed (see e.g. Bogucki 2009): example-based translation methods (EBMT) coupled with fuzzy logic principles have been developed in parallel with rule-based translation (RBT) systems. In the early 2000s these methods were replaced with statistical machine translation (SMT) and, most recently, with neural machine translation (NMT).

[1] Hence, with regard to machine translation, we will also use the term *transprocessing* here in contrast to (human) translation.

Despite all these advances in the integration of various areas of research in artificial intelligence, the natural language content in translation applications is still processed without any sensory perception (i.e. without recognizing the image, voice, taste, smell or even the place where the message is transmitted) and without considering the components of the communicative act, such as a pragmatic context, cultural context, the encyclopaedic knowledge of the translator, the target audience (Chesterman 1997), the assumed knowledge of the intended recipient (Tabakowska 1999: 54), etc. Within the last decade, several models representing meaning as high-dimensional numerical vectors, or vector-space models of semantic representation, have been developed (see e.g. Mikolov et al. 2013) to better capture the use of ambiguous expressions in a specific conceptual domain, yet automatic processing of meaning and text is still quite far from the human ability to differentiate between contexts. Basically, natural language processing algorithms could easily transcode any message into other sentences in the same language (intralingual transfer) or transcode the content into images or sounds (intersemiotic transfer). It can be assumed that at the turn of the second and third decades of the 21st century, machine translation of natural language is still predominantly limited to transcoding the text without the use of cognitive functions and without understanding and interpretation of the message taking into account its situational or cultural contexts (cf. Quah 2006: 18).

However, with the vast amount of training data widely available, MT is slowly becoming a mature technology. In a paper describing an experiment carried out in 2018, Popel et al. (2020) claim that machine-human parity was reached when translating isolated sentences from newspapers in selected language directions.[2] In a recent study conducted in the English-Polish language pair (Kur 2020), the feasibility of implementation of three

[2] Unfortunately, since the public service Lindat where Popel's CUBBITT system is implemented does not offer EN-PL automated translation, these claims cannot be easily validated.

generic MT systems was considered (for translating newspaper articles). As for the specialised translation in the EU context, which is our concern in this paper, the MT service known as e-Translation is used by in-house and external translators of the European institutions. Building an internal MT system might not be a problem for larger organisations and translation companies, yet freelance translators and small LSPs would probably need some help in selecting and assessing such solutions for the purposes of their translation jobs.

In this paper, we will briefly review the methods used for evaluating MT output and try to use some of them for a text from a specialised domain, i.e. an EU legal document. In this way, we should be able to propose simple MT evaluation methods (e.g. potential error indicators for subsequent qualitative assessment) which could potentially be of use to smaller LSPs and freelance translators of specialist texts. The aim is to help them make informed choices as to the evaluation of MT technology, and decide whether to put an MT system in place for their projects.

2. Selection of a text from a specialised domain and commercial MT systems for evaluation

For the purposes of our study, first we needed to choose a pair of reference texts from a specialised domain in the source and target languages, which in our case was English and Polish, respectively. To that end, legal instruments which are available and binding in multiple language versions seemed good candidates. With this in mind, we took an EU Regulation, as it is available in all official language versions and directly applicable in all Member States. Consequently, Regulation (EU) No. 1308/2013 of the European Parliament and of the Council (see Annex; European Union 2013; Unia Europejska 2013) was chosen as the reference text for further examination.

A sample of 26 segments was taken from two sections of the English version of the document. Extract 1 (Segments 1-12)

includes the title and the initial part of the preamble, whereas Extract 2 (Segments 13-26) contains the enacting terms with Articles 59-61 of the Regulation. The text, prepared in this way, was compared with the official Polish version published in the Official Journal of the European Union, downloaded from Eur-Lex (provided in the Annex), which is deemed our reference or 'gold standard' translation.

This English text was then used for the basis in machine processing of text in the EN->PL combination using three commercial MT systems: Microsoft Translator (MST) and Google Translate engines accessed via a single CAT tool plugin and the Amazon Translate (AMZT) engine used in the browser via AWS service. The output from the MT systems was collected in mid-March 2020. The selection of these three MT engines seems justified as they are used by some commercial MT integration services which are particularly targeted at small LSPs and freelancers.[3] As further indicated in the Discussion section, two MT systems were found to be of similar quality and one seemed significantly worse so, for the sake of economy, only the two extremes, i.e. the output of Microsoft and Amazon systems, were chosen to illustrate possible problems in evaluating MT. The worst and best raw machine translation and the official versions of the Regulation in English and Polish are shown in the Annex.

3. From MT metrics and automated MT quality assessment to the dimensions of post-editing effort and full evaluation

Let us now focus on quantitative and qualitative methods of MT assessment. A succinct overview should facilitate further selection of potentially fast and simple methods of MT evaluation. Basically, MT output can be evaluated in an automated way or manually with the help of previously trained humans.

[3] I am grateful to an anonymous Reviewer for mentioning Memsource as one of such services where these three engines are integrated.

By far the most comprehensive, potentially the most objective and also the most demanding method is the full evaluation of MT output by many raters. An indicator of MT quality can, for example, be the postediting effort (PE) as defined by Krings (2001), who distinguishes temporal, cognitive and technical dimensions of PE. As for the temporal effort, measuring and comparing the time needed to translate a text from scratch and postedit an MT version can be a viable option to consider for midsize LSPs, yet even this might still prove overly time- and resource-consuming for a single freelance translator or tiny translation companies. The cognitive dimension of postediting effort is possibly the most difficult to measure as (aside from the think-aloud protocol (TAP) method) it usually requires costly high-resolution eye tracking equipment. Eye tracking technology enables evaluators to identify *fixation points*, or the words and phrases in the text where proofreaders' eyes rested for longer periods of time, which is an indicator of greater cognitive load. Finally, the technical effort is measured by the number of editing operations (such as insertions, deletions, substitutions) and usually obtained by keylogging and screen recording software (or the less handy TAP method).

The level of effort expanded in the proofreading is usually analysed using some error classification. The division of errors into possible categories is quite subjective and there are many typologies used both in research and the translation industry (see e.g. Popović et al. 2014, Daems et al. 2017, Toral and Sánchez-Cartagena 2017). For our purposes we chose the scale and typology used by the Directorate-General for Translation of the European Commission as described by Strandvik (2017). At the same time, we must bear in mind that full evaluation by many raters is infeasible for small LSPs and freelancers and that due to these constraints, the qualitative analysis and error classification must be quite limited and should only complement the automated quantitative analysis.

4. BLEU metric: imperfect but widely used

The translation industry uses many automatic measures, or metrics of machine translation quality, including BLEU, METEOR, F-Measure, chrF, TER, HTER and NIST (see e.g. Sno-ver et al. 2006 or Popović 2015 for correlations of 'best performing metrics'). In the EU context, one of the recently proposed metrics is CharCut (Lardilleux and Lepage 2017), but it has not gained much popularity so far.

The BLEU (Bilingual Evaluation Understudy) metric developed in IBM laboratories (Papineni et al. 2002) is most used nowadays. BLEU is based on matching *n*-grams present in automatic translation to *n*-grams in the reference translation when considering precision and brevity penalty. Though it is not perfect and is often criticised for not being adequately correlated with human judgements, it remains the most popular in the translation industry as the only metric that allows for drawing comparisons with other work over the last two decades (examples of recent research where BLEU is used as the main metric include Läubli et al. 2020, Popel et al. 2020[4]).

For our text, the calculated BLEU values for NMT engines reach the values of 61.63, 72.85 and 73.71 for Microsoft, Google and Amazon MT systems, respectively (explained further in the Discussion section). These scores might be useful indicators suggesting that in the chosen textual domain, Amazon MT and Google MT engines are likely to produce higher quality results than Microsoft Translator. However, automatic metrics should not be regarded as the ultimate evaluation of machine translation output—they are in fact the cheapest and fastest rough estimation of MT quality, so the initial results would need to be corroborated by a subsequent qualitative analysis. The scores often happen to be biased or even erroneous (hence the multitude of various metrics). Furthermore, the aggregate BLEU sco-

[4] My thanks to anonymous Reviewer 1 for pointing out the fresh work by Popel et al. (2020) where BLEU and TER are used as the principal metrics.

re for the engine does not indicate what types of errors occur and where they are to be found in the text.

Another weakness of automated metrics is that they are complicated and not widely accessible. The average freelance translator or smaller LSPs would not be very likely to have at their disposal the tools to calculate BLEU, TER, METEOR or CharCut scores. As a way out, we might try to obtain some indicative results in a spreadsheet. With simple calculations which in a way underlie automatic translation quality metrics, we will try to predict possible problematic segments in MT output using either of the two options shown below and then check if the indicated sentences do indeed contain any errors.

4.1. Quality prediction based on characters

One possible option is to calculate and compare the number of characters in each segment in order to indicate the segments where some content may have been omitted or added by an MT system. The graph in Figure 1 shows the number of characters in segments in the official Polish document and the output of the machine translation engines in question. As we can see, the Microsoft Translator engine seems to differ from both the reference translation and the Amazon Translate engine, offering shorter translations – segments 5 and 15 are worth checking for the quality of the translation and possible omissions.

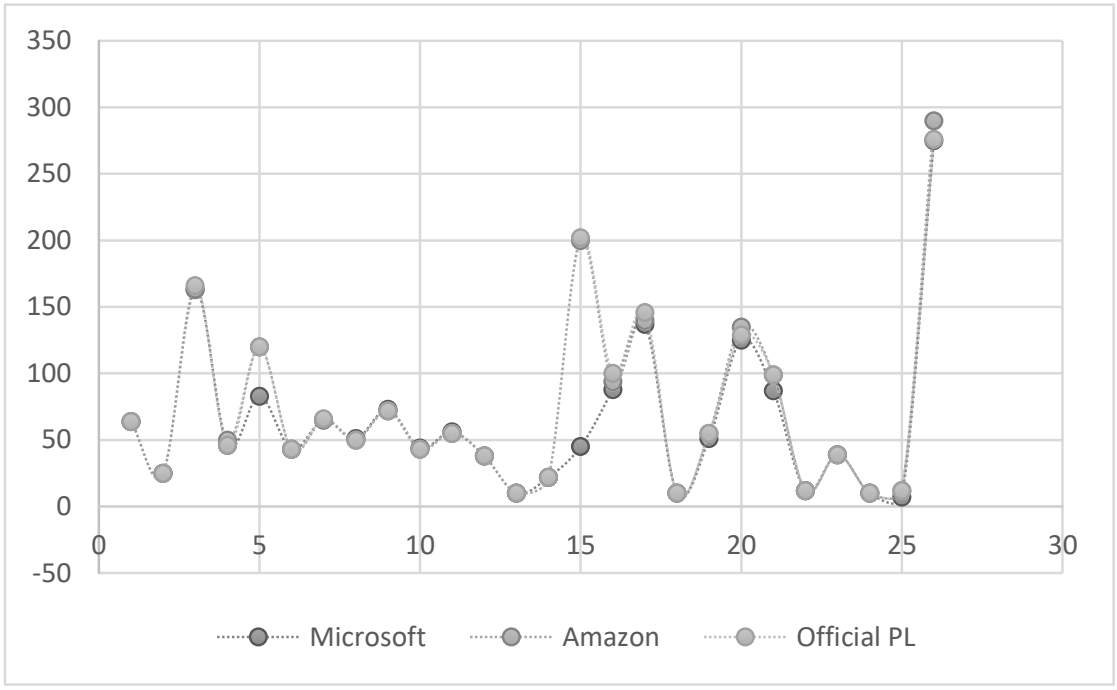

Figure 1
Characters per segment in the official Polish version
and raw MT output in Microsoft and Amazon

4.2. Quality prediction based on words

Metrics may also be based on words[5] rather than characters. To keep the analysis as simple as possible, we could calculate the number of words in each segment and possibly introduce typical statistical calculations (variance, standard deviation). In our case, we stuck to a rough quantitative analysis that allowed us to select segments for a qualitative analysis at a later stage of the assessment. A simple and effective method which consists in calculating the percentage differences in the number of words in segments from the reference text (the official Polish version) sufficed here (Figure 2 and Table 2). As we can see, in this way we could obtain a more detailed image of the differences

[5] An example of such a metric is WER (Word Error Rate), used predominantly in automated speech recognition.

between the segments of the individual versions of the text under analysis.

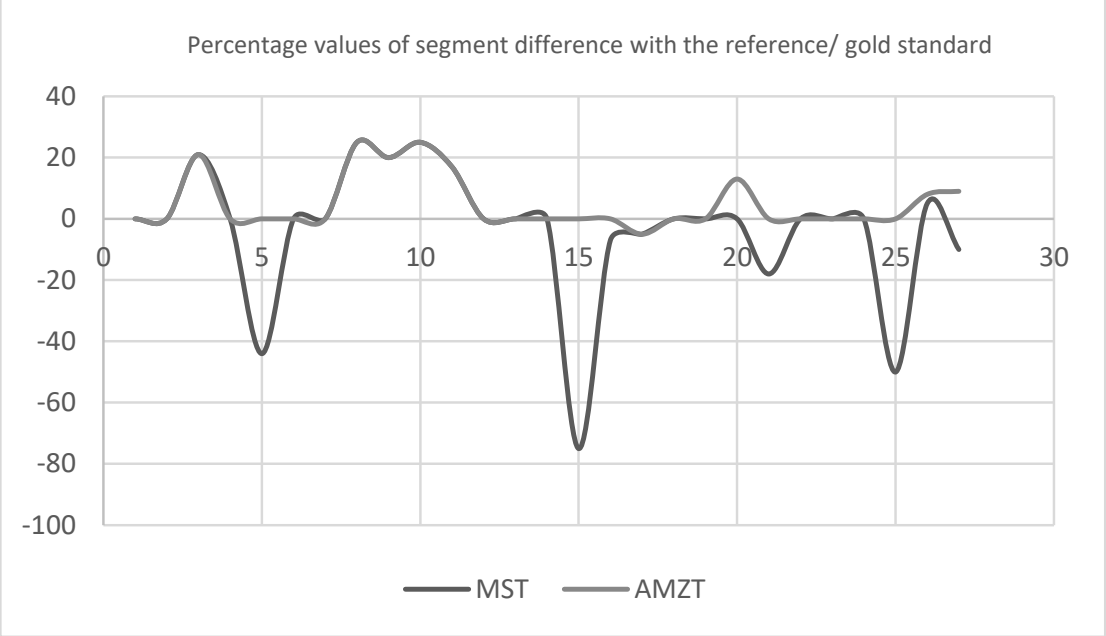

Figure 2
Percentage values of segment difference with the reference text

Table 1
Mean, median and standard deviation
of segmental differences against the reference text

	MEAN	MEDIAN	STD
MST	-3	0	22.00
AMZT	5	0	8.87

As for the segment wordcount, significant percentage divergences from the reference Polish version can be observed for the Microsoft Translator engine, whereas the commercial Amazon MT engine seems closer to the official version published in EU legislation database, EUR-Lex. If we have a look at other statistics (see Table 2), the Microsoft engine appears to use slightly fewer words (3 % less) while Amazon a slightly more (5 % more)

words when compared with the reference text. At the same time, the standard deviation for MST is significantly higher than that of AMZT. The median does not show any differences and seems to be of no prognostic value in our evaluation. The detailed values of percentage differences for individual segments are shown in Table 3. Segments with the same MT output (zero difference) have been omitted.

Table 2
Percentage values of segment difference with the official translation and mean values

Segment # / MT engine	3	5	8	9	10	11	15	16	17	20	21	25	26
MST	21	-44	25	20	25	17	-75	-7	-5	0	-18	-50	5
AMZT	21	0	25	20	25	17	0	0	-5	13	0	0	8

Assuming a cut-off threshold of more than 25 %, a quantitative predictive analysis indicates the following significant differences for individual NMT engines:

(1) Microsoft Translator – possible omissions in segments 5, 15, 25;
(2) Amazon Translate – no segments with the threshold value exceeded (however, the threshold value was reached in two segments).

4.3. Quantitative testing by measuring the edit distance

The edit distance parameter is also commonly used to measure the quality of machine translation. In simple terms, the classic Levenshtein distance is the sum of the operations of removing, inserting and substituting characters in two compared strings of characters. A slightly altered variant of minimum edit distance developed by Levenshtein together with F. J. Damerau (1964) is used more often in the translation industry. This

measure involves inserting, deleting, substituting the character and additionally transposing (shifting) two adjacent characters. To better understand the principle of calculating the edit distance, let us consider two words: GDYNIA and GDANSK. The matrices showing the number of operations necessary to turn one word into another are shown in Figure 3. As we can see, the value of minimum edit distance may equal 3 or 6 (this means 100 % difference!), depending on the variant applied. It is worth mentioning that CAT tools most often use variant b (the Damerau-Levenshtein distance), which always gives smaller values. This distinction may be of importance for freelancers and small LSPs as regards their remuneration for their work in translation and postediting projects.

		G	D	Y	N	I	A
	0	1	2	3	4	5	6
G	1	0	1	2	3	4	5
D	2	1	0	1	2	3	4
A	3	2	1	2	3	4	3
N	4	3	2	3	2	3	4
S	5	4	3	4	3	4	5
K	6	5	4	5	4	5	6

G D A N S K
G D Y N I A

a) Classic Levenshtein distance = 6 (substitution weight 1)

		G	D	Y	N	I	A
	0	1	2	3	4	5	6
G	1	0	1	2	3	4	5
D	2	1	0	1	2	3	4
A	3	2	1	1	2	3	3
N	4	3	2	2	1	2	3
S	5	4	3	3	2	2	3
K	6	5	4	4	3	3	3

G D A N S K
G D Y N I A

b) Damerau-Levenshtein distance = 3 (substitution weight 2)

Figure 3
Calculating the edit distance between two words: *GDANSK* and *GDYNIA*

Multi-part strings, whole sentences and even whole texts can also be analysed in this way. The minimum edit distance (MED) calculated against the Polish version (according to the Damerau-Levenshstein model) for whole texts generated by individual machine translation engines is as follows:

Table 3
Edit distance calculated for both MT versions

	Microsoft	Amazon	Mean value
MED	376	177	277

Owing to specific algorithms, it might be possible to make an initial estimate of the quality of the machine translation before actually embarking on any qualitative analysis. Theoretically, a smaller edit distance means a translation closer to the reference translation, therefore for our sample text we should expect higher quality from Amazon Translation engine. This can be examined in a qualitative examination of MT output.

5. Manual evaluation of the quality of translation according to the European Commission's DGT criteria

In this section we will attempt to compare the official English and Polish versions with the raw output of selected neural machine translation (NMT) engines: the Microsoft Translator generic engine, and the commercial Amazon Translate engine. Each version of the translation will be evaluated using a hierarchy of resources (see Łoboda 2012) and the EC DGT evaluation system as described by Strandvik (2017). In one of its longstanding evaluation models, the Directorate-General for Translation of the European Commission distinguishes two dimensions of errors in categories such as wrong rendering of the sense resulting in mistranslation or unjustified addition of content (SENS), unjustified omission or non-translation (OM),

terminological error (TERM), inconsistency with reference documents (RD), grammatical error (GR), spelling error (SP), punctuation error (PT), and unclear conveyance of meaning (CL).

5.1. Microsoft Translator NMT engine

Microsoft Translator is an engine used for the automated transprocessing of multilingual content in the documentation of Microsoft products, therefore it should be particularly suitable in rendering IT-related texts into another language. This solution is also available free of charge as a generic machine translation engine (implemented in Bing Translator) and commercially (on the Microsoft Azure platform as one of Azure Cognitive Services solutions). The sample included in the attachment was generated via an API plugin installed in one of the CAT tools.[6]

Predictive analysis using the editing distance indicated discrepancies with the official Polish version in almost half of the segments. Segments 5, 15 and 25 were indicated as particularly problematic, and indeed they turned out to be grossly incorrect. The machine-generated text contains very serious omissions and terminological errors, which make the text quality unacceptable in terms of the EC DGT criteria.

(1) OM error category – several major omissions of large sections of text after each first full stop of the MT output (segments S5, S15, S21);
(2) TERM error category – terminological inconsistency (*proce-dura kontrolna* in S19 and *procedura sprawdzająca* in S21);
(3) GR error category – inappropriate form in S17 when continuing in S15 and S16; ungrammatical form *w celu zapewnienia, że* in S15;
(4) SENS/CL error category – *obszar chmielu* instead of *obszar uprawy chmielu* in S19; *czas trwania* [duration] translated as

[6] In our internal tests, the output of the commercial NMT by Microsoft accessed via a CAT-tool plugin proved to be identical with the version obtained with the publicly available free Bing Translator engine.

długość [length] in S25; ambiguous translation of S26 (*w celu oceny… systemu i* [*w celu oceny*] *złożenia wniosków*);

(5) SP/PT error categories – incorrect capitalisation of the line of recital in the preamble (S7); adding a slash before numbers in segments S8-S11.

The quality assessment shows that 12 out of 26 segments 12 are identical with the official Polish version, 6 segments contain errors and minor issues, and 8 contain errors considered grave. This is mainly due to an obvious flaw in the implementation of the engine which results in removing the all of the text that follows any full stop in the raw MT output. Thus, a large number of words were omitted, which resulted in a considerable edit distance in relation to the reference text and the output of the other engine under analysis (ED 376 with mean value 277).

5.2. Amazon Translate NMT engine

Amazon Translate is a recently introduced generic machine translation engine. It is offered commercially and has been used for Amazon, one of the world's largest e-commerce platforms. Amazon operates in many countries, and individual regional websites are available in several languages thanks to machine translation. For example, in addition to the default German language version, Amazon.de regional website provides machine translation in 5 other languages (English, Dutch, Polish, Czech and Turkish). The translation service is also commercially available on a *pay-per-use* basis within AWS (Amazon Web Services) Cloud Platform in 55 languages and seems to be aimed specifically at the e-commerce market. Amazon Translate, unfortunately, is not available as a free open service. The service documentation does not indicate the sources used to build a generic language model and to train the neural network.[7]

[7] https://docs.aws.amazon.com/translate/latest/dg/how-it-works.html.

In the case of this engine, the quantitative predictive analysis did not reveal a single segment that would deviate significantly from the Polish reference text. This is confirmed by a low edit distance value. However, a detailed qualitative analysis reveals the following errors:

(1) GR error category – incorrect grammatical case in segment S4; incorrect grammatical cases in the listed section in segments S16-S17; ungrammatical form *w celu zapewnienia, że* in S15;
(2) CL/SENS error category – as in the case of the other MT engine analysed, *obszar chmielu* instead of *obszar uprawy chmielu* in S19; ambiguous translation of S26 (*w celu oceny… systemu i przedstawienia propozycji*);
(3) SP/PT error category – minor defects due to the change of bracketed references to square brackets (S8-S11); incorrect capitalisation of the line of recital in the pre-amble (S7).

The text generated by the Amazon Translate engine does not contain any significant terminological errors. The raw MT output turns out to be surprisingly similar to the official Polish version, which may suggest that Amazon Translate is a high quality tool and/or the fact that this text has been used to train the NMT engine. The convergence of the official version and NMT output is confirmed by a very low edit distance of 177 with the mean value of 277. Nevertheless, a few grammatical errors were found in the text, which affects the overall quality of machine translation. All in all, our qualitative analysis of errors corroborates our findings from the quantitative predictive analysis. In the case of our reference text, the output of Amazon Translate NMT engine indeed provides a significantly higher quality than Microsoft Translator.

6. Discussion

There are a few issues to consider in this context. First, the values of BLEU which hardly ever reaches 30-40 in general contexts (such as news, see Popel et al. 2020) were found to be significantly higher for our document. Such a high level of correspondence between the reference and MT hypothesis might mean that: (i) our reference text was translated using an NMT engine or (ii) the reference text was used to train the MT system and/or (iii) that the specialised texts in question (EU law) are highly standardized in terms of the terminology and formulaic language so they are processed in a more uniform way by an NMT system. The first option can easily be rejected since EU Regulation No. 1308/2013 was published over 3 years prior to the launch of the NMT services by Google and Microsoft. The second option seems plausible, since the EU institutions (the European Commission and European Parliament) compiled large corpora of EU legislation which have been made available to the public over the last decade. Therefore, it seems advisable to compare the BLEU value for the text in question (our Text 1, or T1) with two other documents: one from the same domain and text type, and another from a related domain and a differing text type. To that end we selected two freshly published documents (texts 2 and 3, or T2 and T3): Commission Implementing Regulation (EU) 2021/28 (European Commission 2021) and a news article from the EU Research Portal CORDIS (Publications Office 2021). We ensured that T2 and T3 were newly published documents in order to minimise the risk of them having been used as the training material for commercial NMT engines. The calculated BLEU scores for T1, T2 and T3 are shown in Table 4.

Table 4
BLEU scores for T1, T2 and T3 without lowercasing the text (the higher, the better)

	Micro-soft	Google	Amazon	Mean MT	Do-main	Text type	Publica-tion year
BLEU T1	61.63	72.85	**73.71**	69.40	EU law	Legisla-tion	2013
BLEU T2	61.67	60.37	**62.53**	61.52	EU law	Legisla-tion	2021
BLEU T3	22.94	**27.87**	26.23	25.68	EU rese-arch	News article	2021

For a news article (T3), where the highest result was obtained by Google Translate, BLEU scores reach typical, significantly lower values than for a specialist document such as EU legislation (T1 and T2). The MT systems by Google and Amazon reach comparable quality, though we found it surprising that for T1 and T2 it was Amazon (a system built primarily for e-commerce) that scored slightly better. It is worth noticing that the texts were not lowercased, as this would deviate from human translation evaluation criteria in EU institutions (such as spelling and capitalization). Otherwise, the scores would be a few points higher. We should also note that for uniformly lowercased texts (a frequent practice in MT evaluation), a higher BLEU score would have been reached by Google rather than Amazon.

The European law and EU-related documents provide fascinating material for the evaluation of machine translation solutions. The amount of data made available for the training purposes by the European institutions over the last two decades are unprecedented, so the quality of generic MT systems can be relatively high for some text types. At the same time, we should bear in mind that EU legal texts (such as Regulations, Directives, Decisions) are highly standardised and written according

to accepted templates. The EU policy-related terminology is also quite uniform, as the crucial and most frequent terms are entered into the IATE database which in turn is a binding source for in-house and external translators of the EU institutions. Such a normalized text structure and terminology is the main reason why NMT systems can give relatively good results and high BLEU scores.

7. Limitations of this study and concluding remarks

We can see that a quantitative analysis can be a useful method for finding general differences between the evaluated MT output and the reference text in some highly conventionalized documents such as EU law. A quantitative analysis of MT makes it easy to detect the number of segments deviating from the adopted version, as well as to assess the scale of such discrepancies.

Such quantitative methods have both their advantages and limitations. First and foremost, they are fast and easy to use. They provide translation project managers with immediate results and statistical data without the need to adhere to more complex MT metrics. The predictive quantitative analysis has a significant prognostic value: some assumptions as to the MT quality can be made before the proper evaluation of MT quality by professional translators. As for the limitations, we should pay attention to the low efficiency in finding grammatical errors which are always considered grave. All the metrics we have discussed here in detail share the same principle which underlies the NMT technology: the algorithms treat texts as sequences of individual sentences/segments rather than coherent texts (see Läubli et al. 2020).

The method described here is restricted to specific, highly conventionalised types of texts from a specific domain such as EU law where the use of synonyms would be limited. In other contexts (e.g. newspaper articles such as T3), our methods would be less reliable but fit for our purpose (and as we can see from the table above, MT engines also fare worse). Quality

prediction based on the number of characters or words is a very simple solution but hardly meant to replace BLEU, METEOR or human-targeted metrics. We believe, though, that in certain contexts such a procedure could be useful for translators or small LSPs who do not have access to tools offering such metrics. While other solutions are usually less accessible or offered as paid solution,[8] with a limited number of language combinations and not always disclosed quality estimation algorithms, calculation of characters or words in the segments can be carried out for free in any spreadsheet application.

The quantitative and qualitative analysis, which is primarily of a technical and linguistic nature, could be further combined with measuring the temporal effort of the post-editing process. This is generally possible to accomplish with the most popular CAT tools (e.g. Qualitivity plugin in Trados Studio), PET (Aziz et al. 2012) or ROE (Farrell 2018). These more advanced solutions allow for filling in translation evaluation forms according to selected translation quality standards and for examining the time spent on post-editing. However, the methods analysed in this paper should be sufficient for freelance translators and small LSPs, enabling them to make informed choices as to whether to put specific MT engines in place in the context of their projects.

References

Aziz, Wilker, Sheila C. M. de Sousa, Lucia Specia (2012). "PET: A tool for post-editing and assessing machine translation". In: Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, Stelios Piperidis (eds.). *Proceedings of the 16th Annual*

[8] Examples include Intento or Memsource, which offers its paid MTQE solution. However, the English-Polish combination is not officially supported when this paper is written. The algorithms behind MTQE values (which are similar to fuzzy bands) are not revealed by the company. (I would like to thank one the anonymous Reviewers for drawing my attention to MTQE by Memsource).

Conference of the European Association for Machine Translation. Istanbul: European Language Resources Association (ELRA), 3982–3987.

Bogucki, Łukasz (2009). *Tłumaczenie wspomagane komputerowo.* Warszawa: PWN.

Bojar, Ondřej (2017). *English-to-Czech MT: Large Data and Beyond. Habilitation Thesis*. Prague: Institute of Formal and Applied Linguistics, Charles University.

Chesterman, Andrew (1997). *Memes of Translation. The Spread of Ideas in Translation Theory*. Amsterdam – Philadelphia: John Benjamins.

Damerau, Frederick J. (1964). "A technique for computer detection and correction of spelling errors". *Communications of the ACM* 7 (3): 171–176.

Daems, Joke, Sonia Vandepitte, Robert J. Hartsuiker, Lieve Macken (2017). "Translation methods and experience: A comparative analysis of human translation and post-editing with students and professional translators". *Meta* 62/2: 245–70.

European Union (2013). "Regulation (EU) No 1308/2013 of the European Parliament and of the Council of 17 December 2013 establishing a common organisation of the markets in agricultural products and repealing Council Regulations (EEC) No 922/72, (EEC) No 234/79, (EC) No 1037/2001 and (EC) No 1234/2007". *Official Journal of the European Union,* OJ L 347, 20.12.2013, 671–854.

European Union (2021). "Commission Implementing Regulation (EU) 2021/28 of 14 January 2021 amending Council Regulation (EC) No 1362/2000 as regards the Union tariff quota for bananas originating in Mexico". *Official Journal of the European Union.* OJ L 12, 15.1.2021, 1–2.

Farrell, Michael (2018). "Raw output evaluator, a freeware tool for manually assessing raw outputs from different machine translation engines". In: David Chambers, Joanna Drugan, João Esteves-Ferreira, Juliet Margaret Macan, Ruslan Mitkov, Olaf-Michael Stefanov (eds.). *Proceedings of the 40th Conference Translating and the Computer, London, UK, November 15-16, 2018*. London: International Society for Advancement in Language Technology Asling, 38–49.

Fischer, Lukas, Samuel Läubli (2020). "What's the difference between professional human and machine translation? A blind multi-language study on domain-specific MT". In: André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa

Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, Mikel L. Forcada (eds.). *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. Lisboa: European Association for Machine Translation, 215–224.

Krings, Hans P. (2001). *Repairing Texts: Empirical Investigations of Machine Translation Postediting Processes*. Kent: The Kent State University Press.

Kur, Maciej (2020). *Feasibility of DeepL, Google and Microsoft MT Systems Implementation into the Translation Process*. Gdańsk: Wydawnictwo Uniwersytetu Gdańskiego.

Lardilleux, Adrien, Yves Lepage (2018). "CHARCUT: Human-targeted character-based MT evaluation with loose differences". In: Sakriani Sakti, Masao Utiyama (eds.). *Proceedings of the 14th International Workshop on Spoken Language Translation, Tokyo, Japan, December 14th-15th, 2017*. Tokyo: IWSLT, 146–153.

Läubli, Samuel, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, Antonio Toral (2020). "A set of recommendations for assessing human–machine parity in language translation". *Journal of Artificial Intelligence Research* 67: 653–672.

Łoboda, Krzysztof (2012). "Praktyczne i dydaktyczne aspekty przekładu dokumentów instytucji Unii Europejskiej: charakterystyka tekstów, narzędzi i problemów terminologicznych". In: Maria Piotrowska, Joanna Dybiec-Gajer (eds.) *Przekład – teorie, terminy, terminologia. Język a komunikacja 30*. Kraków: Tertium, 161–169.

Martínez Mateo, Roberto (2014). "A deeper look into metrics for translation quality assessment (TQA): A case study". *Miscelánea: A Journal of English and American Studies* 49: 73–94.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean (2013). "Distributed representations of words and phrases and their compositionality". In: Christopher J. C. Burges, Léon Bottou, Max Welling (eds.). *Proceedings of the 26th International Conference on Neural Information Processing Systems – Vol. 2, December 2013*. New York: Curran Associates, 3111–3119.

Moorkens, Joss, Sharon O'Brien (2017). "Assessing user interface needs of post-editors of machine translation". In: Dorothy Kenny (ed.). *Human Issues in Translation Technology: The IATIS Yearbook.* Florence: Taylor and Francis, 110–130.

Papineni, Kishore, Salim Roukos, Todd Ward, Wei-Jing Zhu (2002). "BLEU: a method for automatic evaluation of machine translation".

In: Pierre Isabelle, Eugene Charniak, Dekang Lin (eds.). *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics.* Pennsylvania: Association for Computational Linguistics, 311–318.

Popel, Martin, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, Zdeněk Žabokrtský (2020). "Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals". *Nature Communications* 11, 4381. Available at <https://doi.org/10.1038/s41467-020-18073-9>. Accessed 14.03.2021.

Popović, Maja, Arle Lommel, Aljoscha Burchardt, Eleftherios Avramidis, Hans Uszkoreit (2014). "Relations between different types of post-editing operations, cognitive effort and temporal effort". In: Mauro Cettolo, Marcello Federico, Lucia Specia, Andy Way (eds.). *Proceedings of the 17th Annual conference of the European Association for Machine Translation.* Dubrovnik: European Association for Machine Translation, 191–198.

Popović, Maja (2015). "CHRF: character n-gram F-score for automatic MT evaluation". In: Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, Pavel Pecina (eds.). *Proceedings of the Tenth Workshop on Statistical Machine Translation.* Lisbon: Association for Computational Linguistics, 392–395.

Publications Office (2021). Long-lived worms hold the secret for healthy ageing in humans. *CORDIS EU Research Results.* Available at <https://cordis.europa.eu/article/id/428745-long-lived-worms-hold-the-secret-for-healthy-ageing-in-humans>. Accessed 14.03.2021.

Quah, Chiew Kin (2006). *Translation and Technology*. Basingstoke / New York: Palgrave Macmillan.

Rinsche, Adriane, Nadia Portera-Zanotti (2009). *The Size of the Language Industry in the EU. Studies on Translation and Multilingualism*. Brussels: European Commission.

Rossi, Caroline, Jean-Pierre Chevrot (2019). "Uses and perceptions of machine translation at the European Commission". *The Journal of Specialised Translation* 31: 178–200.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, John Makhoul (2006). "A study of translation edit rate with targeted human annotation". *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA).* Cambridge, USA: AMTA, 223–231.

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, Krzysztof Łoboda (eds.). *Quality Aspects in Institutional Translation*. Berlin: Language Science Press, 123–137.

Tabakowska, Elżbieta (1999). *O przekładzie na przykładzie*. Kraków: Znak.

Toral, Antonio, Sheila Castilho, Ke Hu, Andy Way (2018). "Attaining the unattainable? Reassessing claims of human parity in neural machine translation". In: Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, Karin Verspoor (eds.) *Proceedings of the Third Conference on Machine Translation: Research Papers*, *Vol. 1*. Brussels: Association for Computational Linguistics, 112–123.

Toral, Antonio, Víctor M. Sánchez-Cartagena (2017). "A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions". In: Mirella Lapata, Phil Blunsom, Alexander Koller (eds.). *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia: Association for Computational Linguistics, 1063–1073.

Unia Europejska (2013). "Rozporządzenie Parlamentu Europejskiego i Rady (UE) nr 1308/2013 z dnia 17 grudnia 2013 r. ustanawiające wspólną organizację rynków produktów rolnych oraz uchylające rozporządzenia Rady (EWG) nr 922/72, (EWG) nr 234/79, (WE) nr 1037/2001 i (WE) nr 1234/2007", *Dziennik Urzędowy Unii Europejskiej*, Dz.U. L 347, 20.12.2013. 671–854.

Krzysztof Łoboda
ORCID iD: 0000-0002-9575-5080
Jagiellonian University
Institute of Linguistics
and Translation Studies
al. Mickiewicza 9a/408
31-120 Kraków
Poland
krzysztof.loboda@uj.edu.pl

#	EN [official version in EUR-Lex]	PL [official version in EUR-Lex]	Microsoft Translator NMT	Amazon Translate NMT	#
1	REGULATION (EU) No 1308/2013 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL	ROZPORZĄDZENIE PARLAMENTU EUROPEJSKIEGO I RADY (UE) NR 1308/2013	ROZPORZĄDZENIE PARLAMENTU EUROPEJSKIEGO I RADY (UE) NR 1308/2013	ROZPORZĄDZENIE PARLAMENTU EUROPEJSKIEGO I RADY (UE) NR 1308/2013	1
2	of 17 December 2013	z dnia 17 grudnia 2013 r.	z dnia 17 grudnia 2013 r.	z dnia 17 grudnia 2013 r.	2
3	establishing a common organisation of the markets in agricultural products and repealing Council Regulations (EEC) No 922/72, (EEC) No 234/79, (EC) No 1037/2001 and (EC) No 1234/2007	ustanawiające wspólną organizację rynków produktów rolnych oraz uchylające rozporządzenia Rady (EWG) nr 922/72, (EWG) nr 234/79, (WE) nr 1037/2001 i (WE) nr 1234/2007	ustanawiająca wspólną organizację rynków produktów rolnych i uchylająca rozporządzenia Rady (EWG) nr 922/72, (EWG) nr 234/79, (WE) nr 1037/2001 i (WE) nr 1234/2007	ustanawiające wspólną organizację rynków produktów rolnych i uchylające rozporządzenia Rady (EWG) nr 922/72, (EWG) nr 234/79, (WE) nr 1037/2001 i (WE) nr 1234/2007	3
4	THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION,	PARLAMENT EUROPEJSKI I RADA UNII EUROPEJSKIEJ,	PARLAMENT EUROPEJSKI I RADA UNII EUROPEJSKIEJ,	PARLAMENTU EUROPEJSKIEGO I RADY UNII EUROPEJSKIEJ,	4
5	Having regard to the Treaty on the Functioning of the European Union, and in particular the first	uwzględniając Traktat o funkcjonowaniu Unii Europejskiej, w szczególności jego art. 42 akapit pierwszy i art. 43 ust. 2,	uwzględniając Traktat o funkcjonowaniu Unii Europejskiej, w szczególności jego art.	uwzględniając Traktat o funkcjonowaniu Unii Europejskiej, w szczególności jego art. 42 akapit pierwszy i art. 43 ust. 2,	5

	subparagraph of Article 42 and Article 43(2) thereof,				
6	Having regard to the proposal from the European Commission,	uwzględniając wniosek Komisji Europejskiej,	uwzględniając wniosek Komisji Europejskiej,	uwzględniając wniosek Komisji Europejskiej,	6
7	After transmission of the draft legislative act to the national parliaments,	po przekazaniu projektu aktu ustawodawczego parlamentom narodowym,	Po przekazaniu projektu aktu ustawodawczego parlamentom narodowym	Po przekazaniu projektu aktu ustawodawczego parlamentom krajowym,	7
8	Having regard to the opinion of the Court of Auditors (1),	uwzględniając opinię Trybunału Obrachunkowego (1),	uwzględniając opinię Trybunału Obrachunkowego \[1],	uwzględniając opinię Trybunału Obrachunkowego [1],	8
9	Having regard to the opinions of the European Economic and Social Committee (2),	uwzględniając opinię Europejskiego Komitetu Ekonomiczno-Społecznego (2),	uwzględniając opinie Europejskiego Komitetu Ekonomiczno-Społecznego \[2],	uwzględniając opinie Europejskiego Komitetu Ekonomiczno-Społecznego [2],	9
10	Having regard to the opinion of the Committee of the Regions (3),	uwzględniając opinię Komitetu Regionów (3),	uwzględniając opinię Komitetu Regionów \[3],	uwzględniając opinię Komitetu Regionów [3],	10
11	Acting in accordance with the ordinary legislative procedure (4),	stanowiąc zgodnie ze zwykłą procedurą ustawodawczą (4),	stanowiąc zgodnie ze zwykłą procedurą ustawodawczą \[4],	stanowiąc zgodnie ze zwykłą procedurą ustawodawczą [4],	11
12	Whereas:	a także mając na uwadze, co następuje:	a także mając na uwadze, co następuje:	a także mając na uwadze, co następuje:	12
13	Article 59	Artykuł 59	Artykuł 59	Artykuł 59	13
14	Delegated powers	Przekazane uprawnienia	Uprawnienia delegowane	Uprawnienia delegowane	14

15	In order to ensure that the aid referred to in Article 58 finances the pursuit of the aims referred to in Article 152, the Commission shall be empowered to adopt delegated acts in accordance with Article 227 concerning:	W celu zapewnienia finansowania z pomocy, o której mowa w art. 58, realizacji celów, o których mowa w art. 152, Komisja jest uprawniona do przyjmowania zgodnie z art. 227 aktów delegowanych dotyczących:	W celu zapewnienia, że pomoc określona w art.	W celu zapewnienia, że pomoc, o której mowa w art. 58, finansuje realizację celów, o których mowa w art. 152, Komisja jest uprawniona do przyjmowania zgodnie z art. 227 aktów delegowanych dotyczących:	15
16	(a) | aid applications, including rules on deadlines and accompanying documents;	a) | wniosków o przyznanie pomocy, w tym przepisów dotyczących terminów i dokumentów towarzyszących;	a) | wniosków o pomoc, w tym przepisów dotyczących terminów i dokumentów towarzyszących;	a) | wnioski o przyznanie pomocy, w tym zasady dotyczące terminów i dokumentów towarzyszących;	16
17	(b) | rules on eligible hop areas and the calculation of the amounts to be paid to each producer organisation.	b) | przepisów dotyczących kwalifikujących się obszarów uprawy chmielu i obliczania kwot, które mają być wypłacone każdej organizacji producentów.	b) | zasady dotyczące kwalifikujących się obszarów chmielu oraz obliczanie kwot, które mają być wypłacone każdej organizacji producentów.	b) | zasady dotyczące kwalifikujących się obszarów chmielu oraz obliczanie kwot, które mają zostać wypłacone każdej organizacji producentów.	17
18	Article 60	Artykuł 60	Artykuł 60	Artykuł 60	18
19	Implementing powers in accordance with the examination procedure	Uprawnienia wykonawcze zgodnie z procedurą sprawdzającą	Uprawnienia wykonawcze zgodnie z procedurą kontroli	Uprawnienia wykonawcze zgodnie z procedurą sprawdzającą	19
20	The Commission may adopt implementing acts laying down the measures	Komisja może przyjmować akty wykonawcze ustanawiające środki niezbędne do stosowania	Komisja może przyjąć akty wykonawcze ustanawiające środki niezbędne do stosowania	Komisja może przyjmować akty wykonawcze ustanawiające środki niezbędne do stosowania	20

	necessary for the application of this Section concerning the payment of aid.	niniejszej sekcji dotyczącej wypłaty pomocy.	niniejszej sekcji dotyczące wypłaty pomocy.	niniejszej sekcji w odniesieniu do wypłaty pomocy.	
21	Those implementing acts shall be adopted in accordance with the examination procedure referred to in Article 229(2).	Te akty wykonawcze przyjmuje się zgodnie z procedurą sprawdzającą, o której mowa w art. 229 ust. 2.	Te akty wykonawcze przyjmuje się zgodnie z procedurą sprawdzającą, o której mowa w art.	Te akty wykonawcze przyjmuje się zgodnie z procedurą sprawdzającą, o której mowa w art. 229 ust. 2.	21
22	CHAPTER III	ROZDZIAŁ III	ROZDZIAŁ III	ROZDZIAŁ III	22
23	Scheme of authorisations for vine plantings	System zezwoleń na nasadzenia winorośli	System zezwoleń na nasadzenia winorośli	System zezwoleń na nasadzenia winorośli	23
24	Article 61	Artykuł 61	Artykuł 61	Artykuł 61	24
25	Duration	Czas trwania	Długość	Czas trwania	25
26	The scheme of authorisations for vine plantings established in this Chapter shall apply from 1 January 2016 to 31 December 2030, with a mid-term review to be undertaken by the Commission to evaluate the operation of the scheme and, if appropriate, make proposals.	System zezwoleń na nasadzenia winorośli ustanowiony w niniejszym rozdziale stosuje się od dnia 1 stycznia 2016 r. do dnia 31 grudnia 2030 r.; Komisja przeprowadzi przegląd śródokresowy w celu ewaluacji funkcjonowania systemu oraz, w stosownych przypadkach, przedstawi wnioski.	System zezwoleń na nasadzenia winorośli ustanowiony w niniejszym rozdziale stosuje się od dnia 1 stycznia 2016 r. do dnia 31 grudnia 2030 r., przy czym Komisja dokona przeglądu śródokresowego w celu oceny funkcjonowania systemu i, w stosownych przypadkach, złożenia wniosków.	System zezwoleń na nasadzenia winorośli ustanowiony w niniejszym rozdziale stosuje się od dnia 1 stycznia 2016 r. do dnia 31 grudnia 2030 r., przy czym Komisja ma przeprowadzić przegląd śródokresowy w celu oceny funkcjonowania systemu i, w stosownych przypadkach, przedstawienia propozycji.	26