Target

Machine Translation Quality in an Audiovisual Context

--Manuscript Draft--

Manuscript Number:		
Full Title:	Machine Translation Quality in an Audiovisual Context	
Short Title:		
Article Type:	Article	
First Author:	Aljoscha Burchardt, Ph.D.	
Other Authors:	Lindsay Bywood Kim Harris Arle Lommel Maja Popovic	
Corresponding Author:	Aljoscha Burchardt, Ph.D. DFKI Berlin, Berlin GERMANY	
Funding Information:	DG Connect (BE) (Grant Agreement No. 645452.)	Dr. Aljoscha Burchardt Kim Harris Arle Lommel
Section/Category:	Special issue	
Keywords:	Audiovisual Translation; Machine Translation; Translation Quality Assessment	
Abstract:	Audiovisual Translation (AVT) is increasingly common 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 of this article 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.	
Author Comments:	This article is submitted for: AUDIOVISUAL TRANSLATION: Theoretical and Methodological Challenges TARGET - Special issue 28:2 (2016) EDITORS: Sara Ramos Pinto (University of Leeds) Yves Gambier (University of Turku)	
Order of Authors Secondary Information:		

Machine Translation Quality in an Audiovisual Context (working title II)

1 Introduction

Audiovisual translation (AVT) has become a fundamental need 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 increasingly small, while the demand for information in every corner of the globe is ever 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 including MUSA², eTitle³, and SUMAT.⁴

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 early to predict 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 projects in this field include QTLaunchPad⁵, QTLeap⁶, and QT21⁷, as well as applied research projects involving industry such as MMT⁸.

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.⁹

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 the fact 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 potentially useful methods and tools to the AVT community for measuring MT quality—ones that have been developed for text translation. The ultimate objective is to bridge the gap between two worlds: that of the tech-savvy 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

1
2
3
4 in the implementation of MT technologies, eventually providing access to increasing amounts of
5
6
7 multimodal content in as many languages as possible.
8
9

10 In Section 2, we provide a high-level background about the technical ingredients of MT systems
11
12 that should help the reader when we explain the limitations and prospects of using MT in the
13
14 context of AVT in Section 3. Section 4 provides an overview of tools and techniques for
15
16 measuring MT quality. Section 5 closes this article with a short summary.
17
18
19

20 2 Background: Statistical Machine Translation in a Nutshell 21 22 23

24 This section gives a very brief introduction to the technical components of MT systems in order
25
26 to provide a basis for discussion in subsequent chapters. *Statistical MT systems* (SMT) such as
27
28 Google Translate, Microsoft Translator, and open-source Moses systems represent the most
29
30 widely used approach to MT today.¹⁰ These systems use complex algorithms that learn how to
31
32 transfer strings from one language to another using the probabilities derived from parallel
33
34 bilingual texts. The basic components of the most widely used systems are:
35
36
37

- 38 1. A *phrase table*, a database containing words or phrases in the target language and the
39
40 probability that they correspond to words or phrases in the source language
41
42
- 43 2. A *re-ordering model* with probabilities for different word orders in the two languages.
44
45
- 46 3. A *monolingual language* model containing probabilities for sequences of words (n-
47
48 grams) in the target language.
49
50

51 The statistical probabilities are learnt automatically by analysing of huge parallel corpora that
52
53 contain sentences in the source language and their respective (human) translations in the target
54
55 language. Simply put, these probabilities are estimated as relative frequencies of
56
57
58 bilingual/monolingual words and phrases in the given texts where phrases are defined as simple
59
60
61
62
63
64
65

groups of words, without taking into account any linguistic aspects. Basically, 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:

(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, this translation may have appeared repeatedly in the training material. It could also be 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.

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

1
2
3
4 sentence, etc. The systems (in their simplest and very common form) do not have any explicit
5
6 linguistic intelligence whatsoever: they do not “know” what a part of speech is or what negation
7
8 is.
9

10
11
12 While the basic principles are easily explained, SMT systems are highly sophisticated, both in
13
14 terms of mathematical and algorithmical complexity, as well as in computing power and the
15
16 required data resources. SMT is an active field of research that explores several approaches to
17
18 improve the state of the art, such as adding linguistic and semantic knowledge to systems and
19
20 enhancing the mathematical models.
21
22
23

24 25 2.1 The Challenge of Assessing MT Quality 26

27
28 MT systems are frequently confused with translation memory systems (TMs) that contain
29
30 previous translations by professionals. In a way, MT can be seen as an extension of TM
31
32 technology. However, while TMs only retrieve *existing* translations previously produced by
33
34 humans, MT is able to flexibly generate *new* translations based on these translations.
35
36
37

38
39 One major practical drawback is the fact that usefulness or “fitness for purpose” of a given
40
41 machine translation is difficult to estimate. As a consequence, post-editors are often confronted
42
43 with MT output that is not useful, which decreases productivity and efficiency. To remedy this
44
45 situation, a research approach, known as *quality estimation*, is currently being developed to
46
47 assess the quality of MT output (see Section 4).
48
49
50

51
52 It is interesting to note that, despite the relatively high level of technological support for the AV
53
54 translator (e.g., specialist subtitling software and software for the preparation and recording of
55
56 dubbing scripts), the actual act of translation remains fairly unsupported in this domain. AV
57
58 translators do not routinely use TMs, despite their widespread use in text translation.
59
60
61
62
63
64
65

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 is 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 ambiguous pronouns exceeds 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?

Full sentence: Are we meeting at your mother's house?

MT (DE): Treffen wir uns im Haus Ihrer Mutter?

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.

MT: Again a weekend pass.

Full sentence: Das Wochenende ist wieder vorbei.

MT: The weekend is over again.

3 Problems impacting the automatic translation of subtitles

AVT poses a number of particular challenges for MT.¹¹ 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.

1
2
3
4 If the MT engine translating the text has been trained on traditional written text, the features used
5
6 in spoken text may not be represented accurately in the training data and the engine will
7
8 therefore have no relevant examples from which it can produce an accurate translation. It is
9
10 therefore important for the quality of MT that any system intended for use on AV material be
11
12 trained using AV texts. One issue that arises here is the relative difficulty of obtaining such a
13
14 corpus, particularly in lesser-resourced language pairs.
15
16
17
18
19

20 3.1 Domain and genre

21
22 One problem facing the use of MT in AVT is that AVT is an open domain, i.e. audiovisual
23
24 content covers the broadest spectrum of subject matter possible, from a very precise and lexically
25
26 challenging technical documentary to tabloid celebrity news. As a result, even large amounts of
27
28 content are often insufficient to satisfyingly calculate predictability owing to the inconsistent
29
30 nature of the content at all levels, including grammar, structure and vocabulary.
31
32
33
34
35

36 3.2 Lack of visual context

37
38 Competent AVT requires a knowledge of the visual context in which the source text is
39
40 embedded, information to which an MT system has no access. A simple example of this is the
41
42 translation of the English word *hello* into Italian. In most cases, this word would be translated as
43
44 *ciao*, as a greeting during an informal meeting for example, whereas *pronto* would be the correct
45
46 translation for a greeting over the phone. In this example, the previous utterance may provide
47
48 some contextual clues about the respective scenario, but MT technology that uses such inter-
49
50 sentential context cues for translation is still in its infancy. Example (4) is another example taken
51
52 from Swedish subtitles:
53
54
55
56
57
58
59
60
61
62
63
64
65

- 1
2
3
4 (4) Source: The reactions I got in the market **stalls** with the fishermen.
5
6 Translation: Reaktionen på marknaden **toaletter** med fiskarna
7
8 Back translation: Reactions on the market **toilets** with the fishermen.
9

10
11 Here the word *stall* has been wrongly translated by the system as *toilet* in the absence of the
12
13 context that is available to the professional translator.
14
15

16 17 18 3.3 Oral style 19 20

21 As is well documented (e.g., Rubin 1978), spoken language and written text have many
22
23 differences. For example, spoken language has a much higher percentage of grammatically
24
25 incomplete phrases, is more likely to rely on actual physical context (for example using a
26
27 pronoun such as *that* to refer to a noun), and is generally more informal. In addition, the lexicon
28
29 of spoken text (in general) is different from written text, with much more use of verbal discourse
30
31 markers (such as *you know*, *uh-huh*, or *right?*) that are not generally found in written text, in
32
33 addition to slang and colloquialisms. Take, for example, (5), which has been translated using an
34
35 online MT system.
36
37
38
39
40

- 41 (5) Source: Was für 'n Mädel?
42
43 MT: What for 's girl?
44
45
46

47 If we return the condensed pronoun to its full grammatical form, as in (6), then the online MT
48
49 system provides us with a translation which, whilst not correct, is easily post-edited to form a
50
51 correct subtitle by the simple removal of the article *a*.
52
53
54

- 55 (6) Source: Was für ein Mädel?
56
57 MT: What a girl? (correct: What girl?)
58
59
60
61
62
63
64
65

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.

Human: Jedes Mal, wenn ich ihn traf , schien er **voll zuge-dröhnt** gewesen zu sein.

One possible solution is to use corpora made up of subtitles, thus capturing many of the dysfluencies, 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.

Online MT: Je vais avoir un aller.

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.

1
2
3
4 Online MT: Sie sind über den Unfall zu fragen? Nun, es war ein Mann auf der 42.
5
6 Straße. Down by die Brücke. Big Kollegen. Er sah es.
7
8
9

10 The *it* in the final sentence of the spoken example does not have context within a single segment,
11
12 and Google translates it as *es* (neuter gender), rather than the correct *ihn* or *den* (masculine).
13

14 Such results are common when the translation of a word depends on a context that may be a
15
16 number of sentences removed from the word. For similar reasons, it also partially translates *Big*
17
18 *fellow* as *Big Kollegen*, which might imply that the individuals are work colleagues, even though
19
20 the context makes it clear the speaker does not know the man. (A more appropriate translation
21
22 would be something like *großer Kerl* ‘big bloke’). By contrast, a written description would
23
24 probably be more like (10):
25
26
27
28
29

30 (10) Source: There was a big man on 42nd street by the bridge who saw the accident.
31

32 Online MT: Es war ein großer Mann auf der 42. Straße an der Brücke, die den Unfall
33
34 gesehen.
35
36
37

38 While (10) shows other problems (like using the feminine relative pronoun *die* instead of the
39
40 masculine *der* to refer to the man), it is generally more intelligible than (9).
41
42
43

44 Similarly, English *you* can be translated as German *Sie* (formal), *du* (informal singular), *ihr*
45
46 (informal plural) or *man* (impersonal pronoun), and the choice often depends on macro-level
47
48 context (e.g., knowledge of who is speaking with whom) that cannot be easily derived purely
49
50 from the source text. An example can be seen in (11), where the German pronoun *sie* can mean
51
52 *she* or *they* and the MT system gets the wrong one (although here the verb *hat* makes it clear
53
54 which meaning is possible):
55
56
57
58
59
60
61
62
63
64
65

(11) Source: Denn **sie** hat dich auf die Idee gebracht.

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”, i.e., it must fulfill 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.* This method is fast and repeatable and can apply and improve upon automatic metrics using previous results.
2. *Automatic evaluation without professional reference translations for the given MT output, commonly known as quality estimation.* 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 professional translators.* This method provides information about the relative performance of certain systems (or system variants).
4. *Post-editing of MT output by professional translators.* This method measures different aspects of post-editing efficiency (time, number of edits, etc.) and processes the acquired data, e.g., 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 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 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 {REMOVED CITATION} 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 ({REMOVED CITATION}).

However, a particular issue facing AVT is the lack of availability of post-editors to provide input for the second and third forms of metric. 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.

1
2
3
4 The first two methods are used to evaluate and estimate the overall performance of a particular
5 system and language pair, often for a particular domain, as well as to compare systems to one
6 another. Automatic evaluation metrics that fall into these categories include BLEU scores
7 (Papineni et al. 2002), F-scores (Popović 2011b), METEOR (Banerjee and Lavie 2005), TER
8 and the like. They can also be used to estimate certain quality aspects. Quality estimation without
9 a professionally translated reference is a relatively new and challenging approach to MT quality
10 evaluation (e.g. Shah et al. 2013). Roughly speaking, the idea is to build a system that uses
11 methods (i.e., algorithms, linguistic tools, training data, etc.) different from those the MT engine
12 itself used to assess the MT output. The systems have been designed for different tasks such as
13 automatic ranking of several alternative MT outputs, estimating the post-editing effort or overall
14 quality of a given MT output. Usually, the systems are trained on human data such as existing
15 human rankings, gradings of system outputs, etc. Automatic analysis of post-edits is also
16 employed (see Popović 2011a) and can provide insights.

17
18 The NER model¹² implemented in the NERstar tool is one of the first AVT specific metrics. It
19 was not designed for assessing MT, but for assessing the accuracy of re-spoken subtitles when
20 compared to the original spoken text. The model is appealing as it only takes into account two
21 types of errors: those made by the human re-speaker and those made by the speech-to-text
22 system. Additional weights indicate the severity of the respective errors. While this tool is a good
23 candidate for every-day quality assurance in computer-aided subtitling, it does not lend itself to
24 assessing MT quality with the goal of improving the MT engines. For this, we need more fine-
25 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) framework¹³. 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.

Figure 1. MQM metric for analyzing MT

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

Figure 2. Proposed MQM metric for AVT

This metric 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:

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

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 translate5¹⁴ 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 provided an overview of MT quality evaluation methods and proposed an extension to the Multidimensional Quality Metrics MQM to include AV-specific issue types.

6 Acknowledgments

to be added in the final manuscript

7 References

- 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*, vol. 29, pp. 65–72.
- Chaume, Frederic. 2004. *Cine y traducción*. Madrid: Cátedra.
- Lommel, Arle, Aljoscha Burchardt, Maja Popovic, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014. "Using a New Analytic Measure for the Annotation and Analysis of MT Errors on Real Data." *Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT)*, pp. 165–72.
- 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*, pp. 311–18. Association for Computational Linguistics.
- 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*, pp. 104-107. Association for Computational Linguistics.

Rubin, Ann D. 1978. "A Theoretical Taxonomy of the Differences between Oral and Written Language." *Center for the Study of Reading Technical Report; no. 035*.

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.

{REMOVED REFERENCE}

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 <http://www.qt21.eu/launchpad/>

⁶ <http://qt leap.eu/>

⁷ <http://www.qt21.eu/>

⁸ <http://www.modernmt.eu/>

⁹ <http://googleblog.blogspot.de/2012/04/breaking-down-language-barriersix-years.html>

¹⁰ *Rule-based MT systems* such as SYSTRAN and Lucy LT do not play a major role for translating AV content.

¹¹ In this article, we will concentrate on the translation of subtitles from one language into another. We thus assume that we can start from a human-quality subtitle in the source language that needs to be translated into one or more target languages. We will not address the issue of condensing text. Although text summarization and shortening techniques exist, we believe it is too early to discuss them.

¹² <http://www.speedchill.com/nerstar/index.php/ner-model.html>

¹³ <http://www.qt21.eu/launchpad/content/multidimensional-quality-metrics>

¹⁴ <http://www.translate5.net/>