Machine translation quality in an audiovisual context

Burchardt, A., Lommel, A., Bywood, L., Harris, K. and Popovic, M.

This is an author's accepted manuscript of an article published in Target, 28 (2), pp. 206-221, 2016. The final definitive version is available online at:

http://dx.doi.org/10.1075/target.28.2.03bur

Machine translation quality in an audiovisual context

Aljoscha Burchardt (DFKI, Berlin)
Arle Lommel (DFKI, Berlin)
Lindsay Bywood (University College London)
Kim Harris (text&form/DFKI, Berlin)
Maja Popović (Humboldt-Universität zu Berlin)

Abstract

The volume of Audiovisual Translation (AVT) is increasing to meet the rising demand for data that needs to be accessible around the world. Machine Translation (MT) is one of the most innovative technologies to be deployed in the field of translation, but it is still too early to predict how it can support the creativity and productivity of professional translators in the future. Currently, MT is more widely used in (non-AV) text translation than in AVT. In this article, we discuss MT technology and demonstrate why its use in AVT scenarios is particularly challenging. We also present some potentially useful methods and tools for measuring MT quality that have been developed primarily for text translation. The ultimate objective is to bridge the gap between the tech-savvy AVT community, on the one hand, and researchers and developers in the field of high-quality MT, on the other.

Keywords: Machine translation, translation quality, evaluation, audiovisual translation

1. Introduction

Audiovisual translation (AVT) has become a fundamental necessity in the 21st century. Media trends such as VHS and LaserDiscs have come and gone, and translation tools have progressed from typewriters to fully integrated real-time web-based translation environments. Our world is becoming ever smaller, while the demand for information in every corner of the globe is growing. Consequently, the sheer volume of data that needs to be accessible in most regions and languages of the world is rising dramatically: every minute, 300 hours of video material is

being uploaded to YouTube.[1] Even assuming that only a small fraction of this content is of interest to a broader global audience, the effort required to publish it in multiple languages is a tremendous challenge. This has been recognized and acknowledged by research bodies and governments that have supported early-adopter projects involving automatic AV translation. Such projects include MUSA[2] and eTITLE,[3] which have used rule-based MT combined with translation memory to investigate the potential of these tools for AVT; SUMAT,[4] which has trained statistical machine translation engines on subtitles in seven bi-directional language pairs and performed an extensive evaluation of the resulting MT quality; EU-Bridge,[5] which has focused on further advancing the state-of-the-art in automatic speech recognition combined with MT with a view to applying this technology in several domains, including AVT; HBB4ALL,[6] which, although mainly focused on accessibility, has carried out research into the reception of automatic interlingual subtitles; and ALST,[7] a project whose aim was to implement existing automatic speech recognition, speech synthesis and MT technologies in audio description and voice-over, part of which included quality assessment of voice-over scripts produced using MT and post-editing.

The emergence of new technologies has also had a significant impact on the translation of text content. In technical translation, translation memory systems (TMs) and integrated terminology support have become indispensable when it comes to ensuring language consistency and streamlining the translation process. Automatic (or machine) translation technology (MT) is one of the most recent developments in the translation equation, and it is still too soon to know just how and to what extent this technology will support the creativity and productivity of professional translators in the future. However, MT is certainly more widely used in text translation than in AV translation, where its application is, as yet, rare.

Machine translation output generally requires substantial editing effort to be fit for publishing. Its quality depends on factors such as language pair, domain and genre, and similarity of the text to be translated and the material for which the machine has been optimised. EC-funded research on improving MT output has a long history. The most recent research

[1] https://www.youtube.com/yt/press/en/statistics.html

[2] http://sifnos.ilsp.gr/musa/index.html

[3] http://www.upf.edu/glicom/en/proyectos/proyectos_finalizados/e_title.html

[4] http://cordis.europa.eu/fp7/ict/language-technologies/project-sumat_en.html

[5] https://www.eu-bridge.eu/

[6] http://www.hbb4all.eu/

[7] http://ddd.uab.cat/record/137941?ln=en

projects in this field include QTLaunchPad,[8] QTLeap,[9] and QT21,[10] as well as applied research projects involving industry such as MMT.[11]

The use of MT is increasingly popular for 'gisting' (information-only translation) through free online systems such as Google Translate or Bing Translator. Google alone automatically translates roughly as much content in one day as all professional translators translate in an entire year, and is used by more than 200 million people every month.[12]

This type of translation is not only helpful for users in search of information on the internet but also for intelligence services and other bodies that need to determine which documents are relevant and require higher-quality translation. As the purpose of gisting translation is different from that of high-quality translation destined for publication, MT systems built for the former are not well suited for supporting professional translators in the latter (although they are used by translators today, often without being acknowledged as a resource).

The goals of this article are twofold. First, we discuss MT technology and why its application to AVT scenarios is particularly challenging. Second, we present some methods and tools for measuring MT quality that could prove useful to the AVT community – tools that have been developed for text translation. The ultimate objective is to bridge the gap between two worlds: that of the AVT community and that of researchers and developers in the field of high-quality MT. Closer cooperation between these two constituencies will promote innovation and improvement in the implementation of MT technologies, eventually providing access to increasing amounts of multimodal content in as many languages as possible.

In Section 2, we provide a high-level overview of the technical ingredients of MT systems that should help the reader when we explain the limitations and prospects of using MT in the context of AVT in Section 3. Section 4 provides an overview of tools and techniques for measuring MT quality. Section 5 closes this article with a short summary.

[8] http://www.qt21.eu/launchpad/

[9] http://qtleap.eu/

[10] http://www.qt21.eu/

[11] http://www.modernmt.eu/

[12] http://googleblog.blogspot.de/2012/04/breaking-down-language-barriersix-years.html

2. Background: Statistical Machine Translation in a nutshell

This section gives a very brief introduction to the technical components of MT systems in order to provide a basis for discussion in subsequent chapters. *Statistical MT systems* (SMT) such as Google Translate, Microsoft Translator, and open-source Moses systems represent the most widely used approach to MT today.[13] These systems use complex algorithms that learn how to transfer strings from one language to another using the probabilities derived from parallel bilingual texts. The basic components of such systems are:

1. A *phrase table*, a database containing words or phrases in the target language and the probability that they correspond to words or phrases in the source language.
2. A *re-ordering model* with probabilities for different word orders in the two languages.
3. A *monolingual language* model containing probabilities for sequences of words (n-grams) in the target language.

The statistical probabilities are learnt automatically through the analysis of large parallel corpora that contain sentences in the source language and their respective (human) translations in the target language. Simply put, these probabilities are estimated as relative frequencies of bilingual/monolingual words and phrases in the given texts where phrases are defined as simple groups of words, without taking into account any linguistic aspects. Essentially, the components learn how words they have seen have been translated, how the word order of the source and target language differed in these translations, and what words are likely to appear next to each other in the target language.

As a general rule, the more training material that is available, the better the translation results will be. The more similarities between the training material (domain, sentence structure and length, etc.) and the texts to be translated, the higher the translation quality. Ten to twenty thousand training sentences may produce good results for some applications, text types and language pairs, while others may require much more material to achieve useful output.

In this statistical translation paradigm, the complex interplay of the different components produces translations that can sometimes be puzzling at first sight, as in (1), where the polarity of the German question has been reversed:

[13] *Rule-based MT systems* such as SYSTRAN and Lucy LT do not play a major role for translating AV content.

(1) Source: *Was stimmt?* [What is right?]
Online MT: What is wrong?

It is very difficult to trace why a certain translation has been produced by MT algorithms. In the example above, the most probable reason is that that the *translation probability* for *stimmt* was erroneously influenced by the more frequent appearance of the negated *stimmt nicht* in the training data. In this case, it is purely accidental that, while the above translation itself conveys the opposite meaning on its own, it may be semantically acceptable in certain contexts. It is also possible that the given translation appeared in the training data.

One common misconception is to think of the statistical systems anthropomorphically and observe that, for example, the system 'did not see' that X is plural, or that it 'missed' an embedded sentence, etc. The systems (in their simplest and very common form) do not have any explicit linguistic intelligence whatsoever: they do not 'know' what a part of speech is or what negation is, for example.

While the basic principles are easily explained, SMT systems are highly sophisticated, both in terms of mathematical and algorithmical complexity, as well as in computing power and the required data resources. SMT is an active field of research that is exploring several approaches to improve the state of the art, such as adding linguistic and semantic knowledge to systems and enhancing the mathematical models.

2.1 The challenge of assessing MT Quality

MT systems are frequently confused with translation memory systems (TMs). In a way, MT can be seen as an extension of TM technology. However, while TMs only retrieve *existing* translations previously produced by humans, MT is able to flexibly generate *new* translations based on these translations.

One major practical drawback is the fact that the usefulness or 'fitness for purpose' of a given machine translation is difficult to estimate. As a consequence, post-editors are often confronted with MT output that is not useful, which decreases productivity and efficiency. To remedy this situation, a research approach, known as *quality estimation*, is currently being developed to assess the quality of MT output (see Section 4).

It is interesting to note that, despite the relatively high level of technological support for the AV translator (e.g., specialist subtitling software and software for the preparation and recording of dubbing scripts), the actual act of translation remains fairly unsupported in this

domain. AV translators do not routinely use TMs, despite their widespread use in text translation.

2.2 What MT does best and why

Like other technologies, MT improves with use. If workflows are set up well, the selection or rejection of MT suggestions by professional translators and the respective post-edits serve as feedback for continued system development and improvement.

Machine translation works especially well in cases where the source and target languages are quite similar in terms of structure, morphology, concepts, etc. For example, a Spanish-to-Portuguese system will generally be easier to develop and will provide higher quality than a system translating from Swahili to Japanese. Another decisive factor is the availability of large amounts of parallel bilingual texts that are similar enough to the material to be translated with respect to domain, text type, etc., so that the systems can extract all of the relevant information.

By nature, MT has a better chance of success processing grammatical and syntactic phenomena that are within a short distance of one another in the sentence (such as noun-verb agreement in English), than it does processing phenomena that span longer distances, such as verb phrases in German whose components can be split over entire clauses. Likewise, phenomena that require extralingual knowledge like discourse and world knowledge when translating (e.g., ambiguous pronouns) exceed the capabilities of the current state-of-the-art.

Interestingly, however, shorter distances do not generally improve results for AV translation using MT, as the spoken text often relies on inferences and context and contains many condensed and incomplete phrases and expressions, as in (2):

(2)	AV transcription:	Your mother's house?
	MT (DE):	*Ihrer Mutter Haus*? [Your mother house?]
	Full sentence:	Are we meeting at your mother's house?
	MT (DE):	*Treffen wir uns im Haus Ihrer Mutter?* [Are we meeting in your mother's house?]

Although neither German machine translation is perfect, the one based on the short AV transcription is incomprehensible while the translation of the more verbose original sentence conveys the meaning quite well. A similar result is seen in (3):

(3) AV transcription: *Wieder ein Wochenende vorbei*. [Another weekend gone by.]
MT: Again a weekend pass.
Full sentence: *Das Wochenende ist wieder vorbei*. [The weekend is over once more.]
MT: The weekend is over again.

3. Problems impacting the automatic translation of subtitles

AVT poses a number of particular challenges for MT.[14] Most MT systems have been developed using large databases of translated *written* (vs. originally *spoken*) texts that are grammatically correct, with proper punctuation, capitalization, etc. In addition, MT is used most frequently for technical texts where the vocabulary and structures are highly predictable and often restricted.

By contrast, AVT of subtitles and dubbing scripts, by its very nature, deals with written representation of spoken dialogue and has characteristics that can make it difficult for MT. (Note, however, that dubbing scripts are "written to be spoken," a phenomenon termed "prefabricated orality" by Chaume [2004]). This situation creates a whole set of new challenges for MT. In Section 3, we will illustrate some of these challenges as starting points for more systematic future investigations.

If the MT engine translating the text has been trained on traditional written text, the features used in spoken text may not be represented accurately in the training data and the engine will therefore have no relevant examples from which it can produce an accurate translation. It is therefore important for the quality of MT that any system intended for use on AV material be trained using AV texts. One issue that arises here is the relative difficulty of obtaining such a corpus, particularly in lesser-resourced language pairs (Bywood et. al 2013).

3.1 Domain and genre

[14] In this article, we will concentrate on the translation of subtitles. We will not address the issue of condensing text. Although automatic text summarization and shortening techniques exist, we believe it is too early to discuss them.

One problem facing the use of MT in AVT is that AVT is an 'open' domain, in that audiovisual content covers the broadest spectrum of subject matter possible, from a very precise and lexically challenging technical documentary to tabloid celebrity news. As a result, even large amounts of content are often insufficient to satisfyingly calculate predictability owing to the inconsistent nature of the content at all levels, including grammar, structure and vocabulary.

3.2 Lack of visual context

Competent AVT requires a knowledge of the visual context in which the source text is embedded (Díaz Cintas and Remael 2007, 51), information to which an MT system has no access. A simple example of this is the translation of the English word *hello* into Italian. In most cases, this word would be translated as *ciao*, as a greeting during an informal meeting for example, whereas *pronto* would be the correct translation for a greeting over the phone. In this example, the previous utterance may provide some contextual clues about the respective scenario, but MT technology that uses such inter-sentential context cues for translation is still in its infancy. Example (4) is another example taken from the SUMAT project, this time from Swedish subtitles:

(4)	Source:	The reactions I got in the market **stalls** with the fishermen.
	Translation:	*Reaktioner på marknaden **toaletter** med fiskarna.*
	Back translation:	Reactions on the market **toilets** with the fishermen.

Here the word *stall* has been wrongly translated by the system as *toilet* in the absence of the context that is available to the professional translator.

3.3 Oral style

As is well documented (e.g., Rubin 1978), spoken language and written text have many differences. For example, spoken language has a much higher percentage of grammatically incomplete phrases, is more likely to rely on actual physical context (e.g., using a pronoun such as *that* to refer to a noun), and is generally more informal. In addition, the lexicon of spoken text (in general) is different from written text, with much more use of verbal discourse markers (such as *you know*, *uh-huh*, or *right?*) that are not generally found in written text, in addition

to slang and colloquialisms. Take, for example, (5), which has been translated using an online MT system:

(5) Source: *Was für 'n Mädel?* [What girl?]
MT: What for ' s girl?

If we return the condensed pronoun to its full grammatical form, as in (6), then the online MT system provides us with a translation which, whilst not correct, is easily post-edited to form a correct subtitle by the simple removal of the article *a*.

(6) Source: *Was für ein Mädel?*
MT: What a girl? (correct: What girl?)

Closely related are colloquialisms such as that seen in (7):

(7) Source: Guy seemed **high as a kite** every time I met him.
MT: *Guy schien **hoch wie ein Drachen**, jedesmal wenn ich ihn traf.*
[Guy seemed high as a kite (child's toy), every time I met him.]
Human: *Jedes Mal, wenn ich ihn traf, schien er **voll zugedröhnt** gewesen zu sein.* [Every time I met him, he seemed to be totally stoned.]

One possible solution is to use corpora made up of subtitles, thus capturing many of the disfluencies, colloquialisms, and oral features that prove problematic for systems trained on general written text. Although such corpora are not widely available, when they are, systems trained on them show promise, as in (8) from the SUMAT project:

(8) Source: I'll have a go.
MT (SUMAT): *Je vais essayer.* [I will try.]
Online MT: *Je vais avoir un aller.* [I will have a to go.]

SMT is actually well suited for dealing with these issues, if there is sufficient training data available.

3.4 Lack of context

Closely related to the previous point, spoken text tends to consist of short segments. While not problematic per se (shorter segment length generally correlates with better translation quality), spoken segments are more likely to rely on context that is not available within a single segment to be intelligible. Since MT engines generally do not look beyond single segments, this important context will not be accessible to them. For example, consider the spoken-style text in (9):

(9) Source: You're asking about the accident? Well, there was a man on 42nd Street. Down by the bridge. Big fellow. He saw it.

Online MT: *Sie sind über den Unfall zu fragen? Nun, es war ein Mann auf der 42. Straße. Down by die Brücke. Big Kollegen. Er sah es.*
[You are about the accident to ask? Well, there was a man on 42nd Street. Down by (untranslated) the bridge. Big (untranslated) colleague. He saw it.]

The *it* in the final sentence of the spoken example does not have context within a single segment, and the system translates it as *es* (neuter gender), rather than the correct *ihn* or *den* (masculine). Such results are common when the translation of a word depends on a context that may be a number of sentences removed from the word. For similar reasons, it also partially translates *Big fellow* as *Big Kollegen*, which might imply that the individuals are work colleagues, even though the context makes it clear the speaker does not know the man. A more appropriate translation would be something like *großer Kerl* [big bloke]. By contrast, a written description would probably be more like (10):

(10) Source: There was a big man on 42nd street by the bridge who saw the accident.

Online MT: *Es war ein großer Mann auf der 42. Straße an der Brücke, die den Unfall gesehen.*
[There was a big man on 42nd Street on the bridge, who seen the accident.]

While (10) shows other problems – like using the feminine relative pronoun *die* instead of the masculine *der* to refer to the man and a missing main verb in the relative clause ('seen' vs. 'has seen') – it is generally more intelligible than (9).

Similarly, English *you* can be translated as German *Sie* (formal), *du* (informal singular), *ihr* (informal plural) or *man* (impersonal pronoun), and the choice often depends on macro-level context (e.g., knowledge of who is speaking with whom) that cannot be easily derived purely from the source text. An example can be seen in (11), where the German pronoun *sie* can mean *she* or *they* and the MT system gets the wrong one (although here the verb *hat* makes it clear which meaning is correct):

(11)	Source:	Denn **sie** hat dich auf die Idee gebracht.
		[Because she gave you the idea.]
	MT:	For **they** gave you the idea.
	Human:	Because **she** put you up to it.

4. Measuring Machine Translation quality

A translation must be 'fit for purpose,' that is, it must fulfil certain objectives as determined by the parties involved. For much user-generated content, the level of expectation is much lower than it is for television broadcast or DVD publishing. As in the text world, it is important to be clear about what form 'acceptable quality' takes in each case. The processes, tools and metrics used to measure translation quality (if it is measured at all) for a particular purpose vary depending on the desired outcome of the task and the constituency performing it.

4.1 Quality evaluation in MT Research

The assessment of MT quality in research is almost always based on input by professional translators or post-editors in various forms. These are the most common forms of evaluation currently applied:

1. *Automatic evaluation of MT output based on algorithmic comparisons of MT output with (professional) human reference translations* (e.g., Papineni et al. 2002, Banerjee and Lavie

2005). This method is fast and repeatable and can apply and improve upon automatic metrics using previous results.

2. *Automatic evaluation without human reference translations for the given MT output, commonly known as quality estimation* (e.g., Shah et al. 2013). This method requires a trained system (based on human translations) and uses rankings or scores assigned by professional translators (to previous alternative translations) to improve quality estimation metrics.
3. *Ranking of MT output from different systems by human evaluators.* Ranking is performed, e.g., by NLP researchers in some of the shared tasks[15] of the Workshop of Statistical Machine Translation (WMT). Avramidis et al. 2012 report a study where ranking is performed by professional translators. This method provides information about the relative performance of certain systems or system variants.
4. *Post-editing of MT output by human evaluators.* Post-editing is performed, e.g., by NLP researchers in some of the shared tasks of WMT. Avramidis et al. 2012 report a study where post-editing is performed by professional translators. This method measures different aspects of post-editing efficiency (time, number of edits, etc.) and processes the acquired data, for instance, to analyse the types of edits that are most frequent (e.g., word order, morphology, insertion, etc.).
5. *Error annotation of the MT output by human evaluators*. (see, e.g., Vilar et al. 2006 where annotation is performed by NLP researchers; in Lommel et al., 2014, annotation is performed by professional translators). This method can provide detailed error analysis of the MT output, including specific accuracy and fluency errors in addition to word order and distance. This information can then be used to improve MT systems.

All methods have been and continue to be applied in the case of MT for subtitling. As described above, it is not particularly easy to acquire parallel corpora containing AV material. There are issues around the ownership of subtitles and dubbing scripts which make collecting quality corpora of any size problematic, and companies are hesitant to share material with researchers. For this reason, the evaluation of MT using reference translations can be a challenge. Quality estimation has been used successfully in the SUMAT project, where previously annotated subtitles were used to train the system to isolate poor-quality subtitles and discard them, supplying the post-editors simply with a box containing the text “FILT” instead (Etchegoyhen

[15] http://www.statmt.org/wmt15/

et al. 2014). All the other forms of evaluation described above were also used in this project, which performed the largest scale evaluation of MT for subtitles to date. However, a particular issue facing AVT is the scarcity of post-editors to provide input for the respective metrics. Since MT is not commonly used in AVT, there is a lack of trained post-editors who can work with AV texts, although training programmes are on the horizon and research (De Sousa et al. 2011) has demonstrated considerable promise in integrating MT, human translator technology, and post-editing.

The first two evaluation methods described above are used to evaluate and estimate the overall performance of a particular system and language pair, often for a particular domain, as well as to compare systems to one another. Automatic evaluation metrics that fall into these categories include BLEU scores (Papineni et al. 2002), F-scores (Popović 2011b), METEOR (Banerjee and Lavie 2005), TER and other similar metrics. They can also be used to estimate certain quality aspects. Quality estimation without a professionally translated reference is a relatively new and challenging approach to MT quality evaluation (e.g., Shah et al. 2013). Roughly speaking, the idea is to build a system that uses methods (i.e., algorithms, linguistic tools, training data, etc.) different from those the MT engine itself used, to assess the MT output. The systems have been designed for different tasks such as automatic ranking of several alternative MT outputs, or estimating the post-editing effort or overall quality of a given MT output. Usually, the systems are trained on human-generated data such as existing human rankings, gradings of system outputs, etc. Automatic analysis of post-edits is also employed (see Popović 2011a) and can provide insights.

The NER model (Romero-Fresco and Martinez, forthcoming) implemented in the NERstar tool is one of the first AVT specific metrics. It was not designed for assessing MT, but for assessing the accuracy of re-spoken subtitles when compared to the original spoken text. The model is appealing as it only takes into account two types of errors: those made by the human re-speaker and those made by the speech-to-text system. Additional weights indicate the severity of the respective errors. While this tool is a good candidate for every-day quality assurance in computer-aided subtitling, it does not lend itself to assessing MT quality with the goal of improving the MT engines. For this, we need more fine-grained analysis of MT errors.

4.2 Multidimensional Quality Metrics (MQM)

One promising approach for the close analysis of errors in AV translation that comes from text translation work is the Multidimensional Quality Metrics (MQM) framework (Lommel et al.

2014)[16]. Originally developed in the EU-funded QTLaunchPad project and based on an examination of existing translation quality metrics, MQM was created to address the need for a way to objectively describe translation errors that was also flexible enough to address specific needs. MQM consists of over 100 translation quality *issue types* that can be used to describe specific problems in translated texts. These issue types are arranged hierarchically, to allow for different levels of granularity in describing issues found in the text.

Figure 1 shows a relatively complex MQM metric used to do detailed analysis of MT errors. The issues in ***bold italic*** are ones that are not in the basic MQM set but instead represent custom user extensions. They do not contradict MQM because they simply provide additional granularity and can be considered types of their parent issue. In this case they provide additional information on problems with 'function words,' such as prepositions, articles, and 'helper' verbs. This metric focuses heavily on grammatical features and on specific types of Accuracy problems.

INSERT FIG 1 HERE

By contrast, Figure 2 shows a much simpler metric that might be suitable for evaluating MT used for AV:

INSERT FIG 2 HERE

This metric is intended for AV translation in general (not just for MT). It adds *Style* (a basic MQM type) that would be highly relevant to AV translation, and removes a number of categories unlikely to be particularly relevant. It also has much less emphasis on *Grammar* and two custom types are added:

a. *Contextual*, for translations that are contextually incorrect
b. *Timing*, for cases where the translations appear at the wrong time.

[16] http://www.qt21.eu/launchpad/content/multidimensional-quality-metrics

As can be seen, the advantages of MQM are that it provides a standardized vocabulary for describing errors and that it allows users to create task-specific metrics (e.g., a metric for evaluating news captions is likely to be quite different from one used to evaluate legal translations). In addition, MQM can be extended to support issues not present in the master vocabulary. MQM is implemented in the open-source translate5[17] editor and is being used and further developed within the QT21 project. Current work on MQM aims to extend it for additional translation types, including AV translation.

5. Summary

In this article, we have tried to pave the way for closer cooperation between AVT specialists and MT experts in order to promote research on higher quality MT for AVT. We have provided some background on the purpose of MT technology currently used in text translation and have discussed some of the challenges when using this technology for translating subtitles. In conclusion, we have provided an overview of MT quality evaluation methods and proposed an extension to the Multidimensional Quality Metrics MQM to include AV-specific issue types.

Acknowledgments

Work on this article has received partial funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 645452 ("Quality Translation 21").

[17] http://www.translate5.net/

References

Avramidis, Eleftherios, Aljoscha Burchardt, Christian Federmann, Maja Popović, Cindy Tscherwinka, and David Vilar. 2012. “Involving Language Professionals in the Evaluation of Machine Translation.” In *Proceedings of LREC 2012*, 1127-1130. http://www.lrec-conf.org/proceedings/lrec2012/index.html, 22.12.2015.

Banerjee, Satanjeev, and Alon Lavie. 2005. “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments.” In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, ed. by Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, 65-72. Michigan, MI: University of Michigan.

Bywood, Lindsay, Martin Volk, Mark Fishel, and Panayota Georgakopoulou. 2013. “Parallel Subtitle Corpora and their Applications in Machine Translation and Translatology.” In *Corpus Linguistics and AVT: in Search of an Integrated Approach*, special issue of *Perspectives: Studies in Translatology* 21 (4), 1-16.

Chaume, Frederic. 2004. *Cine y traducción*. Madrid: Cátedra.

De Sousa, Sheila C. M., Wilker Aziz, and Lucia Specia. 2011. “Assessing the Post-Editing Effort for Automatic and Semi-Automatic Translations of DVD Subtitles.” In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, ed. by Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, and Nikolai Nikolov, 97–103. http://www.aclweb.org/anthology/R11-1014.pdf. Accessed December 22, 2015.

Díaz-Cintas, Jorge, and Aline Remael. 2007. *Audiovisual Translation, Subtitling*. Manchester: St. Jerome.

Etchegoyhen, Thierry, Lindsay Bywood, Mark Fishel, Panayota Georgakopoulou, Jie Jiang, Gerard van Loenhout, Arantza del Pozo, Mirjam Sepesy Maucec, Anja Turner, and Martin Volk. 2014. “Machine Translation for Subtitling: A Large-Scale Evaluation.” In *Proceedings of LREC 2014*, 46-53. http://www.lrec-conf.org/proceedings/lrec2014/index.html , 22.12, 2015.

Lommel, Arle, Aljoscha Burchardt, and Hans Uszkoreit. 2014. “Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics.” In *Tradumàtica: tecnologies de la traducció* 0 (12): 455-463.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. “BLEU: A Method for Automatic Evaluation of Machine Translation.” In *Proceedings of the 40th Annual*

Meeting of the Association for Computational Linguistics, 311–318. http://dl.acm.org/citation.cfm?id=1073083&picked=prox. Accessed December 22, 2015.

Popović, Maja. 2011a. "Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output." *The Prague Bulletin of Mathematical Linguistics* 96: 59-68.

Popović, Maja. 2011b. "Morphemes and POS Tags for N-gram Based Evaluation Metrics." In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 104-107. file:///Users/SRP/Downloads/ngrams.pdf. Accessed December 22, 2015.

Romero-Fresco, Pablo and Juan Martínez Pérez. Forthcoming. "Accuracy Rate in Live Subtitling – the NER Model." In *Audiovisual Translation in a Global Context: Mapping an Ever-changing Landscape*, ed. by Jorge Díaz Cintas, and Rocío Baños Pinero. London: Palgrave Macmillan. http://hdl.handle.net/10142/141892 (draft). Accessed November 4, 2015.

Rubin, Ann D. 1978. "A Theoretical Taxonomy of the Differences between Oral and Written Language." *Center for the Study of Reading Technical Report* 35.

Shah, Kashif, Eleftherios Avramidis, Ergun Biçicic, and Lucia Specia. 2013. "QuEst – Design, Implementation and Extensions of a Framework for Machine Translation Quality Estimation." *The Prague Bulletin of Mathematical Linguistics* 100: 19-30.

Vilar, David, Jia Xu, Luis Fernando d'Haro, and Hermann Ney. 2006. "Error Analysis of Statistical Machine Translation Output." In *Proceedings of LREC 2006*, 697–702. file:///Users/SRP/Downloads/2lrec06_errorAnalysis.pdf. Accessed December 22, 2015.

Authors' addresses

Aljoscha Burchardt
DFKI GmbH
Alt-Moabit 91c
10559 BERLIN
Germany

aljoscha.burchardt@dfki.de

Arle Lommel
DFKI GmbH
Alt-Moabit 91c
10559 BERLIN
Germany

arle.lommel@gmail.com

Kim Harris
DFKI GmbH
Alt-Moabit 91c
10559 BERLIN
Germany

kim_harris@textform.com

Lindsay Bywood
Centre for Translation Studies (CenTraS)
University College London
50 Gordon Square, Room 206
London WC1H 0PQ
England

lindsay.bywood.13@ucl.ac.uk

Maja Popović
Institut für Anglistik und Amerikanistik
Humboldt-University zu Berlin
Unter den Linden 6
10099 BERLIN
Germany

maja.popovic@hu-berlin.de