universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

Comparing an Understanding of Translation Quality in Translation Studies and Machine Translation Quality Estimation

verfasst von / submitted by

Marita Schett, BA

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien, 2021 / Vienna 2021

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 070 331 351
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Translation Deutsch Spanisch
Betreut von / Supervisor:	Ass.-Prof. Mag. Dr. Dagmar Gromann, BSc

Acknowledgements

First of all, I would like to thank my supervisor Dagmar Gromann for offering me advice and support during all the stages of this thesis, from the conceptualisation until the final phase, and for her feedback that contributed to the successful completion of this thesis.

Moreover, I thank my family and friends for always believing in me and for lending a sympathetic ear whenever I need(ed) it. A big thank you also to my flatmates for providing me with food during phases of intensive writing and for putting up with me especially during those last few months that were not easy for all of us and during which we all spent most time at home.

And finally, thank you Bernadette for making these past three years at university so much better and for all the laughs we shared. Without your company reaching this point would have been so much harder and so much less fun. Thanks for pushing me through this time!

Table of Contents

List of Figures

List of Tables

1 Introduction

Ever since people of different languages started to interact with each other, translation has played an essential role in enabling communication and cleared the way for intercultural relations. For many centuries to follow, translation was performed either in spoken form or written on paper, until, in the 20th century, the invention of computers changed the work of translators profoundly.

Since then, researchers have been trying to break the "code" of language and teach computers how to translate between any language pair without human help. Even when Weaver, one of the pioneers of Machine Translation (MT), came to the result that "'perfect' translation is almost surely unattainable" (Weaver 1949: 11) in 1949 already, other scientists kept seeking the key to perfect MT. Since then, a lot has changed. Technologies have progressed substantially and a great many of MT types have been invented, have had their peaks, and have become outdated. For many years, MT systems were not considered useful for a broad public and even less for professional translators due to the low quality of their output. However, the new approaches in MT systems that evolved in the past three decades, most notably Statistical Machine Translation (SMT) and Neural Machine Translation (NMT), have shown big improvements in MT output quality and have therefore started to become widely applied in the translation industry as well as by non-professionals. Yet, the "holy grail of the MT world" (Bennett & Gerber 2003: 181), the creation of Fully Automatic High-Quality Translation (FAHQT), has not yet been achieved.

At the same time as MT became widely used, the tasks of professional translators slowly started to shift away from mere translations from scratch towards post-editing (PE) of MT output, causing the boundaries between human and machine translation to become more and more blurred (Castilho et al. 2018: 27). PE is usually paid less than translating from scratch since it is considered to involve less time and cognitive effort. However, as a matter of fact, many MT texts that post-editors receive are of quite bad quality and PE often takes more effort than a translation from scratch would have done. This situation is frustrating for post-editors as well as for clients because the post-editors get paid less for more work and the clients possibly need to wait longer than agreed or have to put up with translation products of lower quality. This issue, amongst others, has led to the development of Machine Translation Quality Estimation

(MTQE) tools, that can estimate the quality of an MT text without any human reference translation (Specia & Shah 2018: 202).

However, the publications concerning tools for QE of MT that have been published so far show that those tools are usually developed by computational linguists who, of course, are familiar with the domain of language itself, but who might not consider the quality criteria for good translations that researchers in the field of translation studies agree on. Thus, it is of great interest to examine the quality criteria used in the development of MTQE tools as well as the quality criteria prevalent in translation studies in order to find differences and similarities in the way that these two disciplines define the quality of translation. The research question that is posed in this thesis therefore is:

> *To what extent are the quality criteria used for developing quality estimation tools for machine translation different to the quality criteria prevailing in translation studies?*

The intention beyond examining this question is to point out the problem that even though professional translators and project managers working in the translation industry are the main users of MTQE systems, they are not at all part of their development process. Moreover, the research done for this thesis is meant to demonstrate to what extent and how research on translation quality in translation studies can contribute to the development process of MTQE methods. It is expected that due to their experience in the field and various research initiatives, translators can contribute important aspects that are also interesting and useful for computational linguists and can help them get a broader view of what translation quality can mean. Moreover, translators can possibly not only contribute to the definition of the theoretical concept of translation quality but can also bring in their practical experience in the field. As a consequence, the MTQE community could direct its future research more towards methods and technologies that already exist in the industry and aim at combining their own methods and tools with the ones already in use.

In order to be able to give an answer to the beforementioned research question, first, an extensive literature review is conducted, and its results are presented and analysed in the comparative analysis that is performed as a second step. A detailed explanation of the research method is presented in Chapter 2. Chapter 3 discusses the quality of translation: after a brief general introduction, a distinction is made between the quality concept in translation studies

and the quality concept in the translation industry and several approaches and models are presented for both fields. The fourth chapter covers MT, starting with the definition of different degrees of computer-human interaction in translation and the history of MT. Then, the different types of MT are introduced and the applications of MT in translation workflows are discussed. Chapter 5 deals with Machine Translation Quality Assessment (MTQA) and its different shapes. Manual TQA methods and automatic TQA methods are distinguished and their subcategories, as well as examples, are presented. The matter of Chapter 6 is error classification and analysis, which can also be divided into manual and automatic error classification. Again, examples are being given in order to make the explained concepts more comprehensible. In Chapter 7, the focus shifts away from translation towards PE. A brief introduction on the history of PE is given before the types and levels of PE and some PE guidelines are presented. Chapter 8 illustrates the second big subject of this thesis, that is MTQE. First, the possible applications of MTQE are examined, then the different levels of MTQE are presented, and finally, examples of MTQE systems are given. Chapter 9 gives a brief insight into related work and in Chapter 10 the comparative analysis of the results presented in Chapters 3 to 8 is performed. In Chapter 11, the results of the comparative analysis are discussed and, eventually, in Chapter 12, conclusions are drawn from the results of the literature review and the comparative analysis in order to give an answer to the research question and to propose ideas for further research on the topic.

2 Research Method

As has been stated in the introduction, the aim of this thesis is to evaluate the state-of-the-art of MTQE research and to find out whether or not some quality aspects that are prevailing in translation studies might be missing in state-of-the-art QE tools for MT. To serve this purpose, an extensive literature review is conducted on the topic and the results are presented in Chapters 3 to 8.

According to Hart (1998: 1), a literature review gives a better understanding of the topic and provides a better overview on what kind of research has already been conducted in the field and what the key issues are. He defines a literature review as:

> […] the selection of available documents […] on the topic, which contain information, ideas, data and evidence written from a particular standpoint to fulfil certain aims or express certain views on the nature of the topic and how it is to be investigated, and the effective evaluation of these documents in relation to the research being proposed (Hart 1998: 13).

A literature review can serve different purposes depending on the research project. It can be conducted in order to provide historical background, to give an overview of the current context, to discuss relevant theories and concepts, to introduce relevant terminology, to describe related research in the field, and/or to provide supporting evidence for an issue (Ridley 2008: 16f). In a basic literature review, the current state of knowledge on a topic is presented and a position on the state of this knowledge is taken. This is the whole thesis. If the purpose of the literature review is, however, to uncover an issue for further (empirical) research, it is referred to as an advanced literature review (Machi & McEvoy 2012: 2). In the case of this thesis, the purpose of the literature review is to give an overview of the current context as well as to discuss relevant theories and concepts, precisely the concept of quality in the two fields and its application in quality estimation. Since the literature review is also part of the research method of the thesis, a basic literature review will be conducted. For the realisation of the literature review, the instructions given by the step-by-step guide of Machi and McEvoy (2012) is followed. They propose a six-step process that is pictured in Figure 1.

The first step of the literature review is finding a topic by defining an interest that could be a potential subject suitable for study. However, the process is not linear, and a first general idea of the topic will become clearer by starting to look for literature and by reading (Machi & McEvoy 2012: 33f). In the case of this thesis, this step has already been taken and examining the quality criteria used in translation studies and MTQE research as well as comparing them with each other has been chosen to be the topic of the thesis. The second step, that consists of

the literature search, involves the selection of possible material, searching for inclusions in the material and finally refining of the topic statement. The tools that can be used in the literature search are scanning the literature, skimming, and gathering the central ideas in order to better understand the context. The literature search serves as preparation for step three since it explores and catalogues the subject in order to be able to then develop the arguments (Machi & McEvoy 2012: 59f). In step three, the argument of the thesis is developed based on evidence found in the literature. All sides of the question have to be presented in order to justify the argument that is presented in this thesis. The development of the argument serves to organize and form the step that is to come next: the survey of the literature (Machi & McEvoy 2012: 83). The fourth step, which is the literature survey, consists of discovering what is known about the subject of research. The information found in the literature should be assorted either chronologically or thematically or in a combination of the two. It prepares the following literature critique by documenting and discovering issues in the literature (Machi & McEvoy 2012: 109). In the fifth step, the evidence found in the literature survey step is analysed and arranged in a way that the thesis statement can be justified and the research question can later be answered (Machi & McEvoy 2012: 6). The writing done in the first five steps becomes the foundation of the final step, the writing of the review. The written review should illustrate the research that has been done in the literature review process in a way that the intended audience can understand (Machi & McEvoy 2012: 6f).

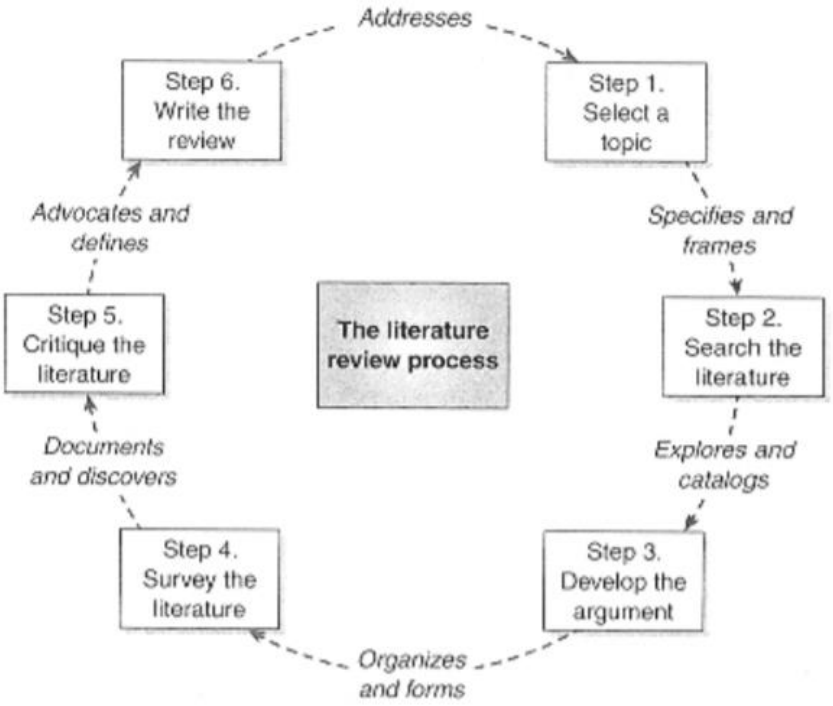

Figure 1: Literature Review Model (Machi & McEvoy 2012: 5)

The platforms that were chosen for the literature search were the u:search tool of the Vienna University Library since as a student of the University of Vienna I have access to their entire collection of publications, as well as Google Scholar since it offers one of the biggest openly accessible collection of academic papers. In order to get an overview of publications on the topic of Machine Translation Quality Assessment and Machine Translation Quality Estimation as well as to be able to properly differentiate between the two concepts, both terms were looked up on u:search as well as on Google Scholar limiting the results to publications from 2015 to date to get an overview of recent approaches in both fields. Another search key was that the publications should be written either in English or in German so that they could be understood and used for this thesis. Of all the approaches on translation quality in translation studies and the approaches to quality estimation, the ones that were referred to most frequently by other publications were chosen to be presented in this thesis since they can thus be considered the most relevant and most widely applied ones. For the chapters on machine translation, its assessment, error classification, and post-editing, the most recent publications giving an overview on the respective topic were chosen in order to include also the latest trends in these fields such as neural machine translation and its assessment. The information that was found in the literature is then organised thematically in a logical order, starting with translation quality in translation

studies and the translation industry since they have been debated for the longest time and some of the central concepts date back to the 1970s and 1980s but are still considered state-of-the-art. Then machine translation is introduced in order to provide a basis for the understanding of the subsequent chapters on translation quality assessment, error classification, post-editing, and quality estimation.

Following the literature review and the presentation of its results, a comparative analysis is performed to find an answer to the research question that was put forward in the introduction. The quality criteria found during the literature review are extracted and put into a list. Then, the criteria found in translation studies and the criteria found in MTQE research are compared to find overlaps but especially to find differences between the two fields.

Since the simple reason that the literature review is already highly extensive and together with the comparative analysis provides interesting results for this thesis, an additional empirical evaluation is beyond the scope of this thesis. Besides, the two fields, quality research in the translation studies on the one and quality estimation of MT on the other hand, have never before been connected and compared in this way, at least not to the best of my knowledge, and therefore the literature review and its comparative analysis already represent a valuable contribution to the field of translation quality from two of its primary perspectives.

3 Quality of translation

The quality of translation has been a topic of much debate for a long time. One reason why it is so hard to find a universally valid definition of translation quality is that it is impossible to determine what exactly is a perfect translation, i.e. what is the goal that should be achieved. Translation scholars and translators, as well as end users of translated texts, will agree that it is, of course, possible to distinguish a good translation from a bad one, yet it is difficult to decide which translation is the best one out of many good options, and it is even more difficult to define which aspects contribute to a high-quality translation.

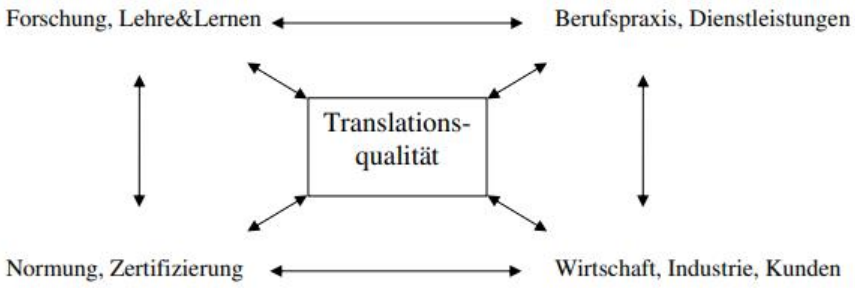

Figure 2: Context model of translation quality (Budin 2007: 59)

Budin (2007) stated that translation quality is always embedded in a broader context as pictured in the model in Fig. 2. Translation quality is positioned in the middle of the model with four influencing dimensions around it. The four dimensions are Research and Training (Forschung, Lehre & Lernen), Translation Services (Berufspraxis, Dienstleistungen), Standardisation and Certification (Normung, Zertifizierung), and Industry and Clients (Wirtschaft, Industrie, Kunden). They all determine translation quality and are determined by translation quality, but also interact with each other (Budin 2007: 59). Research and Training implies prenormative research to develop and improve translation metrics and specifications as well as translator training that involves practising the application of norms and standards in the translation process (Budin 2007: 59f). The dimension of Translation Services states that translation quality should be a central component of the work of translators. Standardisation and Certification implies that norms and standards should look at translation quality in its entirety to achieve acceptance in all parties involved. This can be accomplished by certification and frequent audits. Finally, Industry and Clients stands for convincing the end users of translation products of the importance of norms and standards (Budin 2007: 60).

In the following two sections, different approaches to the definition of translation quality are presented. First, a closer look at quality concepts established in the domain of translation studies is taken, then some metrics and standards used in the translation industry are presented.

3.1 The concept of quality and quality assessment in translation studies

The quality of translation is often perceived as a rather fuzzy concept that is assessed based on subjective judgement. For a long time, the evaluation of translation quality was left to the general public and it was as recently as the 1970s that translation studies began to intensify looking into the tasks and goals of translation criticism to improve the quality of translation services, to awaken the public desire for translation services of better quality, and to raise language awareness (Kaindl 2003: 373).

The concept of translation quality is related very closely to the status of the source text and the target text and thus to the translation type that is deployed. Generally, two different translation types can be distinguished: the first type stays very close to the source text and tries to keep as many of its elements as possible, whereas the second type focusses on the communicative function of a text and intends to achieve the same function in the target culture and for its recipients (Nord 2003: 142). During the history of translation studies, many different scholars have arrived at this distinction. Schleiermacher differentiates between *Verfremdung*, that is the first translation type that focusses on loyalty to the original, and *Einbürgerung*, that is adapting the translation to the norms of the target language and culture (Schreiber 2003: 151), Nida and Taber (1982: 14) distinguish *formal equivalence* (type 1) and *dynamic equivalence* (type 2), and House (1997: 66ff) speaks of *overt translation* (type 1) and *covert translation* (type 2). The choice of the translation type depends on the text type and the text's function and thus has to be decided individually for the translation project at hand. However, a high degree of specification often comes along with a translation that is tailored to the target-language readers, whereas in literary translation the loyalty to the original might play a bigger role (Schreiber 2003: 151). Since MT is usually used in domains of high specification, this thesis focusses more on translation concepts that belong to the second translation type.

Just as many different approaches to translation studies have emerged during the past few decades, a great number of different quality concepts and methods for translation assessment have been developed. While they look at translation quality from various perspectives, many of them share their core aspects. Mossop et al. (2020: 136) state that when it comes to the assessment of translation quality it is most important to ask oneself "is the translation accurate

and is it well written? A longer list might be: accurate, linguistically correct, smooth-reading, adapted to readers" (Mossop et al. 2020: 136). Chatzikoumi (2020: 138) arrives at a similar conclusion when she claims that the core criteria of translation quality

> […] can be summed up as (i) fluency in the target language, which includes grammaticality and naturalness; (ii) adequacy as in semantic and pragmatic equivalence between the source and the target text; and (iii) compliance with possible requester specifications (Chatzikoumi 2020: 138).

This conforms to the definition of translation quality brought forward by Koby et al. (2014), which states that a "quality translation demonstrates accuracy and fluency required for the audience and purpose and complies with all other specifications negotiated between the requester and provider, taking into account end-user needs" (Koby et al. 2014: 416).

This section aims at giving an overview of the most important concepts and assessment methods for translation quality. Based on Kaindl's (2003) work, the distinction between text typological approaches, pragmalinguistic approaches, functional approaches, and polysystematic approaches will be made here. Eventually, other approaches that do not fit into one of these categories will be presented, too.

3.1.1 Text typological approaches

The most important text typological approach to translation criticism was presented by Reiss (1971). According to her work, the most important principle of a translation is fidelity to the intent of the author of the original. Before an overall evaluation of the translated text can be made, it has to be examined from various different angles. First, the text type should be looked at, since it strongly influences the translation method, and then linguistic elements and non-linguistic factors should be considered too (Reiss 2014: 16).

The primary factor that determines which translation method to use is not the intended audience or the purpose of the text, but the text type (Reiss 2014: 17). Each text uses a certain function of language, which may either be representation, expression, or appeal. Depending on which function is dominant, Reiss (2014: 25) distinguishes between three text types: content-focused (or informative) texts, form-focused (or expressive) texts, and appeal-focused (or operative) texts, as can be seen in Fig. 3. Apart from these three function-based text types, she also determines the audiomedial type as a fourth text type. This type includes texts that are written to be spoken or that make use of an extra-linguistic medium (Reiss 2014: 27).

language function	representation	expression	persuasion
language dimension	logic	esthetics	dialogue
text type	content-focused (informative)	form-focused (expressive)	appeal-focused (operative)

Figure 3: Text types and their corresponding language function and dimension (Reiss 2014: 26)

The second perspective that needs to be examined by the translator as well as by the critic are the linguistic components of a text. This involves looking at the linguistic features of the original and their equivalents in the target language (Reiss 2014: 48). Translation is only possible because there are parallels between different languages on the level of *langue* which allows for the translator to find several possible equivalents on the level of *parole* from which s/he has to choose the optimal one in the translation process (Reiss 2014: 49). The linguistic components that must be examined are semantic elements, lexical elements, grammatical elements, and stylistic elements. In terms of the semantic elements, it is most important to achieve semantic equivalence between the source text and its translation. Thus, the linguistic context must be examined very thoroughly to avoid errors such as failure to recognise polysemous words and homonyms, misinterpretations, additions, omissions, and so on (Reiss 2014: 53). Regarding the lexical elements, the goal of the translation is to achieve adequacy. Since the vocabularies of two languages do not coincide completely, no word-for-word accuracy is required, however, the components of the source-language text have to be carried over adequately to the translated text (Reiss 2014: 57f). In terms of the grammatical elements, correctness should be reached in the translated text. Since the differences between the grammatical systems of different languages are often quite big, correctness is satisfied as soon as the translated text observes the grammar rules of the target language system and the relevant semantic and stylistic aspects of the grammatical structure in the original text have been understood correctly and have been rendered adequately in the translated text (Reiss 2014: 60). For the stylistic elements, correspondence between source and target text is the most important factor. The difference between colloquial, standard, and formal language must be recognised, compared, and echoed correctly (Reiss 2014: 63).

Finally, in the last step, the focus shifts away from the language towards extra-linguistic determinants. Reiss (2014: 67) defines them as follows:

> […] extra-linguistic factors enabling the author to make specific choices among the variety of means available in his mother language which would not only be intelligible to the reader or hearer, but under certain circumstances would even permit him to ignore certain linguistic means and still be understood by members of his language group (Reiss 2014: 67).

The extra-linguistic determinants can be divided into the immediate situation, the subject matter, the time factor, the place factor, the audience factor, the speaker factor, and affective implications (Reiss 1971: 71). A comprehensive analysis of the quality of a translated text must include all these factors that influence the translation process.

The immediate situation refers to single passages and moments in the source text where the translator has to place herself or himself into during the translation in order to choose the proper words to express the same semantic concepts. The critic or evaluator also has to imagine herself or himself in those situations to be able to judge whether the translator has chosen the right words not only lexically but also semantically (Reiss 2014: 69). The subject matter determinant indicates that the translator should be sufficiently familiar with the field of the text to be able to translate it adequately (Reiss 2014: 70f). The next two factors, the time and place factors can pose a big challenge for translators. The time factor is important if the source text is related to a particular period of time and contains antiquated words as well as morphological and syntactic structures. These should also be reflected in the translated text, especially if the text is of operative or expressive form (Reiss 2014: 71). The place factor relates to "all the facts and characteristics of the country and culture of the source language, and further also any associations of the scene where the actions described take place" (Reiss 2014: 74). This may confront the translator with culture-bound translation problems, since s/he may encounter abstract ideas that are not easily comprehensible for target-language readers (Reiss 2014: 76). The audience factor refers to the readers or hearers of the source-language text who the author had in mind when writing the original (Reiss 2014: 78). The translator's task is to "make it possible for the reader in the target language to see and understand the text in the terms of his own cultural context" (Reiss 2014: 79), which may sometimes require a sort of decoding of the original or finding a transposed equivalent in the target-language culture. The speaker factor involves all the aspects that influence the language of the author. In expressive texts, the speaker factor determines the style of the author, whereas it is less important in informative texts since their style is more determined by the content than by the author (Reiss 2014: 82). Finally, affective implications are all emotional determinants that affect lexis, style, and grammar. The translator needs to recognise them correctly and render them adequately in the target language (Reiss 2014: 83).

3.1.2 Pragmalinguistic approaches

In pragmalinguistic approaches to translation quality and translation criticism, pragmalinguistic aspects of translation are at the centre of attention (Kaindl 2003: 374). The most important pragmalinguistic approach is House's (1997) model for translation quality assessment. According to House's (1997) work, the essence of translation lies in "the preservation of "meaning" across two languages" (House 1997: 30). The meaning of a text has a semantic, a pragmatic, and a textual aspect and translation can therefore be defined as "the replacement of a text in the source language by a semantically and pragmatically equivalent text in the target language" (House 1997: 31). A target text is equivalent to its source text if its function, i.e., "the application or use which the text has in the context of a situation" (House 1997: 36), is equivalent to the function of the original text. Thus, in order to determine the function of a text at hand, it is important to not only look at the text itself but to also refer to its context. This is done by conducting a pragmalinguistic analysis of the source text and creating a textual profile. This procedure is displayed in Fig. 4.

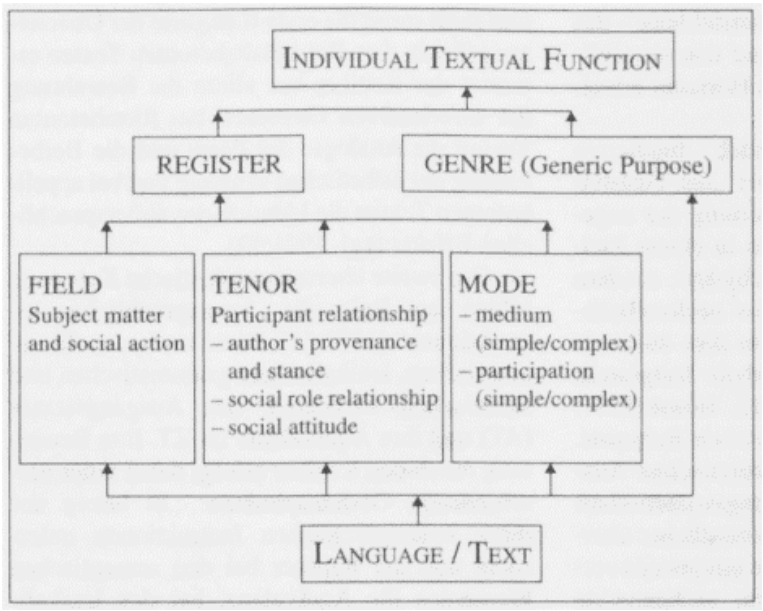

Figure 4: A model for assessing translation quality (House 1997: 108)

The basis of the pragmalinguistic analysis is the text itself, represented as "Language/Text" in Fig. 4. The text is then divided into two parts: register and genre (House 1997: 108). Register is again divided into three subcategories: field, tenor, and mode. Field refers to the subject matter of the text as well as its social action, whereas tenor describes the situational factors of the participants, such as the relationship between the participants, the author's provenance and

stance, the authority relationship between addresser and addressee, as well as the degree of proximity and the resulting degree of formality between them (House 1997: 108). Mode, on the other hand, refers to the medium, which can be either simple, e.g., written to be read, or complex, e.g. written to be spoken, as well as to the participation of addresser and addressee(s) in the communicative act (House 1997: 109). These three categories are analysed on a lexical, syntactic, and textual level. The second part, genre, is defined as "a socially established category characterized in terms of occurrence of use, source and a communicative purpose or any combination of these" (House 1997: 107) and refers to the text type (Kaindl 2003: 375). After considering all these aspects, the individual text function can be determined, which is composed of the register and the genre. The function can either be a cognitive-referential one that refers to the content of the text, or a non-cognitive, or emotive-expressive, one that produces a certain reaction in the addressee (House 1997: 35f).

The judgement of whether a translation is good or bad is then based on the equivalence between the aspects of the source text and the aspects of the translated text. The decision on which aspects and categories should be looked into the most detailed depends on the type of translation that is required. A covert translation is not marked as a translation of a source text in another language and enjoys the status of an original text in the target culture. An overt translation, on the other hand, is not a second original and a direct match of the original function of the source text is not possible since the source text has a unique status in the source culture or is linked to a specific, non-repeatable event. In this case, a cultural filter can be applied to achieve a similar second-level function without hiding the fact that the text is a translation (House 1997: 111ff).

3.1.3 Functional approaches

Whereas text typological and pragmalinguistic approaches focus rather on the source text, functional approaches emphasise that the translated text itself should be in the centre of attention, and only in the second instance it should be considered as a translation of a source text (Reiss & Vermeer 1984: 113).

The most well-known functional approach is certainly the skopos theory that was established by Reiss and Vermeer (1984). The Greek word *skopos* means purpose or objective, thus the skopos theory aims at preserving the purpose of a text in its translated version (Reiss & Vermeer 1984: 96). Reiss and Vermeer state that "[e]s ist wichtiger, daß ein gegebener Translat(ions)zweck erreicht wird, als daß eine Translation in bestimmter Weise durchgeführt wird"

(Reiss & Vermeer 1984: 100), that is achieving the purpose of a translation is more important than the translation method. In order to determine the required skopos of a translated text, it is necessary to know the intended addressee(s). Without this knowledge, it is impossible to know which purpose should be followed (Reiss & Vermeer 1984: 102).

Ammann (1990) developed a concept of translation criticism that is based on the skopos theory and aims at a target text-oriented perspective. The proposed procedure of translation criticism can be divided into five analysis steps: (1) determining the function of the translated text, (2) determining the intra-textual coherence of the translated text, (3) determining the function of the source text, (4) determining the intra-textual coherence of the source text, and (5) determining an inter-textual coherence between the source text and its translation. Coherence refers to the accordance of the content, the accordance of the form, as well as the accordance between content and form. Moreover, intended incoherence can also be part of coherence (Ammann 1990: 212).

Along the lines of the skopos theory, Ammann also states that for determining the function of the translated text it is essential to know the intended addressee(s). Their text comprehension can be captured by using the concept of *scenes and frames.* Scene refers to the conception that is created inside the human brain based on perceptions (Vermeer & Witte 1990: 51), whereas a frame is any phenomenon that may be understood to carry information of some kind (Vermeer & Witte 1990: 66).

3.1.4 Polysystematic approaches

Based on the translation concept established by Descriptive Translation Studies, van den Broeck (1985) developed a method for translation criticism that involves a three-step model of a comparative analysis of the source and the target text as well as an evaluation of the translation (van den Broeck 1985: 56).

The first step of the comparative analysis is to create an adequate translation by reconstructing the text-internal relations and functions of the source text (van den Broeck 1985: 57). This reconstruction includes the identification of elements that possess any function in the text like "phonic, lexical, and syntactical components, language varieties, figures of rhetoric, narrative and poetic structures, elements of text convention (text sequences, punctuation, italicizing, etc.), thematic elements and so on" (van den Broeck 1985: 58). This adequate translation which possesses all the functional elements of the source text is then compared to the actual translation. Potential differences in the elements of the two texts are called shifts and can be of two

different kinds. Obligatory shifts are forced by the rules of the system of the source language and culture, whereas optional shifts result from the translator's decisions (van den Broeck 1985: 57). Finally, the last step involves the description of the differences between the original and its translation to determine the degree of equivalence. This should be embedded in the polysystem of the source-language and target-language cultures (van den Broeck 1985: 58).

After this analysis, the results can be evaluated. During the evaluation, it is important to consider the norms of the translator, the translation method that the translator has chosen with regard to her or his intended addressees, as well as the translation strategies s/he has employed to achieve the intended purpose. Based on these aspects, the quality of the translated text can be rated (van den Broeck 1985: 61).

3.1.5 Other approaches

Apart from the beforementioned four approaches, a great many other approaches to determine and assess the quality of translations have been developed. These include, amongst others, a change in perspective (Mossop et al. 2020) or taking the comprehensibility of a text for the destined readers as the core criterion (Göpferich 2008).

Mossop et al. (2020) present a concept of translation quality that is based on the reviser's perspective. They determine a list of fourteen revision parameters, i.e. "the things a reviser checks for" (Mossop et al. 2020: 136), which they assign to five groups: Transfer, Content, Language, Presentation, and Specifications (Mossop et al. 2020: 136f). This list is presented in detail in Table 1. Consistency is mentioned as a possible fifteenth parameter, however, Mossop et al. (2020: 137) assert that consistency issues are always related to one of the other parameters, such as terminology or layout.

Table 1: Overview of revision parameters (Mossop et al. 2020: 136f)

Group	**Parameter**	**Definition of revision parameter**
Transfer	Accuracy	The translated text renders the content of the original adequately.
	Completeness	All the elements of the original have been translated to the target language and no omission or addition occurred.
Content	Logic	The translated text makes sense and contains no logical errors.

	Facts	The translated text contains no factual, conceptual, or mathematical errors.
Language	Smoothness	Each sentence's wording connects it with the wording of the previous sentence, every sentence is structured to focus on its new and relevant information, the relationships between different parts of a sentence are clear and there are no badly structured or overly wordy sentences.
	Tailoring	The language, tone, and register are adapted to the intended users, the use they will make of it, and the medium where the translation appears.
	Sub-language	The language is suited to the genre, the phraseology is the same as in original target-language texts on the same topic, and the terminology is correct.
	Idiom	Individual words are used in the correct meaning they have in the target language, all the word combinations are idiomatic, the translated text observes stylistic and rhetorical norms in the target language, and no words or phrases of rare or archaic meaning are used.
	Mechanics	The rules of grammar, punctuation, spelling, as well as in-house style and correct usage have been observed.
Presentation	Layout	All the genre rules for page layout have been observed and no problems with spacing, indentation, margins, columns, footnotes, graphics, or lists occurred.
	Typography	No typography problems, such as problems with bolding, underlining, font type or size, italicization, colour, or caps occurred.
	Organization	The structure and organisation of the document do not present any issues.
Specifications	ClientSpecs	The translated text complies with the client's specifications regarding terminology, layout, etc.
	EmployerPol	The translated text complies with the employer's policies regarding Translation Memories, spelling practices, etc.

An entirely different approach was brought forward by Göpferich (2008). She claims that an important criterion for a good text is that it can be understood by its readers. Even though Göpferich speaks of texts in general and does not limit her comprehensibility concept to translated texts, it can be deployed to translations as one of many quality criteria. However, it must always be kept in mind that the comprehensibility of a translation is not only influenced by the translator and her or his translation method but also strongly depends on the comprehensibility of the original. If the original lacks some essential comprehensibility factors, it is difficult to create a comprehensible translation too. Yet, in the best case, the original is a perfectly comprehensible text and Göpferich's comprehensibility factors can be deployed by the translator in the translation process.

As Fig. 5 shows, the comprehensibility of a text depends on text-internal factors as well as on external ones. The external factors include the communicative function of a text, which depends on the text's purpose (Zweck), its expected addressees (Adressaten) and their age, degree of education, social class, etc., as well as the sender (Sender) (Göpferich 2008: 156f). Apart from the communicative function, the text production parameters play an important role too. They include the mental denotation model (mentales Denotatsmodell), i.e., the picture that the sender has in mind while writing and that should be evoked in the recipient's mind when reading the text (Göpferich 2008: 159), the mental convention model (mentales Konventionsmodell), which refers to the conventions of a suitable text type that limit the possibilities of codification of the mental denotation model (Göpferich 2008: 160), the medium on which the text is transmitted (Medium), and potential legal or editorial guidelines that need to be observed (juristische und redaktionelle Richtlinien) (Göpferich 2008: 162).

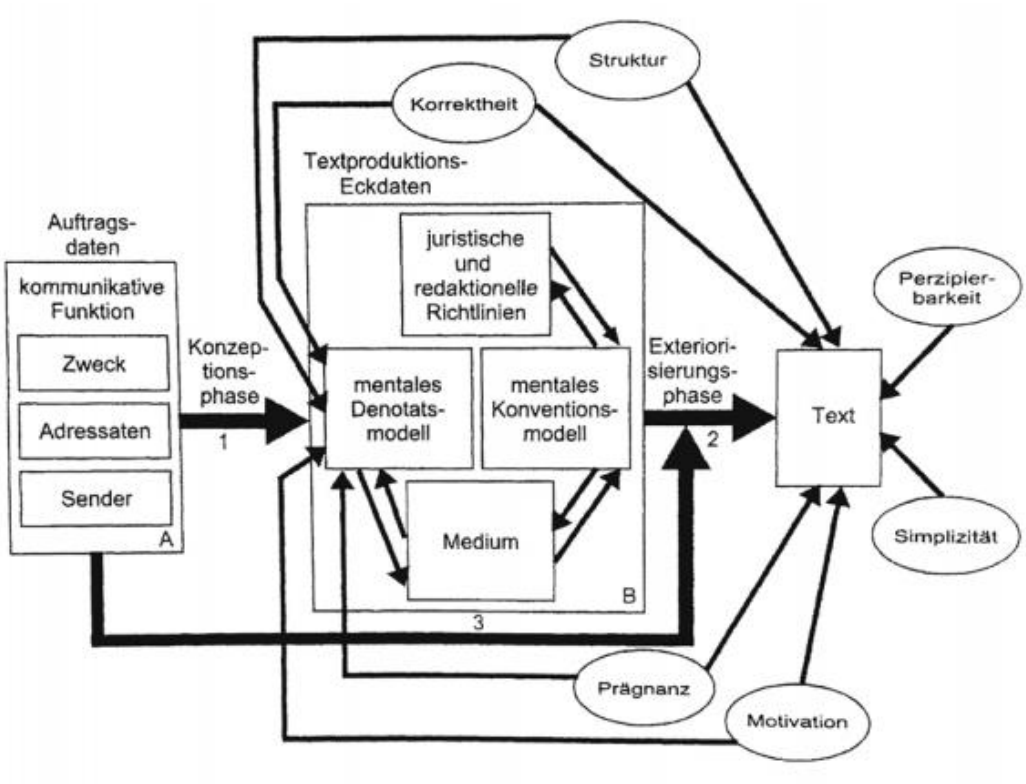

Figure 5: Reference frame for text comprehensibility (Göpferich 2008: 155)

Apart from these external factors that need to be taken into account in order to create a text that can be understood by its addressees, comprehensibility also depends on a range of text-internal factors, which Göpferich calls "Verständlichkeitsdimensionen" (Göpferich 2008: 163), that is dimensions of comprehensibility. These six dimensions are defined as follows:

1. *Concision (Prägnanz):* To assess the concision of a given text, two aspects need to be considered: the mental denotation model and its codification in the text. A text has reached maximum concision when (1) the mental denotation model only contains the minimum of information that is necessary to satisfy the communicative function of the text, and (2) it has then been expressed with the least possible number of characters (Göpferich 2008 163f). Possible cases, where full concision has not been reached, are texts where essential details of the denotation model are missing or superfluous details have been included, where long formulations are used instead of shorter ones that would carry the same amount of information, and where tautologies and redundancies appear (Göpferich 2008: 165ff).
2. *Correctness (Korrektheit):* The correctness of a text refers to whether the previous knowledge on the text's content of the addressee was judged correctly, whether the mental model of the described content was communicated correctly, whether the

text complies with genre conventions, whether the used medium is adequate, and whether there are linguistic mistakes in the text (Göpferich 2008: 168).

3. *Motivation (Motivation)*: The dimension of "Motivation" refers to the ability of a text to arouse interest and to keep it up while reading. This can be achieved, for example, by using stimulating vocabulary, vivid examples, or a personal style (Göpferich 2008: 169ff).
4. *Structure (Struktur)*: With this dimension, Göpferich (2008) refers to the structuring of the text's content. On the macro level, i.e., on the level of paragraphs, chapters, and bigger sections, good structure can be achieved by organising information according to the expected previous knowledge of the addressee(s) and by relating different units of information with each other (Göpferich 2008: 172ff). On the micro level, i.e. two contiguous sentences at most, the order of codification of the essential concepts and schemata as well as their logical connection should be assessed (Göpferich 2008: 174).
5. *Simplicity (Simplizität)*: The simplicity of a text refers to the lexical and grammatical simplicity of a text, as well as to the adequate degree of directness, the precision of the used words and sentence constructions, and consistency in lexis and syntax (Göpferich 2008: 176f). Lexical simplicity can be achieved by avoiding technical terms that might be unknown to the addressee wherever it is possible, and by avoiding long compound words (Göpferich 2008: 177f). Measures to achieve grammatical simplicity include, amongst others, using the active voice rather than the passive voice, using affirmative sentences rather than negative ones, and avoiding concentrations of prepositions and conjunctions in a sentence (Göpferich 2008: 178ff).
6. *Perceptibility (Perzipierbarkeit)*: The dimension of "Perceptibility" includes all layout and design characteristics, nonverbal and paraverbal aspects, and typographic features. These aspects contribute to how easily a text can be perceived by one's senses and processed cognitively (Göpferich 2008: 186).

3.2 The concept of quality in the translation industry

Translation quality is an important topic not only in academics but also for companies and institutions that make use of translation services in their daily workflows. Alongside official standards such as ISO 9001 (Beuth 2015) and ISO 17100 (Beuth 2016), many companies and institutions have also defined their own quality models for translation services. One of the best-known models is the one created by the Society of Automotive Engineers (SAE) in 1999 to be

able to compare, define, and provide proof of translation quality of automotive service information (Schütz 1999: 278). Another model, that was created as a general starting point from where quality criteria can be chosen depending on the task at hand, is the Multidimensional Quality Metric (MQM) that was put forward by Lommel et al. (2015).

3.2.1 The SAE J2450 quality model

The so-called SAE J2450 quality model contains five main quality factors: terminology, grammar, style, content, and structure (Schütz 1999: 280). For each of these quality factors, the SAE J2450 model assigns five quality criteria that are defined as follows:

- *Accuracy*: This criterion is met if the translation product provides the intended results or effects.
- *Compliance*: The translation product should comply with standards, conventions, regulation, etc. If a specialized corporate writing style or controlled language is required, the translation product should meet this requirement.
- *Consistency*: This criterion evaluates if the translation product maintains a certain level of human language performance and competence.
- *Understandability*: This criterion is met if the end user of the translation product is able to understand its content and can perform the described procedures and processes.
- *Interpretation*: The translation product should be clear and unambiguous and should leave no room for individual interpretation (Schütz 1999: 280).

If any of these criteria is not met, an error must have occurred. The SAE J2450 model distinguishes seven error types (see Table 2) and assigns a numeric weight to every error depending on whether it is classified as serious or minor. The total score, which decides on the quality of a translated text, is the sum of the error scores normalised by the number of words in the text (Hui 2017: 58).

Table 2: Overview of the error types in the SAE J2450 Translation Quality Metric

Error type	**Definition**	**Score**
Wrong term	1) The target-language (TL) term does not correspond to the client's glossary, or 2) the TL term does clearly not correspond to the standard translation, or 3) the TL term is not consistent with other translations, or 4) the TL term expresses another concept than the source-language (SL) term does (Hui 2017: 58).	Serious: 5 Minor: 2
Syntactic error	1) The SL term has been related to the wrong part of speech in the TL, or 2) the TL text's phrase structure is incorrect, or 3) the order of TL terms is not in accordance with the syntactic rules of the TL (Hui 2017: 59).	Serious: 4 Minor: 2
Omission	1) SL terms do not appear in TL text and therefore the semantics of the SL is missing, or 2) a graphic containing SL text is missing in the TL text (Hui 2017: 59).	Serious: 4 Minor: 2
Word structure or agreement error	1) The TL term appears in an incorrect morphological form, or 2) two or more TL terms do not agree even though they are supposed to (Hui 2017: 59).	Serious: 4 Minor: 2
Misspelling	1) The spelling of the TL term does not correspond to the spelling in the client's glossary, or 2) the spelling of the TL does not correspond to the spelling norms in the TL, or 3) the TL term is written in a wrong writing system (Hui 2017: 59f).	Serious: 3 Minor: 1
Punctuation error	An error related to the punctuation rules in the TL occurred (Hui 2017: 60).	Serious: 2 Minor: 1
Miscellaneous error	Errors that cannot be clearly assigned with one of the other error types can be marked as miscellaneous error (Hui 2017: 60).	Serious: 3 Minor: 1

3.2.2 The Multidimension Quality Metric (MQM)

Another proposal for an error typology that can be used by companies and institutions working in the translation industry is the Multidimensional Quality Metric (MQM) that was brought forward by Lommel et al. (2015). Its creators examined existing specifications and tools and merged their approaches into one general metric that seeks to offer a flexible catalogue of error types from which users can choose the types that are relevant for their needs and apply them to their project (Lommel et al. 2015). This allows for a selection in compliance with the requirements that are expected of the parties involved in the translation process. Checking for other errors, that are not part of the requirements, would be unnecessary work. In the MQM, the error types are grouped hierarchically, and the categories are represented in a tree-like structure. Fig. 6 shows the core issues, i.e., the most common issues, of the MQM (Lommel et al. 2015). They are grouped on two levels, the level-one-categories being accuracy, fluency, design, locale convention, style, terminology, and verity. Those categories can then be divided further into several subcategories. There are also many subcategories of level one and two categories that are, however, not represented in the core issues since they do not occur as frequently.

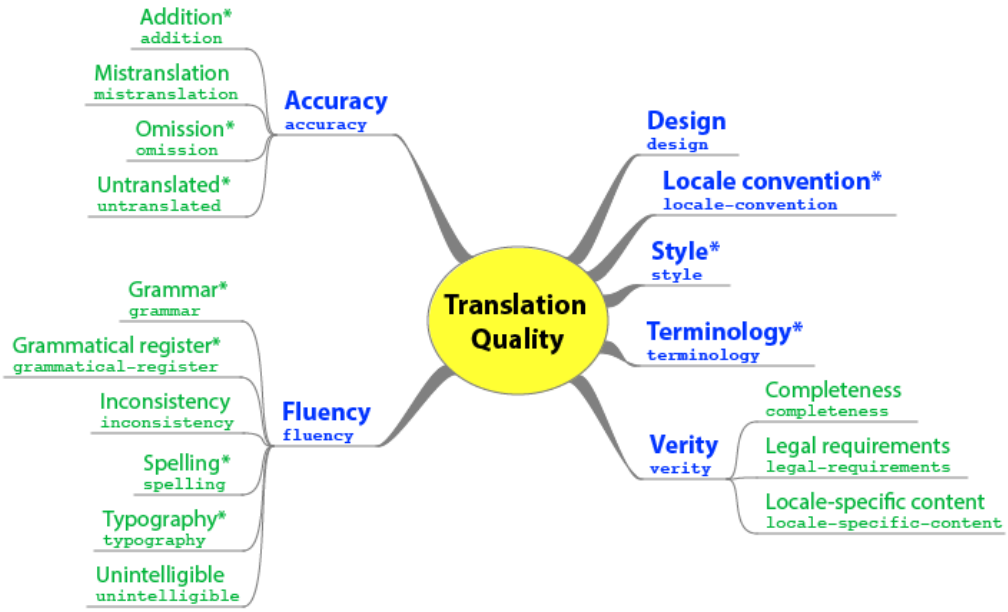

Figure 6: The core issues of MQM (Lommel et al. 2015)

Apart from annotating errors in translated texts, the MQM can also be used to calculate a score for evaluation of the translation product. To calculate a score, it is not enough to know how

many errors are present in the text but also how severe they are and how important the error type is for the task at hand (1998: 1). The severity of errors can be divided into four levels:

1. *None*: Errors that are labelled as none (Lommel et al. 2015) or null (Lommel 2018: 121) are changes that are marked for further attention but that are not wrong per se. This label can be used, for example, if a requester would like to change a term after the translation. In this case, terminology issues can be marked with this label (Lommel 2018: 121). The suggested severity value for this error level is 0 since a translation should not be penalized by these issues (Lommel et al. 2015).
2. *Minor*: Minor errors do not impact the understandability or usability of the translated text. They are often not even noticed by the intended user(s), yet they should be corrected. For example, if *to who it might concern* is written instead of *to whom it might concern*, many readers might not even notice the slight grammatical error and if they do, they will be able to correct it by themselves. The severity value for minor errors is 1 (Lommel et al. 2015; Lommel 2018: 121).
3. *Major*: Major errors are issues that impact the usability and understandability of the content, yet do not cause any harm. They should be fixed before the release of the translated text since they would otherwise result in negative outcomes. If a text contains, for example, misspelt words whose meaning cannot be recovered from their context, this issue should be labelled as major. Major errors are assigned a severity level of 10 (Lommel et al. 2015).
4. *Critical*: Critical issues render a text unfit for its purpose and may even result in safety or legal implications. Translating, for example, *2 pounds* as *2 kilograms* might result in the destruction of equipment or injury. Therefore, critical issues must necessarily be fixed before the release and the evaluation score is penalised by a severity level of 100 (Lommel et al. 2015; Lommel 2018: 120).

Apart from the severity of errors, their importance also plays a role. The importance refers to "the relative value assigned to different categories of errors, rather than to individual instances" (Lommel 2018: 121). For example, style might not be considered too important for a technical manual as long as it is intelligible, yet for marketing brochures it is very important. In order to express the importance of particular error categories in the evaluation score, each category is assigned a weight. The default weight is 1.0, higher numbers indicate greater importance, lower numbers indicate less importance. If a category is assigned a weight of 0.5, for example, all the errors in the category count only 50% of their value (Lommel 2018: 121).

The MQM score can then be calculated using the values of the errors and the weights of the error categories. The result is typically presented as a percentage, where 100% indicates a perfect translation with no errors. Negative scores are also possible if the penalty points exceed the total number of words in the translated text (Lommel 2018: 122).

4 Machine Translation

In today's globalised world, every day a huge number of texts needs to be translated to another language than their original language. Most of these texts do not have a particularly high literary or cultural status but are rather needed for the everyday life of people, like instruction manuals, commercial and business transactions, administrative memoranda, newspaper reports, etc. Many of these texts are highly repetitive and require a high degree of accuracy and consistency (Hutchins & Somers 1992: 2). In order to be able to keep up with the demand for such translations, professional translators often rely on Machine Translation (MT).

The following sections first explain what MT is, how it differs from computer-assisted translation and how it developed historically. Then an overview of the different types of MT systems and how they work is presented. Finally, the challenges and limitations of current MT systems are addressed, and their domains of application are introduced.

4.1 Computers and translation

MT can be defined as "computerised systems responsible for the production of translations from one natural language into another, with or without human assistance" (Hutchins & Somers 1992: 3) or as translation that "is performed by the computer, with no human intervention in the process (although […] there may be need for human intervention before or after it)" (Forcada 2010: 215). These definitions present some important aspects that characterise MT. It is a computer that performs the translation, and no human being is directly involved in the translation process. Humans may, however, play a role before or after the MT process in the form of pre-editing, i.e., the preparation of texts to make them easier to translate for the MT system, or in the form of post-editing, i.e. the correction of errors in the MT output (see also Chapter 7).

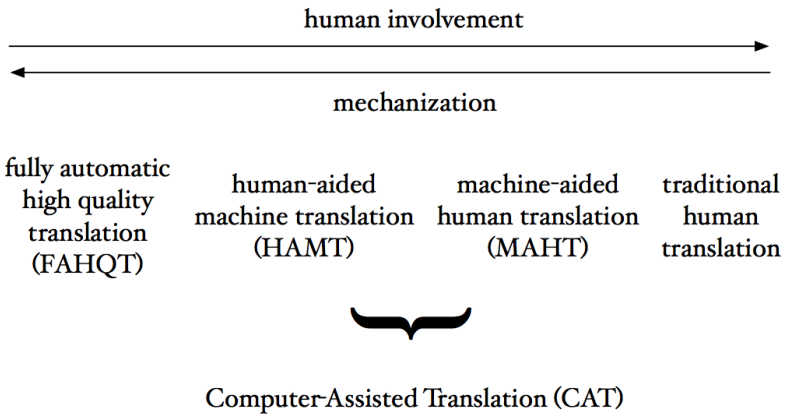

Figure 7: An overview of human and machine translation and their hybrid forms (Hutchins & Somers 1992: 148)

Fig. 7 shows the spectrum of human-computer interaction possibilities in translation as presented by Hutchins and Somers (1992). On the right side of the spectrum is traditional human translation which uses maximum human involvement and no mechanisation at all. With the development of MT new forms of translation emerged that can be divided into Fully Automatic High Quality Translation (FAHQT), Human-Aided Machine Translation (HAMT), and Machine-Aided Human Translation (MAHT), depending on their degree of reliance on human involvement and mechanisation.

In the best case, the raw MT output is of such high quality that it can be published directly. The so-called Fully Automatic High Quality Translation (FAHQT) produces results that are comparable to human translations, which has been the aim of MT researchers from the outset (Hutchins & Somers 1992: 148). However, FAHQT has not yet been achieved and usually computers and humans interact in the translation process. This interaction of computers and humans is called Computer-Aided or Computer-Assisted Translation (CAT) and covers both Machine-Aided Human Translation (MAHT) and Human-Aided Machine Translation (HAMT). In MAHT, a human translator performs the translation by the help of some sort of computational aid. At the most basic level, a word processing programme on a computer can be seen as a form of computational aid. In the last decades, however, more advanced aids have developed, such as spelling checkers, grammar checkers, and style checkers that have been integrated to most word processing programmes, as well as dictionaries, thesauri, and encyclopaedias that are commercially available online (Hutchins & Somers 1992: 149). Moreover, a great many different CAT tools which support the work of translators by proposing them translation suggestions from translation memories and termbases have been developed during the

past years. In HAMT, on the other hand, a computer performs the translation, but a human intervenes before or after the translation process in the form of either pre-editing or post-editing (Hutchins & Somers 1992: 150).

4.2 The history of machine translation

The first idea of using mechanical dictionaries to overcome language barriers traces back to the 17th century when both Descartes and Leibniz came up with the idea of creating dictionaries based on universal numeric codes. Their ideas were inspired by the universal language movement whose proponents were trying to create an unambiguous language based on logical principles and iconic symbols that would allow for every person in the world to communicate with each other without misunderstanding (Hutchins & Somers 1992: 5).

Several other concepts that adopted the idea of a universal language appeared throughout the subsequent centuries, with Esperanto as the best known (Hutchins & Somers 1992: 5). Yet, there were only few attempts to mechanise translation until 1933 when two patents for a kind of multilingual translation dictionary were filed almost simultaneously by Georges Artsrouni in France and Petr Trojanskij in Russia (Kenny 2019: 429). Of the two approaches, Trojanskij's was the more developed one. His idea was that a monolingual human operator would insert the source text into the device in the form of a certain code, where the grammatical functions for the use of every word were added (Kenny 2019: 429f). The machine would then work like a translation dictionary producing target-language words encoded with their respective grammatical information. In a third step, a monolingual target language speaker would decode the words and create the target-language text. This model, as simple as it is, was a forerunner for three-stage MT architectures that were to come in the following decades.

In 1949, Weaver, a mathematician and member of the Rockefeller Foundation, published a memorandum in which he stated that "a multiplicity of language impedes cultural interchange between the peoples of the earth" (Weaver 1949: 1) and suggested trying to solve this problem by the use of computers. The reason, why this memorandum stands out, is because Weaver was mapping out quite accurate goals and ideas of MT before it was even known what computers would soon be able to accomplish. Also, he was the first person to introduce the idea of using a context of *n* words, where *n* represents a user-defined number, instead of translating word by word.

After Weaver's memorandum, MT research programs were launched in many different countries. The first breakthrough was when a functioning MT system that could translate from Russian to English was presented at Georgetown University in 1954 (Kenny 2019: 430). This success provoked the first hype cycle in MT history and led to an emergence of numerous MT research groups in many places, but especially in the US, the United Kingdom, and the Soviet Union. But despite that first phase of innovation and enthusiasm, big successes failed to appear and in the early 1960s disillusionment was beginning to set in. In 1966, eventually, the famous report of the Automatic Language Processing Advisory Committee (ALPAC) (ALPAC 1966) was published, that concluded that MT was slower, less accurate, and more expensive than human translation. As a consequence, funding for MT research was drastically cut almost everywhere (Kenny 2019: 431).

During the years that followed, MT research was continued to be conducted to a reduced extent mainly in Canada and Europe. In the 1980s, hand in hand with the emerging personal computer market, first commercial MT systems were coming up (Kenny 2019: 432). In 1988, eventually, IBM researchers developed the first concept of MT that was not linguistics-oriented but based on probabilistic models of translation. Statistical Machine Translation (SMT) was born. The technological progress as well as the rapid expansion of the internet promoted the growth of SMT in the 1990s and 2000s and it quickly became widely applied all over the world (Kenny 2019: 433). From the mid-2000s on, SMT was considered state of the art for more than a decade until a new approach called Neural Machine Translation (NMT) had its breakthrough in 2015/16 and outperformed SMT (Kenny 2019: 434).

4.3 Types of machine translation

Over the years, two main approaches to machine translation have been developed: rule-based approaches and data-driven approaches. Rule-Based Machine Translation (RBMT), or knowledge-based machine translation, processes linguistic knowledge in the form of lexical, grammatical, and translation rules to create a target language segment. These rules need to be defined and computed by humans. The rule-based approach was state of the art until the 2000s and is still developed and used, especially for languages where less data is available (Kenny 2019: 434). Data-driven MT, on the other hand, is also called corpus-based MT, since its approach is that translation knowledge can be learned from bitexts. Those are parallel corpora in which source texts are aligned with human translations (Kenny 2019: 435) and from which

linguistic patterns can be learned automatically. Data-driven MT is the newer one of the two approaches and has become state of the art during the past two decades due to their comparatively short development time and the mostly higher quality of their results. However, they require enormous amounts of data to learn how to translate and therefore they tend to be very computer-resource intensive. This results in higher costs for creating the system, but also for maintaining it when it is in use (Bennett & Gerber 2003: 179f).

Both, RBMT and data-driven MT, can be subclassified into three different types each. Whereas the types of RBMT differ in the number and kinds of intermediate steps they take between the source-language and the target-language version of a text, the types of data-driven MT use three very different methods of extracting information from bitexts. All six types of MT systems are presented in the following subsections.

4.3.1 Rule-based machine translation

The three types of RBMT that can be distinguished are the direct type, the transfer type, and the interlingual type (Kenny 2019: 434). Each of these three types follows another approach and uses a different architecture that involves different tasks. The pyramid diagram in Fig. 8 used by Arnold (2003) gives an overview of the three architectures of RBMT and the tasks involved in each of them.

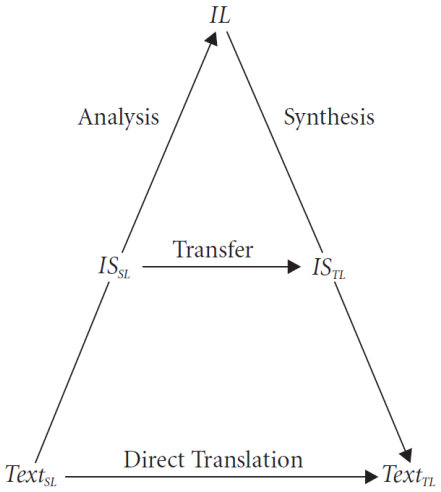

Figure 8: The pyramid diagram of MT architectures (Arnold 2003: 123)

The direct translation type is the simplest one and was mainly used in first generation, or pre-ALPAC, MT systems. In this approach, the source-language text ($Text_{SL}$) is directly converted into a target-language text ($Text_{TL}$) in the style of a dictionary without assigning any linguistic

structure. Therefore, the translation is carried out on a word-by-word basis, which, quite obviously, limits the quality of the translation output (Arnold 2003: 122; Kenny 2019: 434).

Transfer systems are more complex since they use a three-step architecture like the one originally proposed by Trojanskij in 1933, including analysis, transfer, and synthesis (or generation) (Kenny 2019: 435). In the first step, a syntactic and semantic analysis of the source text is undertaken, resulting in an abstract representation, or interface structure (*IS*), of the source-language segment (IS_{SL}). In the transfer step, the source-language representation is converted into a similar representation in the target language (IS_{TL}). In the last step, the synthesis or generation, a target language segment is created from its abstract representation (Arnold 2003: 122). The MT systems that use a transfer architecture are mainly second generation, or post-ALPAC, systems. However, despite their more complex architecture, the results they achieve are only slightly better than the ones achieved by direct MT systems (Kenny 2019: 433).

Interlingual systems operate similar to transfer systems, but they use a language-independent representation scheme, the so-called interlingua (*IL*). The interlingual representation of the source-language content has no traces of the source language, so that, ideally, sentences in any target language can be generated directly, eliminating the transfer step. However, the approach has never really been deployed on a large scale (Arnold 2003: 123; Kenny 2019: 435).

4.3.2 Data-driven machine translation

The data-driven, or corpus-based, approach on MT can also be divided into three subcategories: example-based MT, statistical MT, and neural MT. Each of these types uses a different way of extracting information from bitexts and of using this information in the translation process.

Example-Based MT (EBMT) involves the automatic identification and extraction of segments from bitexts, which are then stored in a database. In the translation process, the system looks for existing segments in the database and combines them to a target-language text (Kenny 2019: 435). The translation process of EBMT systems can be divided into three phases, analogue to the analysis, transfer, and synthesis phases in the transfer approach described in the previous section:

1. *Matching:* The source-language text is split into segments. For each segment, an identical, or at least similar, segment in the source-language side of the aligned bilingual corpus is searched.

2. *Alignment:* The corresponding target-language segments are matched to the source-language segments identified in the first step to create translation units.
3. *Recombination:* The target-language sides of the translation units are combined in order to form target-language sentences (Forcada 2010: 219).

Statistical MT (SMT) systems, on the other hand, are based on a probabilistic model for aligning source and target-language segments and a probabilistic model of the target language. The aligning model assigns a probability to every target-language suggestion for each segment of the source language based on data learned from a parallel corpus. The target-language model, on the other hand, is based on a monolingual corpus and assigns a probability to every segment of the target language. During translation, SMT systems analyse the probabilities of both models and decide on the most probable translation for a given source-language segment based on the joint probability of each word occurring given the previous word (Arnold 2003: 139; Kenny 2019: 435). So if, for example, in a French text the sentence *Quelle heure est-il?* appears, two possible translations that appear in the parallel corpus might be (1) *What time is it?* and (2) *What hour is it?*. However, the probability of Alternative (1) will be higher since it will appear more often aligned to the French sentence and in the monolingual corpus. While the first SMT models were still based on the translation of single words or unigrams, later models could learn translations for n-grams, that is strings of *n* words that "appear contiguously in the training data used" (Kenny 2019: 436). A string of two words is called bigram, for example, and a string of three words is called trigram. However, the limited amount of co-text used to build the probabilistic models as well as the fact that text is split into n-grams that are translated independently of each other and that do not always correspond to any kind of structural linguistic unit result in difficulties of SMT systems when it comes to the handling of discontinuous dependencies or grammatical agreement. Other problems are omissions of source language words in the target language as well as inconsistency (Kenny 2019: 436).

Neural MT (NMT) systems also learn to translate from bitexts but use very different methods to do so. Instead of looking for examples in the database, like EBMT models do, and instead of being based on probabilistic models, like SMT systems, NMT systems use artificial neural networks with artificial neurons, that operate analogously to neurons in the human brain. If one neuron gets activated it stimulates other neurons depending on the strength or weight of the connections between them, finally creating an abstract representation of individual words and the contexts in which they appear. The training of an NMT system, therefore, consists in learning the weights for each connection in order to be able to produce the best output

translation possible even for sequences that are not part of the training corpora. Unlike SMT systems, NMT systems process full sentences rather than n-grams which allows them to handle morphology, lexical selection and word order phenomena like discontinuous dependencies better than SMT (Kenny 2019: 436).

4.4 Applications of machine translation

Even though MT systems have improved a lot during the past few years and the results that can be obtained are very promising, especially for common language pairs where a lot of parallel training data is available, there is still a lot of work to be done in order to reach Fully Automatic High Quality Translations (FAHQT) one day. Yet, the use of MT is already quite common in human-computer interaction contexts, as has been shown in Section 4.1. When using MT, two main purposes to follow can be distinguished: assimilation, or gisting, tasks, and dissemination tasks (Bennett & Gerber 2003: 180f; Forcada 2010: 217).

In assimilation, the MT user does not understand the source-language text and looks for a fast translation in order to comprehend the main ideas of the content, its gist. The MT output is not destined for publication, thus translation errors are not too important for this cause and grammar and style do not need to be perfect as long as the general sense of the text is preserved (Forcada 2010: 217).

In dissemination, on the other hand, MT is just an intermediate step in the production of a text in the target language. Here, the MT user is a professional translator or post-editor, and the final target-language product is designed not only for personal use but for being published (Forcada 2010: 217f). Depending on the situation, the utilization of MT in the translation process may be done in different stages:

1. *Machine translation followed by post-editing*: The source-language text first is translated to the target language by an MT system. In the next step, professional post-editors (ideally specially trained translators who know both the source and the target language) edit the raw MT output into adequate text that can be published (Forcada 2010: 217).
2. *Pre-editing*: In some cases, especially if a text is being translated into more than one target language, it can be helpful to pre-edit the source-language text before the MT step, i.e. to change the source-language text in order to make it easier for the MT system to translate it. This may avoid having to make the same modifications in each

language after the translation. However, pre-editing does not replace post-editing, since not all MT problems can be anticipated (Forcada 2010: 217).

3. *Controlled languages*: In situations where heavy repetitive pre-editing would occur, it may be helpful to define a controlled language in advance that is restricted in its lexicon and syntax. This language then must be used when the source-language text is written. However, designing a controlled language is quite costly and therefore only profitable if heavy pre- and post-editing would occur otherwise (Forcada 2010: 217f).

In order to ensure success in using an MT system, three key factors must be taken into account. First, the required language direction must be available in a commercially robust MT system. This is usually the case for language pairs with high market demand, however, less widely spoken and less demanded language pairs may not be available in any commercially robust system. A second factor is whether the source text is suitable for MT or not, i.e., whether the post-editing effort will be much lower than the effort of translating from scratch or not. While some text types, like technical documentation or business correspondence, are more suitable for MT, others, like literary or marketing texts, cannot yet be handled adequately by MT systems. The post-editing effort would be beyond any acceptable level (Bennett & Gerber 2003: 186). Finally, the third factor is dictionary coverage, i.e., whether the MT system's dictionary covers the relevant terminology (Bennett &Gerber 2003: 187).

5 Machine Translation Quality Assessment

The quality of translation is a topic of debate as has also been shown in Chapter 3 and therefore its assessment is an important aspect in the translation industry as well as in translation studies. The growing demand for translation and the resulting increasing usage of commercial MT that has been going on over the past decades require standards for the assessment of MT performance and output quality in order to be able to successfully integrate MT systems into the translation workflow.

According to Doherty (2017), "[t]he main goal of Translation Quality Assessment (TQA) is to ensure a specified level of quality is reached, maintained, and delivered to the client, buyer, user, reader, etc., of translated texts" (Doherty 2017: 131). Therefore, researchers have been trying to find a way to put subjective judgements about translation quality into a general set of rules and measures that can be used for an objective assessment of translation quality. The result are many different methods and metrics, some of which are presented later in this chapter.

Translation quality can be assessed either by humans, in the case of manual TQA methods, or by computer programmes, in the case of automatic TQA methods. This differentiation seems quite self-defined, however, in practice, the two categories are not always that easy to distinguish. Automatic TQA methods usually use either human reference translations or human annotators, and manual TQA methods make use of computational tools and automated processes. Moreover, there are also semiautomated metrics that are located somewhere between those two categories (Chatzikoumi 2020: 139). In order to present an overview of the different TQA methods, the categorisation presented by Chatzikoumi (2020: 139) is adopted here. According to this categorisation, automatic TQA methods include all techniques "in which human intervention is limited to the development of the evaluation system itself, not taking place during the evaluation process" (Chatzikoumi 2020: 139). The category of manual TQA methods, on the other hand, involves all techniques "in which humans intervene manually during the evaluation phase, regardless of possible interventions in previous phases" (Chatzikoumi 2020: 139). Semiautomated methods are included in the section on human TQA methods since they also do not work fully automatically.

Both methods have their advantages and disadvantages. While manual TQA approaches are generally considered to be subjective, relatively slow, and too cost-intensive, they have the

benefit of being useful when it comes to assessing more complex linguistic phenomena of the translated text and evaluating the type of errors that have occurred. Automatic TQA methods, on the other hand, work faster, are cheaper since they require (almost) no human labour, and are considered more objective. However, the results of automatic TQA are mostly less comprehensive than the results of manual TQA and they often do not indicate the type of problems that occurred (Castilho et al. 2018: 25). Nevertheless, despite the drawbacks that automatic TQA presents, it has gained much popularity in the MT community when it comes to the assessment of MT quality since there is a high demand for frequently performed, low resource iterative assessments, which is not possible for manual TQA methods (Doherty 2017: 134).

In the literature, both terms Machine Translation Quality Assessment and Machine Translation Quality Evaluation are used mostly synonymously (Castilho et al. 2018: 23). In this thesis, however, only the term Machine Translation Quality Assessment (MTQA) will be used in order to avoid confusion, since the abbreviation of Machine Translation Quality Evaluation would be the same as the abbreviation of Machine Translation Quality Estimation (MTQE), one of the core topics of this thesis.

5.1 Manual assessment of MT quality

MTQA performed by humans, or manual MTQA, can be done using a range of different methods. They can generally be divided into methods that use directly expressed judgement (DEJ-based) and such that do not use it (non-DEJ-based) (Chatzikoumi 2020: 139). In DEJ-based methods, humans perform tasks like rating a translation on how good or bad it is or ranking different MT outputs from best to worst. Those methods are usually prone to a higher degree of subjectivity as they rely on the perception of quality of the human judges. Moreover, their judgement is also exposed to indirect comparison, i.e., a segment is more likely to get a good score if the previous segment was also considered good and vice versa. In non-DEJ-based methods, on the other hand, humans are asked to perform more task-oriented processes like the classification and correction of errors, the answering of questions, or gap filling (Chatzikoumi 2020: 139f).

Manual TQA can be performed by either professional or amateur evaluators, depending on the given scenario. While professional translators or linguists can be assumed to create more reliable results, amateur evaluators like students and lay(wo)men are much more often involved in manual TQA tasks due to resource constraints, since professional evaluators are much more expensive. The negative side effect of this is that amateur evaluators often have an undefined,

or self-rated, language proficiency, unknown expertise with the text type and no training concerning the use of the TQA measures employed in the task. In order to make up for this lower level of expertise, manual TQA is usually not conducted by one single evaluator but rather by larger groups of individuals. By averaging the results of these groups, strong personal biases (positive and negative ones) can be moderated, and possible mistakes can be evened out. On the other hand, if professional evaluators are hired, they are usually working alone or in very small groups to save costs (Castilho et al. 2018: 23).

5.1.1 Assessment methods based on directly expressed judgement

In DEJ-based assessment methods, human judges directly express their opinion on how good or bad an MT output is. Usually, the criteria that are used for assessing MT quality are adequacy (or accuracy; both terms are used synonymously in the literature) and fluency. They can be assessed using different methods that all include comparing either the source text with the target text or the target text with a reference translation. The most common method is to rate the adequacy and fluency of a translation on a five-point scale. Other methods include ranking the output of several MT outputs, comparing two MT systems, or expressing a direct judgement on translation quality (Chatzikoumi 2020: 147). All these methods are discussed more in detail below.

5.1.1.1 Assessing adequacy and fluency

The adequacy of a translation can be defined as "the extent to which the translation transfers the meaning of the source-language unit into the target" (Castilho et al. 2018: 18) and can be assessed by looking at both, the source and the target text. Therefore, bilingual evaluators are needed for adequacy assessment. Fluency, on the other hand, is defined as "the extent to which the translation follows the rules and norms of the target-language" (Castilho et al. 2018: 18) or "is a 'good' exemplar of the target language" (Doherty 2017: 133) and can be assessed by monolingual evaluators by looking only at the target text.

Both adequacy and fluency are typically assessed using ordinal scales on which the evaluators need to rate different aspects, like the scale in Fig. 9. This is usually done at sentence level and the extended context is mostly not taken into consideration (Castilho et al. 2018: 18).

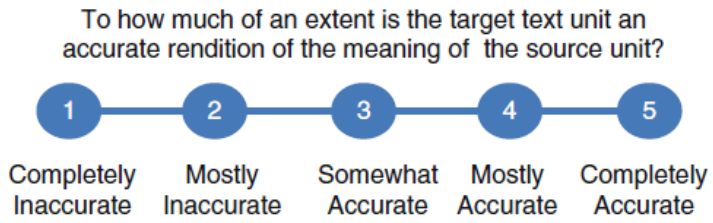

Figure 9: Example of an ordinal scale for measuring adequacy (Castilho et al. 2018: 18)

5.1.1.2 Ranking

Another method that is frequently used for the quality assessment of MT output is ranking. Ranking is especially useful for scenarios where the outputs of different MT systems should be evaluated comparatively in order to find the best option. In such scenarios, evaluators are provided with a source-text unit as well as two or more MTs whose origin rests unknown and which are listed in an unpredictable, random order. The evaluators then must either pick one translation that they consider best or rank all the options from best to worst (Castilho et al. 2018: 21).

The main advantage of ranking methods is their conceptual simplicity, i.e. they are easy to understand for annotators. On the other hand, the inter-annotator agreement is relatively low for ranking methods and they require a lot of annotation time. Moreover, usually only up to five MT outputs are compared at the same time, since ranking more options can impair the judgements of the human annotators. Therefore, ranking methods are not useful for the evaluation of many MT systems at a time (Chatzikoumi 2020: 148).

5.1.1.3 Direct Assessment

Direct Assessment (DA) methods consist of judging the quality of an MT output on a continuous rating scale rather than on a five-point scale. This way, the degree to which one translation is better than the other can be assessed. DA methods are mainly focused on adequacy, even though they can assess both adequacy and fluency, and they allow for more fine-grained analysis than five-point scales do (Chatzikoumi 2020: 148f). The assessors used for DA methods are usually not experts. In order to assure reliability, quality control items, such as bad reference translations or repeated MT outputs, appear in between the segments of the text that is being assessed. This way, the intra-annotator agreement, i.e., whether the assessor works consistently or not, can be yielded (Chatzikoumi 2020: 149).

5.1.2 Assessment methods not based on directly expressed judgement

In manual assessment methods that are not based on directly expressed judgement (non-DEJ-based evaluation methods), human judges also assess MT output but express their judgement only indirectly. The methods that are used can be either semiautomatic metrics, tasks where the comprehension of MT output is required, or the classification and analysis of errors (Doherty 2017: 134). All these methods are presented below.

5.1.2.1 Semiautomatic metrics

Semiautomatic metrics are variations of automatic metrics where the computer does not perform the evaluation alone, but a human annotator intervenes at some point (Chatzikoumi 2020: 140). Some examples of these so-called human-in-the-loop assessment methods are the HBLEU and HTER metrics (Snover et al. 2006), the human-targeted variations of the BLEU and TER metrics that will be presented in Section 5.2.1.

5.1.2.2 Task-based evaluation

The quality of MT output can also be assessed by letting humans perform a particular task and evaluating how good this works. Some examples for such tasks are asking people to summarize the most important ideas of a text, asking them to answer questions on the content, and filling gaps in the text (Chatzikoumi 2020: 149). These methods show indirectly to what degree the content of the source text has been transferred to the target language and if the target text is intelligible or not, i.e., they evaluate the adequacy and fluency of the translation without asking for a direct judgement. However, task-based evaluation methods are mainly used for translations for gisting purposes since they only evaluate if the basic concepts of the source text have been translated correctly without going more into detail (Chatzikoumi 2020: 149f).

5.1.2.3 Error classification and analysis

While most TQA methods that are presented in this section give either an overall score on the quality of a translated text or demonstrate better-or-worse ranking decisions, error classification and analysis can help in giving additional, more detailed information on the types of errors and what influence those error types have on the performance of an MT system. Whereas an overall score or a comparison of several MT systems represent valuable information for MT developers, it may be interesting to know more about the particular strengths and weaknesses their systems have in order to be able to improve it (Popović 2018: 130). Error classification and analysis can, for example, give answers to questions like: What are the most serious issues in

an MT system? Are there any aspects where a worse-ranked system is able to outperform a better-ranked one? Are some error types easier to post-edit than others (Popović 2018: 130)? In order to answer such questions, error classification and analysis techniques have emerged that can be performed either manually, automatically, or semiautomatically. Since this is a very extensive topic that is at the same time very relevant for this thesis, it will be discussed in a separate chapter (Chapter 6).

5.1.2.4 Post-editing

Another way of assessing an MT system is to post-edit its output since post-editing (PE) involves the correction of errors and the improvement of parts of the translation that are of low quality and issues in the MT output are thereby made visible. Since PE is an extensive topic just like error classification, it will be discussed in a separate chapter too (Chapter 7).

5.1.3 Advantages and disadvantages of manual TQA

The manual assessment of translation quality has some clear advantages but also some well-known disadvantages. On the one hand, translations are usually generated for human end users and should therefore stand up to human judgement. For this reason, using human judges in the assessment process before passing the translated text to its end users stands to reason. Moreover, not every translation error really is one and needs to be corrected. Yet, only human judges can decide if an error is important and needs correction or not based on their comprehension of the real world (Turian et al. 2006: 5). On the other hand, the main disadvantage of manual TQA methods is their subjectiveness. Other drawbacks include the high cost of human resources, the lack of repeatability, the amount of time needed, the low inter-annotator agreement, and the fact that the beforementioned evaluation tasks are generally performed at sentence level, which makes it hard to evaluate inter-sentence cohesion (Chatzikoumi 2020: 156).

Some of these disadvantages, like the high costs, the lack of repeatability, and the amount of time needed, can be overcome by the use of automatic assessment methods that are presented in the next section. However, it is important to state that subjectiveness still cannot be completely eliminated using automatic methods, since they are all computed by humans and most of them rely on the comparison of translations with human references.

5.2 Automatic assessment of MT quality

Since there is a growing awareness in the MT community of the limitations of human assessment methods, automatic assessment methods are becoming more and more popular. Automatic

assessment is mainly based on automatic evaluation metrics (AEMs), which can be defined as "scripts or software programs that perform TQA using an explicit and formalised set of criteria to evaluate the MT output, typically against a human reference translation" (Doherty 2017: 134) or "metrics that yield a score for the MT output based on the degree of similarity to reference translations" (Doherty 2017: 134). This means that a human reference translation is needed to compare the MT output with, producing a score based on the similarity between the two texts. This score is usually either a percentage of matched n-grams, i.e., a sequence of n words in a text, between the MT output and its reference, or the edit distance between them (Chatzikoumi 2020:141). Two of the most popular AEMs are the BiLingual Evaluation Understudy (BLEU) metric (Papineni et al. 2002) that is based on n-gram precision, and the Translation Edit Rate (TER) metric (Snover et al. 2006) that is based on edit distance.

Whereas reference translation-based metrics are by far the most popular and most widely used ones, there are also other approaches to the automatic assessment of MT quality. In 2008, Zhou et al. presented their model of diagnostic evaluation based on linguistically motivated features they call checkpoints which goes without the need for reference translations. Another approach are confidence or quality estimation (QE) metrics which try to predict the quality of the output of an MT system without having any information about the expected output. They are not evaluation metrics per se but can be seen as proxies for them (Chatzikoumi 2020: 140) and will therefore also be briefly introduced in this section. More detailed information about how they work will be given in Chapter 8 and different QE systems will be introduced at a later point of this thesis.

When it comes to the evaluation of automatic TQA methods, the most important criterion is that they judge the MT output the same way a human evaluator would (Turian et al. 2006:5) in order to guarantee acceptability. Other criteria are consistency and correctness of the results, reliability of the method, speed and usability, low cost, delivery of scores that can be interpreted intuitively, sensitivity to nuances of quality among the compared MT systems, the possibility of direct system performance optimisation, and usability for a great range of fields (Chatzikoumi 2020: 140). A satisfactory TQA method should meet all these criteria, yet the current methods still have some flaws as will be shown in the following sections.

5.2.1 Reference translation-based metrics

Reference translation-based TQA metrics rate MT output based on its similarity to one or more human reference translations that are considered gold standard. Since automatic methods are

not (yet) capable of evaluating the beforementioned concepts of accuracy and fluency like humans can (Doherty 2017: 134), they use different approaches. Some metrics are based on the principles of precision and recall. Precision is defined as "the fraction of words in the MT output that are correct (as indicated by the gold-standard reference translation)" (Doherty 2017: 134), i.e., the ratio of the n-grams of the MT output that are found in at least one of the human reference translations to the number of n-grams in the MT output (Chatzikoumi 2020: 141). Recall, on the other hand, is understood as "the fraction of correct words (also indicated by the gold-standard reference translation) that are produced in the MT output" (Doherty 2017: 134), i.e., the ratio of the n-grams of the MT output found in at least one of the human reference translations to the number of n-grams in the reference translation that is the ideal number of n-grams (Chatzikoumi 2020: 141) (. Other metrics, on the other hand, are based on the edit distance, i.e., the minimum number of operations needed to make the MT output match the reference translation(s) (Chatzikoumi 2020: 140). For each type of reference translation-based metric one example will be presented here: the BiLingual Evaluation Understudy (BLEU) metric developed by Papineni et al. (2002) as an example of an AEM based on precision and recall, and the Translation Edit Rate (TER) metric developed by Snover et al. (2006) as an example of an AEM based on edit distance.

5.2.1.1 BiLingual Evaluation Understudy (BLEU)

One of the first automatic evaluation metrics that attracted great interest was the BiLingual Evaluation Understudy (BLEU) metric, presented by Papineni et al. (2002). According to its developers, the BLEU metric is "a method of automatic machine translation that is quick, inexpensive, and language-independent, that correlates highly with human evaluation, and that has little marginal cost per run" (Papineni et al. 2002: 311). It is proposed as a substitution to human evaluators in scenarios when quick or frequent evaluations are needed (Papineni et al. 2002: 311).

The central idea behind the BLEU metric is that the closer the MT output is to a professional human translation, the better it is. Therefore, the system aims for measuring the closeness of an MT to one or more human reference translations according to a numerical metric, the BLEU metric (Papineni et al. 2002: 311). There are three important aspects that influence the closeness of one text segment to another: word choice, word order, and segment length. The BLEU metric incorporates all of them, while, at the same time, trying to allow some legitimate differences (Papineni et al. 2002: 311).

The BLEU metric measures closeness in terms of word choice and word order by using *modified n-gram precision.* This means that the maximum number of times an n-gram occurs in any reference translation is first counted. Next, the total count of this n-gram is clipped by its maximum reference count and is then divided by the total number of candidate n-grams (Papineni et al. 2002: 312). This procedure ensures that a reference n-gram is considered exhausted after a matching candidate n-gram has been found for it in order to avoid giving high precision scores to bad translations like the one in the following example taken from Papineni et al. (2002: 312):

Candidate: *the the the the the the the*
Reference 1: *The cat is on the mat.*
Reference 2: *There is a cat on the mat.*

The modified unigram precision here is 2/7 since the word *the* is found twice in Reference 1 and the total number of candidate words is 7. Without adding the restriction of considering a reference n-gram exhausted after it has been matched, the unigram precision score would be 7/7 because every candidate word can be found in the reference translations. This, of course, would not be a useful score since the candidate translation does not make much sense.

Whereas the n-gram precision penalises words that appear in the candidate translation, but do not appear in the reference translation, as well as words that appear more frequently in the candidate translation than they do in the reference translation, it does not know how to cope with recall problems and fails to identify the proper translation length (Papineni et al. 2002: 314). The candidate translation in the following example, taken from Papineni et al. (2002: 314) shows an excellent unigram and bigram precision, but is not a good translation, of course:

Candidate: *of the*
Reference 1: *It is a guide to action that ensures that the military will forever heed Party commands.*
Reference 2: *It is the guiding principle which guarantees the military forces always being under the command of the Party.*

In order to avoid such a short answer getting a good score, a *brevity penalty factor* is introduced to the metric and is multiplied by the n-gram precision. A candidate translation must now have the same length as the reference translation and match it in word order and word choice in order

to obtain a high score (Papineni et al. 2002: 315). The scores of the BLEU metric range from 0 to 1, 1 corresponding to candidate translations that are identical to a reference translation. This means that (1) even a human translator will probably not obtain a score of 1, and (2) the more reference translations per sentence there are, the higher the score will be (Papineni et al. 2002: 315).

5.2.1.2 Translation Edit Rate (TER) and Human-targeted Translation Edit Rate (HTER)

Snover et al. (2006) introduced a new, more intuitive automatic measure for evaluating MT output that is based on "the amount of editing a human would have to perform to change a system output so that it exactly matches a reference translation" (Snover et al. 2006: 223): the Translation Edit Rate (TER). TER is defined as "the minimum number of edits needed to change a hypothesis so that it exactly matches one of the references, normalized by the average length of the references" (Snover et al. 2006: 225). This means that only the number of edits needed to match the closest reference are counted since the goal is to obtain the minimum number. The possible edits are the insertion, deletion, and substitution of single words as well as shifts of word sequences. Moreover, punctuation marks are treated as words and fixing errors in capitalisation is counted as an edit. All edits have equal weight, they all have a cost of one (Snover et al. 2006: 225). In the following example taken from Snover et al. (2006: 225), a candidate translation (or hypothesis, as Snover et al. call it) is compared to one reference translation:

Reference: *SAUDI ARABIA denied THIS WEEK information published in the AMERICAN new york times*

Candidate: *THIS WEEK THE SAUDIS denied information published in the new york times*

The candidate translation is fluent and means the same thing as the reference translation, yet they do not completely coincide, and four edits have to be made: one shift (of the word sequence *this week*), one insertion (of the word *American*) and two substitutions (*the* and *Saudis* need to be changed to *Saudi* and *Arabia*). By dividing the number of edits by the total amount of reference words, we obtain a TER score of 4/13 or .31. Since TER is an edit-distance metric, the lower its score is the better the candidate translation is. A TER score of 0 would mean that the candidate translation is identical to its reference and that no edits are needed.

The weakness of the TER score is, however, that it does not reflect the acceptability of a candidate translation (Snover et al. 2006: 225). The candidate shown in the example above semantically matches the reference translation, even though it uses slightly different words to express the same content. However, human knowledge is needed to detect such semantic equivalence. Therefore, Snover et al. refined the TER metric by employing human annotators in a human-in-the-loop process in order to make TER a more accurate method for evaluating translation quality. They called this new method Human-targeted Translation Edit Rate, or HTER (Snover et al. 2006: 226). For the HTER score, human annotators, who are fluent in the target language, are used to create a targeted reference translation. They receive a candidate translation and its reference and edit the candidate until it expresses the same meaning as the reference but is not necessarily identical with it. This new targeted reference is then used as a new human reference in another run of obtaining the minimum TER (Snover et al. 2006: 226).

5.2.1.3 Advantages and disadvantages of reference translation-based AEMs

The most important advantages of reference translation-based AEMs are their speed and their low cost when in usage since they require fewer human resources than manual assessment methods do. Also, their reusability in the development of MT systems is an important factor, since it enables modifications and improvements to be made in the systems as well as re-evaluation (Chatzikoumi 2020: 142ff).

However, reference translation-based AEMs are far from being perfect and still have many drawbacks. The most obvious one is the need for reference translations created by humans, which impedes the full automation of the process, time and cost minimisation, as well as full subjectivity. Other currently existing drawbacks, that could however be overcome in the future, include the metrics' inability to distinguish between nuances, the fact that they work well at document level but are not reliable at sentence level, the difficulty in interpreting the scores, as well as their inability to identify the exact strengths and weaknesses of an MT system (Chatzikoumi 2020: 144).

5.2.2 Diagnostic evaluation based on checkpoints

Zhou et al. (2008) came up with another method that provides a diagnostic evaluation based on linguistic checkpoints. They claim that their method is able to overcome the drawbacks of reference translation-based evaluation methods, such as their inability to compare MT systems that use different architectures or to inform MT developers on the exact strengths and

weaknesses of their systems, since the capability of a system in handling different linguistic categories is evaluated rather than just a general score being assigned (Zhou et al. 2008: 1121).

The first step of diagnostic evaluation consists of creating a checkpoint taxonomy like the one in Fig. 10 that consists of the relevant checkpoints for evaluating a translation of a particular domain and a particular language pair. A checkpoint is a linguistic unit at either word level (e.g. ambiguous words, prepositions, etc.), phrase level (e.g. collocations, subjective-predicate phrases, etc.), or sentence level (e.g. adverbial clauses, noun clauses, etc.) (Zhou et al. 2008: 1122).

Word level		
Noun	Verb (with Tense)	Modal verb
Adjective	Adverb	Pronoun
Preposition	Ambiguous word	Plurality
Possessive	Comparative & Superlative degree	
Phrase level		
Noun phrase	Verb phrase	Adjective phrase
Adverb phrase	Preposition phrase	
Sentence level		
Attribute clause	Adverbial clause	Noun clause
Hyperbaton		

Figure 10: Example of a checkpoint taxonomy (Zhou et al. 2008: 1123)

When the taxonomy is set, a large number of parallel sentences is collected from different corpora. In the next step, the sentences of both source and target language are parsed and aligned. After the alignment, the parsed sentence pairs are searched, and the checkpoints are extracted. Finally, the references for each checkpoint have to be determined in the source languages. The whole creation of the checkpoint database is a fully automatic process. With the checkpoint database at hand, the diagnostic evaluation can then be performed. To do so, the number of matching n-grams of the references against the MT output sentence is calculated and a numerical value is assigned to each category. The score of a category group or a whole system can be obtained by adding the scores of the different categories they contain (Zhou et al. 2008: 1122).

By comparing the results of diagnostic evaluation based on checkpoints with the results of state-of-the-art metrics such as BLEU, Zhou et al. showed that their method performs better when it comes to providing information of the capability of an MT system in translating important linguistic features. This can help developers to know which part of their MT system needs to be improved.

6 Error classification and analysis

As mentioned in the previous chapter, the manual and automatic assessment methods presented before providing valuable information about the overall performance of MT systems and can help in their improvement, yet there are some more in-depth questions that researchers need answers for. It might, for example, be of great interest for them to know what the most serious problems of a certain MT system are, what the particular strengths and weaknesses of a system are, or if some error types are more difficult to post-edit than others. In order to be able to answer those questions, error classification and analysis techniques have been developed. By means of those techniques, errors can be identified and classified in a translated text and more advanced decisions can be taken based on the results (Popović 2018: 130). The distribution of a defined error class in a translated text at hand can be analysed and different translation outputs can be compared for each error class. Moreover, more specific aspects, such as relations between certain error types and the preferences of users and post-editors, or the impact of different error types on different aspects of the post-editing process can be analysed (Popović 2018: 131).

Just like TQA, the classification and analysis of translation errors can be carried out either manually or automatically, or using combined, semi-automatic methods (Popović 2018: 131). First, a suitable set of error classes, i.e. an error typology or taxonomy, needs to be defined. This is a challenging task in itself since the relevant error categories and their level of detail vary with every task, depending on the domain of the translated text but also on the language pair. Once the error typology is set, the errors can be annotated and analysed. In manual error classification and analysis settings, the annotator looks into the MT output, marks each erroneous word and puts an error tag to it. Apart from the MT output, the annotator needs to get at least one correct text for reference, either the source text or a reference translation (Popović 2018: 131). If automatic error classification is used, a computer programme compares the MT output with a reference translation in order to find errors. However, in contrast to AEMs that are used for TQA purposes, automatic error classification aims for indicating the number of different error types rather than for producing a single overall score (Popović 2018: 142).

6.1 Manual error classification

Manual error classification and annotation are performed by human annotators who look into the MT output and mark and classify the errors that they find. For reference, they get either the source text in the original language or a reference translation, depending on whether they are monolinguals or bilinguals.

The general procedure of manual error annotation, from the source-language text until the error analysis, is illustrated in Fig. 11. The rectangles denote automatic processes, like MT, and the ellipses denote manual processes (Popović 2018: 132). First, the source-language text is translated by an MT system and the output is either directly used for error annotation, or post-editing is performed as an intermediate step and the errors in the post-edited MT output are annotated by assigning error tags to the performed post-editing operations. Depending on whether the annotator is monolingual or not, a human translation might need to be created for reference. Once the annotator has all required texts and the error taxonomy for the given task, the error annotation can be carried out. The subsequent analysis of the annotated errors can be performed either by a human or by a computer.

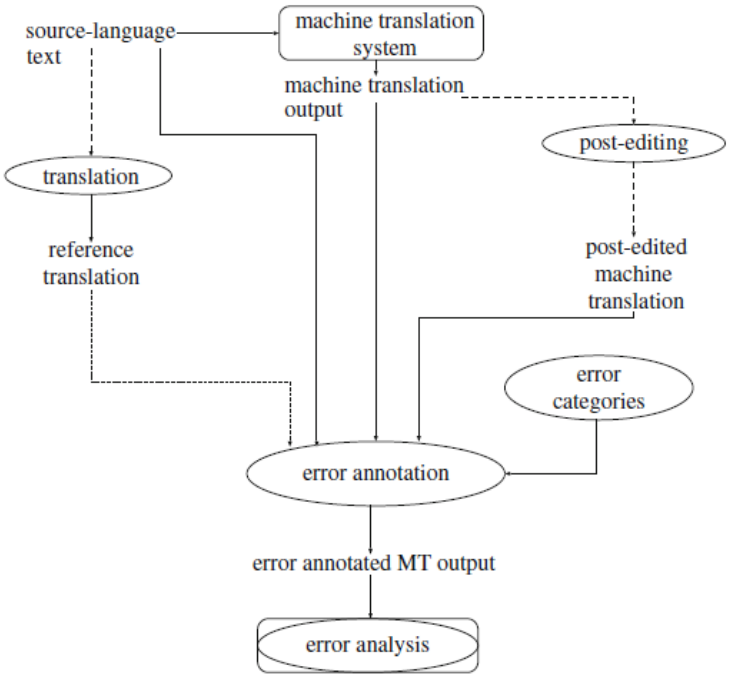

Figure 11: General procedure of manual error annotation (Popović 2018: 132)

6.1.1 Challenges of manual error classification

In order to achieve the best possible result through error annotation and analysis, the error categories should be clearly defined beforehand in the form of an error typology or taxonomy. These categories should reflect all strengths and weaknesses of the MT system that are relevant for the task at hand and the language pair(s) involved, and both linguistic, as well as translation aspects, should be covered. This is no simple task and even though there is work in progress, no general rules for defining error categories has been presented yet (Popović 2018: 140). During the past few years, a lot of different possible error typologies have been suggested, ranging

from very basic ones with just a handful of categories to more advanced ones that contain different levels of error categories, from more general errors to very specific ones. Some of the challenges they have all been dealing with are (in)consistency issues, the number of error classes and the definition of error classes (Popović 2018: 139).

Precise guidelines and intensive training of the annotators, regardless of whether they are bilinguals or monolinguals, are necessary in order to minimise inconsistencies in the error annotation. However, it is impossible to completely avoid all inconsistencies, mainly because of two reasons. The first one is the differing perception of the exact position of erroneous words in a text, especially when it comes to word-order or phrase-order errors. The second reason is the different perception of certain error classes in different contexts. The more detailed the error taxonomy is, the lower the consistency gets in terms of the perception of error classes (Popović 2018: 139).

However, more detailed error typologies usually also provide better information about the errors, which is also important. For the annotators, it might, for example, be much easier to work consistently, if they can just mark any morphological error with the label *morphological error*. This, on the other hand, does not provide detailed information about whether the error was an inflectional, derivational, or compositional one. But if we add a second level of error categories, where *morphological error* is separated into *inflectional error*, *derivational error*, and *compositional error*, the consistency of one annotator as well as the results of different annotators, the so-called inter-annotator agreement, will get lower (Popović 2018: 139). It is therefore essential to find an exact definition for each error class to minimise those issues.

The definition of error classes, however, depends on the task as well as on the language pair(s) again (Popović 2018: 140). If the translated text is from a specific domain, *terminology error* is a very important class, whereas it is not for a text from a more general domain. Similarly, the error class *verb inflection error* is very important for languages like German or the Romance languages, but it is less important if English is the target language. Besides, without a clear definition and intensive training, it might in some cases be difficult to distinguish between an error class and its subclass, as it is, for example, the case for *mistranslation* and its subclass *terminology error* (Popović 2018: 140).

6.1.2 Possibilities for facilitation of manual error classification

In order to overcome those challenges and to facilitate the process of error classification, different approaches have been found by researchers that aim at turning the time and resource-

intensive task of manual error classification into a rather simple and reliable process with reduced effort and inconsistencies. Popović (2018: 140) groups those approaches into three main lines of work: (1) the "unification and generalisation of error typologies" (Popović 2018: 140) to be used as a starting point, (2) the "annotation of post-edited operations instead of using raw translation output" (Popović 2018: 140), and (3) "the automation of (a part of) the error classification process" (Popović 2018: 140).

Establishing a general error typology that unifies the most important error categories from different approaches can be useful for reducing the effort of finding error categories and minimising inconsistencies related to the definition of error classes. These general typologies usually offer a very detailed error set with several different subsets. The idea is to use them as a starting point and select the relevant categories for the task at hand (Popović 2018: 140).

Fig. 12 shows such a general error typology as suggested by Popović (2018: 141). It contains error classes on three levels, ranging from more general classes to more specific ones. The error classes on Level 1 are *lexis*, *morphology*, *syntax*, *semantic*, *orthography*, and *too many errors*. All of the classes apart from *too many errors* can then be further divided into several subclasses, for example, *syntax* can be divided into *capitalisation*, *punctuation*, and *spelling*. The category *too many errors* can be used for very low-quality segments, where it is hard to classify the errors, yet, it should not be overused (Popović 2018: 141).

Level 1	Level 2	Level 3
Lexis	Mistranslation	Terminology
	Addition	
	Omission	
	Untranslated	
	Should not be translated	
Morphology	Inflection	Tense, number, person
		Case, number, gender
	Derivation	Part of speech
		Verb aspect
	Composition	
Syntax	Word order	Range
	Phrase order	Range
Semantic	Multi-word expressions	
	Collocations	
	Disambiguation	
Orthography	Capitalisation	
	Punctuation	
	Spelling	
Too many errors		

Figure 12: A possible general error typology (Popović 2018: 142)

Aside from general error typologies, the annotation of post-editing operations can also be deployed in order to facilitate manual error classification. Usually, post-editing and error classification are considered and carried out as two separate tasks. Fig. 11 has shown that in general, either raw or post-edited MT output is given to the annotators for error classification. However, post-editing can also be seen as a form of implicit error annotation, since every correction that is made in the post-editing process shows where errors are located in the MT output. Merging post-editing with error annotation by annotating every change that is made during the post-editing process could therefore help to facilitate the annotation process and reduce error position inconsistencies (Popović 2018: 141).

Another measure that can be taken in order to reduce the effort of error classification is to automate (parts of) the process. The advantages and disadvantages of automatic error classification in comparison to manual error classification are similar to the ones of automatic TQA. Whereas automatic error classification is faster, cheaper, and more consistent, it is also less precise and strongly dependent on a given (human) reference translation. In addition, the risk of error tags being assigned incorrectly is much higher than for manual error classification (Popović 2018: 141). A detailed overview of the automatic error classification process and its methods is given in the next section.

6.2 Automatic error classification

As mentioned in the previous section, manual error classification is very time and resource-intensive and prone to inconsistencies. Therefore, the idea arose to use a computer programme to compare the translation output with a reference and label the errors (Popović 2018: 142). The procedure is roughly the same as the one for manual error classification shown in Fig. 11 with the only difference that error annotation and analysis are fully automatic processes (Popović 2018: 146). Yet, the error categories still have to be chosen manually, thus the same challenges as in the creation of error typologies for manual error classification have to be faced.

7 Post-Editing

As mentioned in the previous chapters, the growing need for translated texts has entailed an increased use of MT to save time and costs. Professional translators are no longer just asked to translate texts from scratch, but to post-edit translations made by MT systems. Post-editing (PE) is "the correction of raw machine-translated output by a human translator according to specific guidelines and quality criteria" (O'Brien 2011: 197f). This chapter first gives an overview of how PE developed as a separate field of research. Then the different types and levels of PE are presented along with some guidelines.

7.1 Development of post-editing

Since PE is by definition connected with MT, the development of MT systems and PE have always been going hand-in-hand. Many of the first MT systems that appeared around the middle of the 20th century made use of pre-editors, that were experts in the source language, and post-editors, that were experts in the domain and the target language. However, when MT financing experienced a cut after the publication of the ALPAC report in 1966, PE activities and empirical, academic research on PE were also reduced as the report stated that PE was not cost-effective (Nitzke 2019: 14). In the 1980s, PE research was initiated by non-academics and many companies and institutions, like the Commission of the European Communities, began using MT and PE in their translation workflows. The development of PCs gave an extra boost to the use of PE since functionalities like search and replace functions helped to accelerate the PE process and thus made it more profitable. Around the same time, PE also started to become a recognised field in translation studies and was studied academically again (Nitzke 2019: 15). Finally, in the past few years, Computer Assisted Translation (CAT) tools were developed that combine Translation Memories (TM) with MT (Nitzke 2019: 15). Translators that are using CAT tools can either translate segments from scratch or choose to use MT and post-edit the MT output by means of TM. Thus, PE is a part of the work of many translators nowadays, either because they are assigned to perform PE tasks for a client or because they use MT systems in the translation process.

7.2 Types and levels of post-editing

PE can be performed on many levels, ranging from the use of raw MT output with no PE at all to full PE of MT output that can, in the worst case, require more effort than translation from scratch. The selection of the adequate level of PE depends on several reasons: the intended

user(s), the amount of text to be post-edited, the required quality level of the final product, the translation turn-around time, the expected life expectancy and perishability of the information in the translated text, and the context in which the final product will be used (Allen 2003: 301). Most of these factors are taken into account if a look is taken at the objective or purpose of the translated text. Depending on whether the translation is used for assimilation purposes or dissemination purposes (see also Section 4.4), PE can be divided into different types and levels that are listed below.

7.2.1 Post-editing of MT for assimilation

As has already been stated in Section 4.4, the most important factor of MT for assimilation purposes is speed. The users need a quick translation of content in a language they do not understand and freedom from errors is not very important as long as the content of the MT output is intelligible (Allen 2003: 302). Therefore, PE is not very important in assimilation contexts and raw MT output is often used for gisting purposes with no PE being performed at all. Yet, there are some contexts where it is essential that a group of people quickly understands the content of a translated text, however, the information has a very short life span. This might be the case for documents like working papers for internal meetings, minutes of meetings, technical reports, etc. Such documents are usually not intended for public use or wide circulation thus it would be a waste of resources to perform PE that goes beyond the absolutely necessary changes. In such situations, rapid PE can be performed where only a minimal amount of corrections is undertaken in order to fix the most obvious and significant errors. Possible stylistic issues should not be considered in rapid PE. That way, a fully comprehensible element of information can be produced by using only the least possible amount of work (Allen 2003: 302).

7.2.2 Post-editing of MT for dissemination

MT for dissemination purposes is MT output that is intended to be published. Therefore, the raw MT output needs to be post-edited to render it appropriate for being read by many people (Allen 2003: 303). In the best case, the raw MT output would be good enough to be published without any corrections. Yet, almost no domain has so far been able to produce MT results that are consistently good enough to be published without PE thus usually at least a minimum amount of PE needs to be performed (Allen 2003: 304). Depending on the expected quality of the results, either light PE or full PE can be performed.

Light, or minimal, PE means that only “the least amount of comments possible for producing an understandable working document, rather than producing a high-quality document”

(Allen 2003: 305) is made. The translated document should be corrected to the extent that it is adequate for dissemination and the meaning of the content is conveyed adequately in the target language, yet it can still contain some minor errors especially in grammar and style (Allen 2003: 305). However, the decision on which corrections are absolutely necessary often depends on the subjective perception of the post-editors. This often leads to either over-correcting, where the post-editor spends too much time and cognitive effort on the PE process, or under-correcting, where the post-editor does not sufficiently correct the MT output and significant errors still appear in the result (Allen 2003: 305).

Full PE, on the other hand, produces text of very high quality, similar or equal to a good human translation (Allen 2003: 306). Depending on the quality of the raw MT output, full PE thus requires a high level of effort and the question arises whether or not it might be faster to let a human translate the text from scratch.

7.3 MT Post-editing guidelines

As has already been mentioned, the judgement of where light PE stops and where full PE begins is often a rather subjective one. In order to avoid over-correcting and under-correcting, clear guidelines are necessary for every PE project for post-editors to know exactly what their task is. Also, guidelines help customers and language service providers to know what they can expect from the post-edited product. Despite this, there are no widely accepted general PE guidelines and most organisations and companies that use PE in their translation workflows have their own internal guidelines and thus not many publications are available on this topic (TAUS 2010). Yet, in order to give an example, the PE guidelines as defined by the Translation Automation User Society (TAUS) are presented here.

In 2010, the Translation Automation User Society (TAUS) published two sets of guidelines for PE along with some recommendations about how to reduce the amount of PE required. These pre-editing recommendations include, amongst others, appropriate tuning of the MT system, ensuring high quality of the source text, integration of terminology management, and training of post-editors in advance (TAUS 2010). Once these recommendations are followed and the best possible MT output is generated, a set of PE guidelines has to be chosen depending on the desired level of quality of the end product.

The first set of guidelines is for creating a text of "good enough" (TAUS 2010) quality, which corresponds to light PE. The aim is to create a text that is comprehensible and accurate,

but that may sound like it was generated by a computer due to its unusual syntax, grammar, and/or style (TAUS 2010). The suggested guidelines are:

- Aim for semantically correct translation.
- Ensure that no information has been accidentally added or omitted.
- Edit any offensive, inappropriate or culturally unacceptable content.
- Use as much of the raw MT output as possible.
- Basic rules regarding spelling apply.
- No need to implement corrections that are of a stylistic nature only.
- No need to restructure sentences solely to improve the natural flow of the text (TAUS 2010).

The second set of guidelines that TAUS presents is for PE where the end product should be of "quality similar or equal to human translation" (TAUS 2010) which corresponds to full PE. These texts should also be comprehensible and accurate, but in difference to texts of good enough quality syntax and grammar should also be on par with human quality and the punctuation should be correct. The style should also be accurate, even though it may not be as good as that of a human translation (TAUS 2010). The suggested guidelines for full PE are:

- Aim for grammatically, syntactically and semantically correct translation.
- Ensure that key terminology is correctly translated and that untranslated terms belong to the client's list of "Do Not Translate" terms.
- Ensure that no information has been accidentally added or omitted.
- Edit any offensive, inappropriate or culturally unacceptable content.
- Use as much of the raw MT output as possible.
- Basic rules regarding spelling, punctuation and hyphenation apply.
- Ensure that formatting is correct (TAUS 2010).

8 Machine translation quality estimation

As has been stated in the previous chapters, the quality assessment of MT is essential for the continuous improvement and development of MT systems. In order to overcome the drawbacks of manual evaluation methods, like time and cost-intensiveness as well as subjectivity, automatic TQA methods have been developed. However, most automatic evaluation methods that are deployed on a large scale share two major limitations: (1) their score is based on a comparison of the MT output against a human reference translation which impedes full automation of the process, minimisation of cost and time resources, as well as complete objectiveness, and (2) their scores reach good correlation with human judgement, and are therefore reliable, only at the corpus level, whereas the results at sentence level are not very reliable (Specia et al. 2010: 39f). In an attempt to overcome these limitations, Quality Estimation (QE) of MT systems has been developed. The task of MTQE "consists in estimating the quality of a system's output for a given input, without any information about the expected output, that is, without reference translations" (Specia et al. 2010: 40) and it does so by "using Machine Learning (ML) models trained on labeled data points" (Scarton et al. 2016: 831). So instead of comparing the MT output to a human translation, QE predicts the quality that can be expected from the output of a particular MT system. MTQE tools are trained on large sets of data with manually or automatically assigned quality labels and learn from these data sets by making use of ML models. It is important to state that QE does not aim at estimating a score for the overall performance of an MT system, but rather at giving an estimation for an individual translation project (Specia & Shah 2018: 202).

The following two subsections first give an overview of the different possible applications of MTQE, although the two most common applications are certainly the estimation of PE effort and the selection of the best MT system since they are the most useful applications for end-users outside of academia. Then, the different levels of MTQE are explained in detail and finally, a selection of four different MTQE approaches is presented.

8.1 The applications of MTQE

The QE of MT systems has four main applications: (1) estimating the effort it takes to post-edit a particular segment in the MT output, (2) choosing between translation alternatives of different MT systems, (3) deciding on whether the output of an MT system can be used for its own learning process, and (4) selecting translation samples for quality assurance by manual inspection (Specia & Shah 2018: 203).

8.1.1 Quality estimation for predicting post-editing effort

In dissemination scenarios, whenever MT is used it is usually followed by some level of manual PE in order to render the document fit for publication. In most cases, this will accelerate the translation process since it will be faster to post-edit a given segment instead of translating it from scratch. However, there may be segments that the MT system cannot translate properly and for which PE will take more effort and time than translating the segment from scratch. If a document contains many segments like this, the amount of work for the post-editors increases remarkably which may lead to a lot of frustration, especially if the PE task gets more time-intensive than a translation from scratch would have been. Moreover, the fees for PE are usually considerably lower than for translation services as it is considered to require less time and cognitive effort. On the other hand, the situation is also frustrating for clients and language service providers as they might have to wait for longer than agreed to receive a product of lower quality than expected. To solve this problem, QE of MT output has been created, alongside other purposes.

QE aims at supporting human post-editors by predicting the level of quality that can be expected and by showing them which segments are good enough for being post-edited and which segments should be translated from scratch. Every QE tool should distinguish between at least three different levels of estimated quality: (1) segments that are of such high quality that they do not require PE, (2) segments that contain minor errors which can be corrected by PE, and (3) segments that are of such low quality that it requires less effort to translate them from scratch than to post-edit them (Specia & Shah 2018: 205).

8.1.2 Quality estimation for system selection

Another area of application for QE is when it comes to selecting the best MT system to translate a text at hand. Unlike automatic TQA methods like BLEU, QE does so again by estimating the expected quality of the MT output (Specia & Shah 2018: 208). The advantage of this method is that the ranking of the different MT systems is not based on reference translations. Therefore, QE allows for the selection of the best MT system to perform the translation of a particular text at hand, whereas reference-based TQA methods like BLEU choose the one with the best overall score (Specia & Shah 2018: 214).

8.1.3 Quality estimation for self-training

Data-driven MT systems can best be trained and improved by adding supplementary parallel training data. Yet, creating this kind of data is rather expensive since it requires human translators to translate monolingual content. It is therefore desirable to find a technique that allows for selecting the most useful sentences for improvement from a monolingual dataset and ignores the less useful ones. This selection process is often referred to as Active Learning (AL) (Specia & Shah 2018: 214f).

One strategy for AL is to use QE to identify segments of the monolingual dataset for which low MT quality can be expected since segments with good translations do not contribute much to the improvement of MT systems. Poorly translated segments, where words or phrases are incorrectly or not at all translated, one the other hand, can help to train MT systems (Specia & Shah 2018: 215). By the help of QE, the segments with a predicted low-quality MT can be identified and translators can be asked to create translations of only those segments.

8.1.4 Sampling QE for quality assurance

When it comes to assessing translation quality, usually the samples are chosen randomly. This is a good method if the dataset of available translations is big enough since a random selection of samples corresponds to the natural distributions of errors in a text. However, if the sample is too small, which is often the case especially for less common languages, some issues may be left out. QE can help to select a representative sample that contains the most important issues based on quality predictions of the segments in the samples (Specia & Shah 2018: 222).

8.2 Levels of quality estimation for MT

Quality estimation for MT can be performed at word, phrase, sentence, and document-level. Whereas systems at word and phrase-level assign labels to single words and phrases in the original and the target text which are often binary or ternary or belong to a few predefined classes, sentence-level and document-level systems usually produce a general numeric quality score either in the form of an absolute score or of a relative ranking of various translations or MT systems (Specia et al. 2018: 3).

The current research in academia focuses on word, phrase, and sentence-level systems but still faces various problems. One of the main problems is the acquisition of useful training sets since every single word or phrase in the training sets needs to be labelled (Specia et al. 2018: 2). However, since 2015, neural network models started to be used in the development

of QE systems which has led to significant improvements. Document-level systems have received less attention in academic research due to a lack of tools to extract discourse-wide information from texts, such as topic or structure, and because, again, only small datasets that have been labelled by humans are available (Specia et al. 2018: 4).

The different levels of quality estimation for MT as well as their architectures and the most important QE systems are presented in the following subsections. Since they are very similar, word- and phrase-level QE are merged under the subsentence level. Sentence-level and document-level QE are, however, described in separate subsections.

8.2.1 MTQE at subsentence level

MTQE systems that operate at subsentence level predict the quality of either single words or short phrases. They are able to give more insightful information about the nature and location of errors than sentence or document-level systems do since they provide a score for every word or phrase instead of for whole sentences or texts.

The score of word and phrase-level MTQE is usually based on the assignment of discrete labels. These labels can either be binary, ternary, or based on multiple classes. Binary systems usually decide between two categories such as *OK* and *BAD*, *GOOD* and *BAD*, or *KEEP* and *CHANGE* (Specia et al. 2018: 9). Ternary systems add a third category that gives more information about what to do with an erroneous segment. They decide, for example, between *KEEP*, *CHANGE*, and *DELETE*. Multi-class systems distinguish between segments with and without errors too. However, they do not assign a simple *DELETE* or *CHANGE* label to every segment but give more information about the type of error that occurred. If the error type is indicated, an error typology needs to be defined in advance.

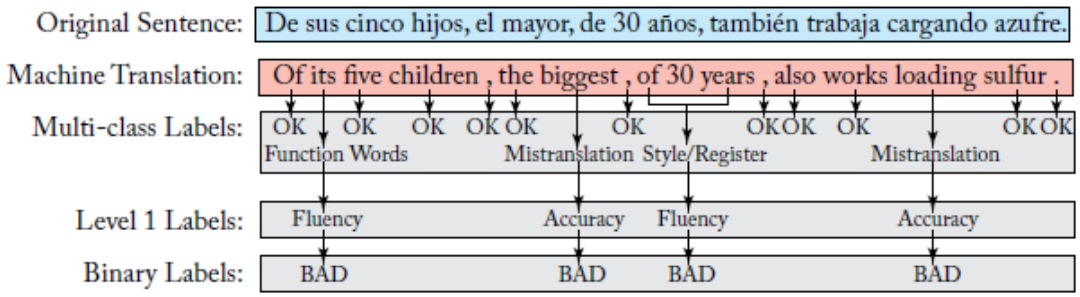

Figure 13: Word-level QE of MT (Specia et al. 2018: 11)

Fig. 13 shows an example of three different ways to perform MTQE on the word level. The Spanish original sentence was translated to English and its quality was estimated by three different MTQE systems. The first system assigned multi-class labels to the MT version. Every correct word is labelled as *OK*, whereas every erroneous word is assigned an error type taken from the second level of the MQM metric. The second system does not go that much into detail and only assigns the level-1 error types *Fluency* and *Accuracy* to the erroneous words. The third system, eventually, only assigns the binary labels *OK* and *BAD*. All three methods have some advantages and disadvantages. Whereas a more fine-grained distinction of errors gives the human translator or post-editor important information on the type of error that occurred, this also increases the complexity of the estimator tool since it needs to distinguish between subtle nuances. In order to learn how to differentiate between error types, a lot of annotated training data is required which is not always easy to get (Specia et al. 2018: 11f).

Subsentence-level QE systems for MT are built and trained based on different kinds of features. The word-level features that are used include fluency, complexity, adequacy, and confidence features. Complexity features apply to the source text, confidence features to the MT system, fluency features to the translated text and adequacy features to both source and translated text (Specia et al. 2018: 52).

Fluency features describe the target words or phrases as well as their surrounding context and allow the QE system to associate lexical patterns with quality labels. They include n-gram features, backoff behaviour features, dependency features, pseudo-reference features, and semantic features. N-gram features examine the target word and its part-of-speech (POS) tag as well as bigrams and trigrams with the target word and its surrounding words and POS tags (Specia et al. 2018: 14). So, in the case of the word *computer* taken from the example in Fig. 14, the POS tag is NN, the possible bigrams are *this computer* and *computer are*, the POS tag bigrams are DT NN and NN VBP, the mixed bigrams are DT *computer* and *computer* VBP. The creation of the trigrams works analogously. After identifying these n-grams, the MTQE system can calculate their probability based on a language model and assign quality labels (Specia et al. 2018: 14). Backoff behaviour features look at the bigrams and trigrams with the target word in it as well as at the words preceding them, compare them with a language model and assign a numeric value to the target word. The lower this value is, the higher is the probability of the target word being erroneous (Specia et al. 2018: 14f). Dependency features check if the target word at hand has a dependency relation with at least one other word in the sentence. If so, a value of 1 is added, if not, the added value is 0 (Specia et al. 2018: 16). In the case of the

word *computer*, the value would be 1 because it has several dependency relations in the sentence. Pseudo-reference features check if the target word appears in a pseudo reference and semantic features evaluate the number of senses of the target word as well as possible synonyms (Specia et al. 2018: 16).

Complexity features refer to how challenging it was for the MT system to translate a given source word or phrase and rely on alignments between the source text and the MT. They use a similar approach as the n-gram features for fluency. They evaluate the source-language word that was aligned to a given target word, its POS tag, bigrams and trigrams with the surrounding words, and bigrams and trigrams with the surrounding POS tags (Specia et al. 2018: 17). For the word *computer* the aligned source word is *computador*, the POS tag of *computador* is NOM, the possible bigrams are *deste computador* and *computador são* as well as PRP+P NOM and NOM V, and so on.

Adequacy features aim at finding out how adequate the MT is with respect to the original text by combining information from the source and the target text. Again, bigrams and trigrams are created but here the information from the source text is mixed with the target word (Specia et al. 2018: 17f). Possible bigrams for the word *computer* and its POS tag are *deste computer* and *computer são* as well as PRP+P NN and NN V, etc.

And finally, confidence features are based on the probability that a given word will be translated into another word. If the probability is very small, the MT is very likely to have made a mistake (Specia et al. 2018: 18). In this example, the probability that *computador* is translated into *computer* will be very high, whereas it is much lower for other words.

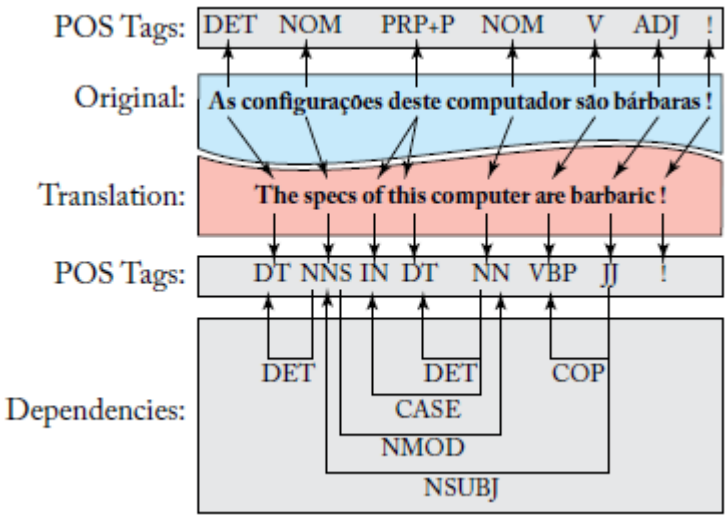

Figure 14: Example of an MT with POS tags and dependencies (Specia et al. 2018: 13)

Phrase-level features also include fluency, complexity, adequacy, and confidence features. In addition, punctuation, language model, alignment, and part-of-speech features are also taken into account. Punctuation features count the punctuation characters in the target phrase and its aligned source phrase in order to find differences. Language model features refer to the number of tokens in the target phrase and the aligned source phrase as well as to the ratio between these two numbers and the average length of the tokens. Alignment features relate to the number of words in the source text that cannot be aligned with any word in the translated text, the number of words that can be aligned with more than one word, and the number of alignments between the source text and the translated text. And finally, part-of-speech features deal with the percentage of certain word types in the source and target phrase (Specia et al. 2018: 20).

Subsentence-level MTQE systems usually use one of three different architectures: there are non-sequential approaches, sequential approaches, and APE-based approaches. Systems using the non-sequential approach are the simplest ones. Here, the quality of each word or phrase is predicted independently from other words and phrases in the text and therefore interdependencies between different words and phrases in the text are ignored (Specia et al. 2018: 21). Sequential approaches, on the other hand, take advantage of the fact that words and phrases usually are part of a sentence and are therefore linked to other words and phrases. MTQE systems relying on sequential approaches take into account the proceeding words and phrases in the sentence and the quality labels they predicted for them (Specia et al. 2018: 24). APE-based approaches, finally, use an Automatic Post Editor (APE) to create a post-edited version of the MT. The APE is trained using translations that have been post-edited by humans. In the next

step, the QE system calculates the TER between the MT version and the APE version and assigns quality labels to the words and phrases in the sentence based on the TER results (Specia et al. 2018: 27). In the example given in Fig. 15, the labels G (for *good*) and C (for *change*) are assigned to every word of the Spanish MT version depending on whether it would need to be changed during PE or not.

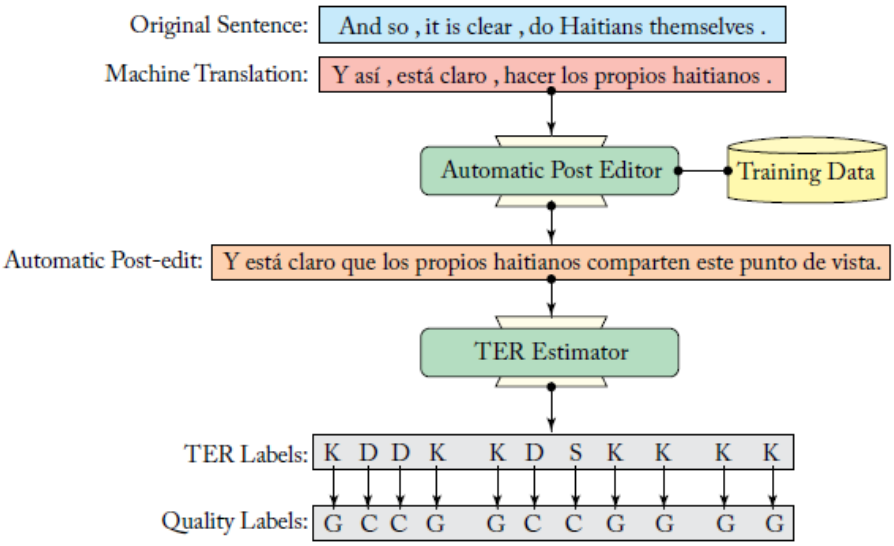

Figure 15: Architecture of an APE-based QE approach (Specia et al. 2018: 28)

Some of the most important subsentence-level MTQE systems to date are the Predictor-Estimator developed by Kim et al. (2017), the hybrid Unbabel approach by Martins et al. (2017), and the QUETCH model by Kreutzer et al. (2015). The three systems will be described in detail in Section 8.3.

8.2.2 MTQE at sentence level

Most research on QE of MT so far has focussed on sentence-level prediction of MT quality. The reason for this is quite simple. First, most MT systems process one sentence at a time and thus it makes sense that QE proceeds the same way. And even though NMT systems translate word-by-word they still take whole sentences as input. Secondly, readers usually consume whole translated sentences or even whole texts at a time and not just single words (Specia et al. 2018: 43).

In general, sentence-level MTQE assigns a numeric score to each sentence based on different criteria. One possibility is to judge a sentence on its overall quality by, for example,

rating it on a scale of 1 to 5 (Specia et al. 2018: 48f). Another possibility is to assign a score based on the edit operations required to obtain a good translation (Specia et al. 2018: 51).

As has already been stated in previous chapters, humans would judge an MT differently depending on its purpose. If it is only used for gisting purposes the evaluation will be less strict than for a text that is destined to be post-edited or published. Therefore, the linguists involved in the development of MTQE systems try to build them in a way that they are able to predict quality based on the text's purpose in the target-language community (Specia et al. 2018: 43).

The features that are used for building sentence-level MTQE systems can be either glass-box features, i.e. the same features as in the MT system that generated the translation, or black-box features, i.e. independently chosen features (Specia et al. 2018: 51). Similar to word-level MTQE systems, sentence-level systems also rely on complexity features that reflect how hard it is to translate a text (Specia et al. 2018: 52), fluency features that "attempt to measure how natural, fluent, and grammatical the translation is" (Specia et al. 2018: 53), confidence features that show how probable the translation is, and adequacy features that refer to the relation of the original and the translation on different linguistic levels (Specia et al. 2018: 54). Complexity features take into account the length and type of a source sentence, the number of relative clauses that appear in it, the average frequency of source words in the source-language corpus, the average number of suggested translations for every source word, etc. (Specia et al. 2018: 52f). Fluency features check the length and type of the target sentence, its coherence and grammaticality, and the number of untranslated words (Specia et al. 2018: 53). Confidence features examine the global score of an MT system, the number of translation hypotheses in the n-best list as well as their length, the frequency of the word in the translation in the n-best list, etc. (Specia et al. 2018: 54). Adequacy features, finally, compare the target sentence with its aligned source sentence and deal with the ratio of the number of tokens in the source and the target sentence, examine their similarity using vector representations, count the maximum number of aligned words and the number of unaligned words, etc. (Specia et al. 2018: 54f).

Apart from those features, there are also pseudo-reference and back-translation features as well as linguistically motivated features that have to be taken into account when building a sentence-level MTQE system. Pseudo references are created when the same sentence is translated by another MT system than the one that the MTQE system wants to predict the quality for and the two results are then compared. The idea is that if the two versions are very similar, the translation is likely to be of high quality (Specia et al. 2018: 55). Features based on back

translation work quite similarly. Here, the translated sentence is translated back into the source language either by the same MT system or by another one. If the two source-language versions of the sentence are quite similar, it is very likely that the translation is good (Specia et al. 2018: 55). Linguistically motivated features are features that are specific for each language and language combination, like the differentiation between the formal and the informal way to address another person in German (du/Sie) that does not exist in English (Specia et al. 2018: 56).

Two of the most important sentence-level MTQE systems are the Predictor-Estimator system by Kim et al. (2017) and the hybrid Unbabel approach by Martins et al. (2017) which can both be deployed at subsentence level and sentence level. As has been stated before, they will both be explained more in detail in Section 8.3.

8.2.3 MTQE at document level

The document level is the least explored granularity level for MTQE because, amongst other reasons, finding an adequate amount of useful training data is rather difficult (Specia et al. 2018: 63). However, performing MTQE at document level would be very useful since most MT systems work at sentence level and are oblivious to information that extends over more than one sentence or even the whole document. MT texts are thus prone to inter-sentence incoherence (Specia et al. 2018: 61).

Document-level MTQE is more than just a combination or single word, phrase, or sentence-level scores, since words, phrases, and sentences can be good in isolation but not when put together as a whole text and vice versa. Moreover, different sentences can be of different relevance in a text, so if a less relevant sentence is of low quality this should not affect the overall document score in the same way a highly relevant low-quality sentence does (Specia et al. 2018: 62).

The creation and training of document-level MTQE systems also rely on complexity, fluency, and adequacy features. Here, the complexity features refer to how difficult it is for the MT system to translate the whole text correctly, which is measured, amongst others, by counting the number of tokens in the source document, the average number of translation suggestions for every source word, and the average frequency of a given source word in a source corpus (Specia et al. 2018: 69). Fluency features capture the similarity of the translated text with other original texts on the same topic in the target language by counting the number of tokens, the number of punctuation marks, and the number of sentences in the target document, by calculating the probability of the n-grams in the target document, etc. (Specia et al. 2018: 70). Adequacy

features combine the information of the source and the target document to see if the translation is adequate, amongst others by comparing the ratio of certain word types in the source document and in the translated document (Specia et al. 2018: 70f). Apart from these features, discourse-aware features, word-embedding features, as well as consensus and pseudo-reference features also play a role in document-level MTQE systems. Discourse-aware features aim at measuring text cohesion by considering word repetitions and/or using latent semantic analysis. The assumption is that a text which presents a high number of repetitions has high lexical cohesion. Latent semantic analysis, on the other hand, captures the topic of a text based on the words that are in it. So, if there are many words in the document that do not fit the topic, the document is probably of low quality (Specia et al. 2018: 71). Word-embedding features compare the word embeddings in the source text and the target text to assess complexity and fluency of the translation (Specia et al. 2018: 73) and pseudo-reference features are again used to find the degree of consensus between different MT systems which can be an indicator for either high or low quality (Specia et al. 2018: 74).

8.3 Recent MTQE systems

Now that the different levels of MTQE have been presented, a closer look can be taken at recent MTQE systems. In the past decade, a great number of different systems for predicting the quality of MT output have been developed. As has been stated in the previous section, most systems focus on the QE at sentence level, whereas fewer systems work at word and phrase level and even fewer ones at document level. This section gives an overview of some MTQE systems that have shown the best results during the past few years.

8.3.1 QUETCH for word-level QE

Kreutzer et al. (2015) present a word-level QE model that uses deep neural networks for learning bilingual feature representations from raw input words. They call their system QUality Estimation from scraTCH, or QUETCH. Instead of relying on manual feature engineering like its forerunners, QUETCH learns representations in a continuous space by taking discrete units like words or n-grams and replacing them by vectors (Kreutzer et al. 2015: 316f).

As can be seen in Fig. 16, the architecture of the QUETCH model consists of three layers: the input layer, one hidden layer, and the output layer. The hidden layer is further separated into a lookup-table layer and a linear layer. First, the source and target words are aligned in the input layer. A word window of a fixed size and centred at a certain target word is extracted

from the target sentence. From the source sentence, a word window of the same size centred at the aligned source word is extracted, too. The two word windows are then combined in order to obtain a context vector for the target word. The context vectors of every target word of the sentence or text at hand are used as input for the lookup-table layer. They get transformed into a binary label and the output layer finally shows either “OK” or “BAD” (Kreutzer et al. 2015: 317).

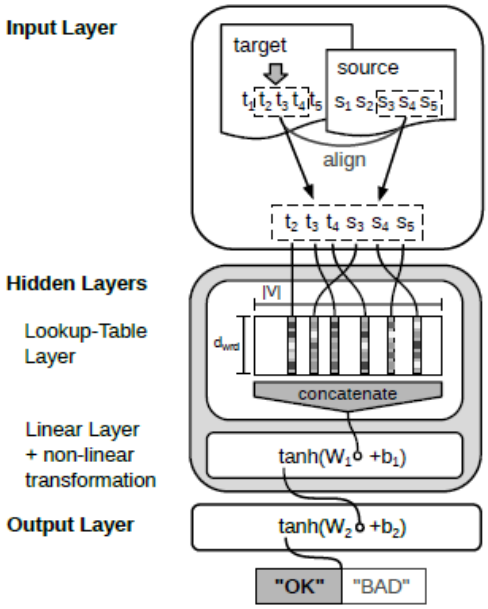

Figure 16: The architecture of QUETCH (Kreutzer et al. 2015: 317)

8.3.2 The Predictor-Estimator model for subsentence and sentence-level QE

Kim et al. (2017) developed a two-stage end-to-end neural QE model known as the Predictor-Estimator model that can be deployed for predicting translation quality at word, phrase, and sentence level. The model consists of two parts: (1) a neural word-prediction model (the Predictor part) which is trained by large parallel corpora, and (2) a neural QE model (the Estimator part) which is trained by quality-annotated parallel corpora. First, the word predictor predicts a target word based on the context where it appears. The target word is represented together with its source and target context by a so-called QE featured vector, which is then used as input for the quality estimator. The quality estimator takes the vector and assigns a quality label to it (Kim et al. 2017: 562). This process is pictured in Fig. 17 below. Moreover, Kim et al. (2017) added stack propagation to this two-stage architecture. This means that information that is learned by the estimator is given back to the word predictor and thus the two components can be trained and improved jointly (Kim et al. 2017: 563).

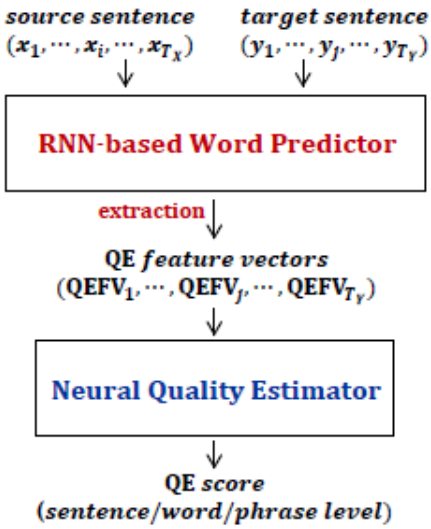

Figure 17: Predictor-Estimator architecture (Kim et al. 2017: 563)

Both, the predictor and the estimator are based on Recurrent Neural Networks (RNN). Like every neural network, RNNs consist of at least three layers: an input layer, one or more hidden layers, and an output layer. Each layer consists of several nodes. Whenever a piece of information, say a given target word in a translation, is passed to a node in an RNN, the node transforms this information and creates a hidden representation of it (Specia et al. 2018: 25). In contrast to non-recurrent neural networks, however, a node in an RNN passes the hidden representation not only to the next layer but also adds it to the information for the next word (Specia et al. 2018: 26).

The advantages of the Predictor-Estimator model are its cost-effectiveness as well as the fact that it does not require feature engineering. Usually, the creation and collection of QE data for training purposes is a rather expensive task. Yet, in the Predictor-Estimator model, the quality estimator learns from data that it obtains from the word predictor which saves costs. And since the predictor learns from embedding representations and the estimator from the vectors it gets as input, there is no need for feature engineering. The big drawback, however, is that the system is very complex. It consists of several parts that can be configured in many different ways and therefore a lot of experience with neural networks is needed in order to use the Predictor-Estimator model (Specia et al. 2018: 51).

8.3.3 Unbabel's hybrid approach for subsentence and sentence-level QE

Martins et al. (2017) propose an MTQE model that is built on a hybrid approach containing a linear model, a neural model, and an APE model. They differentiate between two types of QE: pure QE and APE-based QE, which they both combine in their hybrid model.

The pure QE part also consists of two parts: a linear sequential model and a neural system. The linear sequential model receives an input containing the source sentence, the target sentence, and a set of word alignments and produces binary labels (OK and BAD) for each sentence (Martins et al. 2017: 207). The neural system also receives the source and target sentence and word alignments as input. Moreover, it also receives information about the corresponding part-of-speech tags. From this information, a vector is created for every word and its context and a label (OK or BAD) is assigned to it (Martins et al. 2017: 209).

The APE-based QE part uses APE to overcome the problem of little available training data. It performs a roundtrip translation by first collecting a large corpus of monolingual sentences in the target language of the task at hand. These sentences are seen as artificial post-edited sentences. In the next step, the sentences in the corpus are translated to the source language of the task by an MT system, creating the source sentences. Finally, another MT system translates the sentences back to the target language, creating the target sentences. Then, two systems inside the model can be created: the first one uses the source-language sentence as input and the PE version as output and the second one uses the target-language sentence as input and the PE version as output. Now, HTER labels can be calculated for the sentences in order to obtain quality labels (Martins et al. 2017: 211).

These three components were stacked together by Martins et al. to receive one QE system for predicting translation quality at word and sentence level (Martins et al. 2017: 212). The approach has proven to deliver very reliable results, however, its big drawback is that it is rather complex since it uses three different systems that have to be combined in the right way (Specia et al. 2018: 38).

8.3.4 Referential Translation Machines for subsentence, sentence, and document-level QE

Biçici and Way (2014) propose an approach that uses Referential Translation Machines (RTMs) to predict the quality of a translated text at subsentence, sentence, and document level. An RTM is a "computational model for identifying the translation acts between any two data sets with respect to a reference corpus selected in the same domain" (Biçici & Way 2014: 313). The way

how RTM models work is depicted in Fig. 18. An algorithm first looks at the sentences in the training set and the test set as well as at a parallel corpus and a target-language corpus and selects interpretants, that is data that is close to the training set and the test set. An MT Performance Prediction System (MTPPS) then takes the interpretants and the training data and uses machine learning to generate features of the training set and the test set. By looking at these features, the QE system can measure how well the test set matches the training set and uses this information to give a prediction of the expected translation quality (Biçici &Way 2014: 314; Biçici 2019: 74).

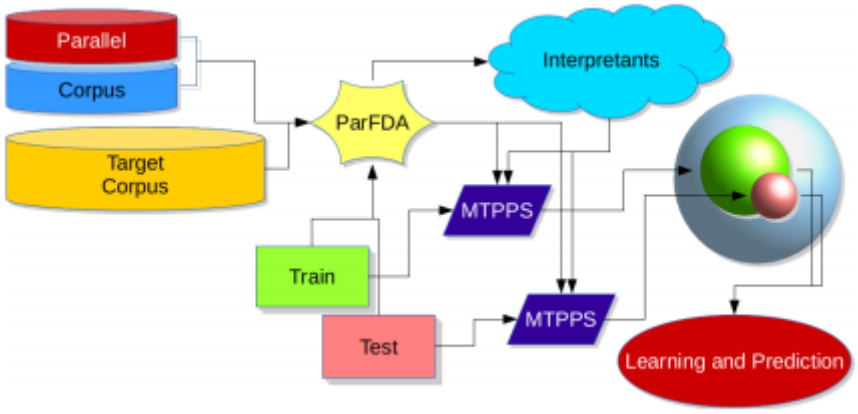

Figure 18: Depiction of the RTM approach by Biçici (2019: 74)

9 Related Work

Since (machine) translation quality and its assessment and prediction have become a topic of much debate in translation studies, the translation industry, and the MT community, especially in the last few years, a lot of research has been conducted on the topic. The studies that are presented in this chapter examine state-of-the-art TQA methods in order to point out universal issues and give recommendations on how to overcome them. Doherty (2017) suggests the incorporation of psychometric principles, Castilho et al. (2018) highlight issues in human TQA in contrast to automated TQA, and Läubli et al. (2020) present a set of recommendations for achieving human-machine parity. Another area of research that is related to the topic of this thesis are studies on the usability and usefulness of MTQE tools. Turchi et al. (2015) conducted an experiment to find out how useful MTQE really is in terms of productivity gain measured in PE time. Parra Escartín et al. (2017) also aimed at assessing the productivity gain by measuring the PE time needed, but also by counting the number of keystrokes.

Doherty (2017) takes a look at various usage scenarios for TQA and their approaches in order to point out some universal issues that can be found in all of them as well as to suggest ways to overcome them. He first presents approaches to TQA in translation studies, MT research, and the translation industry before turning to the shortcomings in these approaches and giving suggestions for improvement by incorporating psychometric principles to state-of-the-art TQA methods.

In terms of TQA in translation studies, Doherty (2017: 132) claims that the discussion on translation quality is a very theoretical one that is centred on source vs. target orientation and the importance of the purpose and function of a translation and that more empirical work in the area is needed. The key issues that he finds are a lack of research in the field of TQA, error proneness and subjectivity, a lack of systematic approaches, inconsistency in terminology, and the choice of evaluators. In terms of TQA in MT research, Doherty (2017: 133f) distinguishes between human and automatic TQA. For human TQA he states that mostly the adequacy and fluency of a text are measured at segment level, without taking into account the broader context of the segment. The issues that he sees in this procedure are forced dichotomies between good and bad translations, the use of interval scales, and the subjectivity of human evaluators (Doherty 2017: 134). For automatic TQA, on the other hand, the automatic evaluation metrics BLEU (Papineni et al. 2002), the General Text Matcher (GTM) (Turian et al. 2006), and TER (Snover et al. 2006) are mentioned. However, Doherty (2017: 134f) criticises that the

results of these metrics often do not correlate with human judgement and that they are often misused at sentence level even though they are intended for document-level evaluation. He also points out that subjectivity is not eliminated in automatic evaluation metrics since they make use of human reference translations. Moreover, MT systems are not completely objective neither since they are built based upon human data (Doherty 2017: 135). In the translation industry, TQA is mostly performed by humans or using semi-automated methods to evaluate translators and checking acceptance as well as for training and pricing. Even though it is customer-focused, the opinions on translation quality of language service providers and their clients often differ a lot, potentially causing mismatched expectations and conflicts (Doherty 2017: 136f).

After listing the most frequent issues in TQA, Doherty (2017) suggests the incorporation of psychometric principles, that is objective psychological measurements, to TQA. The first psychometric principle that is suggested is validity, i.e. assessing whether TQA is doing what it is intended to do. Validity can be divided into internal and external validity. Internal validity refers to the content of the assessment task at hand, whereas external validity means that the items that are being assessed should be items that occur in real-world translation situations and are therefore representative (Doherty 2017: 138). The second psychometric principle that Doherty (2017: 139) proposes is reliability that is both the intra-evaluator and the inter-evaluator consistency. The choice of evaluators is another psychometric principle that should be taken into account. Most studies on TQA do not provide any information on the evaluators and the training or guidelines they were given; however, it is important that evaluators have a certain level of expertise and receive TQA-specific training (Doherty 2017: 140). Apart from that, Doherty (2017: 141) also claims that the instructions given to the evaluators should be thought of. The presentation and context of the task can have a strong impact on the results and for example, the decision on whether or not TQA is conducted at sentence level without considering the context of a segment should be taken consciously.

Castilho et al. (2018) strike a similar path as Doherty (2017) with their work and examine established and developing approaches to translation quality assessment in HT and MT workflows and present an overview of these approaches in order to identify universal issues inherent to all of them. They first make a distinction between TQA in translation studies and TQA in the translation industry and present some of the most important approaches in both domains. The TQA approaches in translation studies are divided into the "equivalence approach" (Castilho et al. 2018: 12) that focuses on a reproduction of the source text that is as close to the original as possible, the "descriptive and markedly target-oriented approach"

(Castilho et al. 2018: 13) that takes the target text as starting point for the assessment, and the "functionalist (or Skopos) approach" (Castilho et al. 2018: 13) that decides on the adequate translation method based on the purpose of a target text. Regarding TQA methods in the translation industry, Castilho et al. (2018: 15) state that there is a tendency to quantify and refer to various standards used in the field.

Then, a distinction between TQA performed by humans and automated TQA is made. In terms of human TQA, the assessment of a translation's adequacy and fluency, its readability and comprehensibility, its accessibility, and its usability and performance are specified. These can be assessed, amongst others, by the use of ordinal scales or rankings (Castilho et al. 2018: 17ff). Furthermore, Castilho et al. (2018: 23) state that the choice of evaluators and their performance should be discussed too. In terms of automated TQA methods, on the other hand, the Word Error Rate (WER) (Nießen et al. 2000), the Translation Error Rate (TER) and the Human-targeted TER (HTER) (Snover et al. 2006), the BiLingual Evaluation Understudy (BLEU) (Papineni et al. 2002), the General Text Matcher (GTM) (Turian et al. 2006), and METEOR (Lavie & Agarwal 2007) are referred to.

Finally, the universal issues that Castilho et al. (2018: 29ff) identify based on their overview of TQA methods are a lack of standardisation of TQA usage, inherent inconsistency in TQA, the relationship between human and automatic measures, social quality and risk, and education and training. They criticise that standardised TQA methods exist only for some separate industry sectors, however, in general, there is no clear definition of what translation quality and its assessment mean, and the opinions of suppliers and clients often differ greatly (Castilho et al. 2018: 30). Furthermore, especially human TQA is often quite inconsistent and the correlation between human and machine TQA is often very low (Castilho et al. 2018: 30). Human evaluators are, moreover, also influenced by external factors such as social, ethical, and qualitative implications as well as by their education and training for the task at hand (Castilho et al. 2018: 31). All these issues should be taken into account by TQA research since they limit the performance of current TQA methods and should therefore be overcome.

Läubli et al. (2020) investigate three aspects of human MT evaluation in response to previous studies that claimed to have achieved human-machine parity for NMT. The aspects that they investigate and empirically test are the choice of evaluators in TQA tasks, the use of linguistic context, and the creation of reference translations.

Both professionals and amateur evaluators are regularly used for TQA tasks. Läubli et al. (2020: 657f) aimed at finding out to what extent the choice of evaluators impacts the results of the TQA task at hand. They conducted an experiment with two professional translators and three non-expert evaluators and asked them to rank sentences that have been translated from Chinese to English by a human translator and two MT systems. The results showed that the professional translators are able to differentiate more clearly between the HT and the MT versions than the non-experts (Läubli et al. 2020: 658).

In terms of linguistic context, Läubli et al. (2020) assumed that providing the evaluators with more context will lead to better results in terms of human-machine parity. They performed another ranking task where one part of the evaluators received only isolated sentences and the other part received entire documents that they had to rank (Läubli et al. 2020: 659). The results showed that document-level context helps with adequacy since the difference of the scores between the HT and MT versions was much wider at document level than at sentence level. In terms of fluency, the HT was clearly considered to be better both at sentence and at document level. This shows that providing linguistic context to evaluators is of great importance for TQA (Läubli et al. 2020: 660f).

For the creation of reference translations, quality and directionality should be considered. In terms of quality, errors can occur since the reference translations are created by humans. Possible errors are misinterpretations of the source text's content, errors in fluency, errors due to limited resources such as limited time, attention, and motivation of the translators, and the effects of post-editing that may cause unusual sounding MT parts remaining in the translation (Läubli et al. 2020: 662). Läubli et al. (2020) examined the same set of translations as before and asked the evaluators to categorise the errors that they found based on a predefined error typology. They found out that the MT versions showed more issues concerning incorrect words, named entity errors, incorrect word order, and items that depend on inter-sentence context. This confirms the assumption that MT is still less effective in choosing correct translations and that the problem of incorrect word order in MT is not solved yet (Läubli et al. 2020: 663). In terms of directionality, Läubli et al. (2020: 667f) showed that using translationese results, that is source texts that are not originals but that have been translated to the source language, results in higher scores for MT, however, their lexical variety is smaller, and they are more explicit and normalised.

After the empirical assessment of the choice of evaluators, the use of linguistic context, and the creation of reference translation, Läubli et al. (2020) state that full human-machine parity has not been achieved yet. They present five recommendations to the MT community in order to come closer to reaching this goal, including the choice of professional translators as evaluators for TQA tasks, the evaluation of documents instead of isolated sentences, the evaluation of fluency in addition to accuracy, the application of only light editing of reference translations, and the use of original source texts instead of translated ones (Läubli et al. 2020: 668f).

Turchi et al. (2015) state that MTQE is thought to support translators using CAT tools during the translation and/or PE process and thus speeds up their work. The aim of their experiment was to find out if this is really true, i.e. whether or not the implementation of QE in CAT tools leads to a productivity gain (Turchi et al. 2015: 530).

The CAT tool that was used for the experiment was the open-source tool MateCat with two small changes. First, only one suggestion was given per segment instead of the three that MateCat normally gives, and second, every suggestion was coloured either in green, implying that post-editing this segment is expected to be faster than translation from scratch, in red, implying that PE would be slower, or in grey, implying that no quality label has been assigned to this segment (Turchi et al. 2015: 531). The document, a user manual from the IT domain, was then split into two parts, the first part being used for training and the second part for the experiment itself. For training, the first part of the document was translated by an MT system and then post-edited by human post-editors. Next, the HTER score for every segment was calculated and quality labels were assigned. Segments with an HTER score of 0.4 or lower were labelled as *good* and were coloured in green, whereas segments with an HTER score of above 0.4 were labelled as *bad* and were coloured in red (Turchi et al. 2015: 531). The other half of the document was then translated from English to Italian by an MT system and sent to four professional translators who were asked to post-edit all the segments they received. Half of the segments of each translator were kept in grey and the other half was coloured, with 75% green segments and 25% red segments (Turchi et al. 2015: 531f).

The results of the experiment showed that QE can indeed increase productivity, however, not to the extent as it was expected by Turchi et al. (2015). For 51.7% of the segments, the average PE time was lower for coloured segments than for grey ones. This shows that QE is useful, but its contribution is rather small. The best results were achieved for sentences of medium length, i.e., sentences containing between six and twenty words, with an HTER score

of above 0.2 and lower than 0.5. For sentences belonging to this group, the PE process was speeded up in 56% of the cases.

Parra Escartín et al. (2017) conducted a user study that was also aimed at finding out how quality estimation of MT can be used to improve the productivity of professional translators (Parra Escartín et al. 2017: 344). Four professional translators, who were all Spanish native speakers and had several years of experience working with CAT tools and PE tasks (Parra Escartín et al. 2017: 349) were confronted with a dataset of 40 English source language segments and Spanish MT output segments that had been evaluated by an MTQE framework (Parra Escartín et al. 2017: 346f). The segments were presented with a traffic light system that should indicate the type of task that the translators were facing: Light yellow meant that no MT output was given to the translators and that they, therefore, had to translate from scratch, light blue meant that MT output was given to the translators but no MTQE information and that they had to decide for themselves whether to post-edit or to translate from scratch, light green meant that the MTQE system suggested post-editing, and light red meant that the MTQE system suggested translating the segment from scratch (Parra Escartín et al. 2017: 348f). The post-editing tool PET was used to record the number of keystrokes pressed by the translators as well as the time needed (Parra Escartín et al. 2017: 347).

The study showed that the translators needed less time for segments with a good quality estimation. The number of keystrokes needed was lower for post-editing than when translating from scratch and it was highest for the segments that offered no QE, whereas it was lower for segments with bad QE. This shows that even bad QE seems to be a better aid than no QE at all in terms of effort measured in keystrokes (Parra Escartín et al. 2017: 350). However, when asked for their opinion after the study, three out of four translators stated that they did not find the MTQE helpful (Parra Escartín et al. 2017: 351). The conclusion that Parra Escartín et al. draw from their study is that MTQE makes translation workflows more efficient in terms of time and effort needed, especially if the MTQE is good and accurate, but that a better post-editing tool might be needed in order to convince the translators of the usefulness of MTQE (Parra Escartín et al. 2017: 352).

10 Comparative Analysis

The previous chapters have demonstrated that a lot of different quality criteria can play a role in the assessment of translation quality, depending on the underlying understanding of translation itself. In the course of the comparative analysis, the understanding of translation quality in translation studies and industry versus the understanding of translation quality used in the development of MTQE tools are contrasted and compared. To do so, the quality criteria found in different approaches are merged into two tables: one that illustrates the quality criteria that are considered important in translation studies (Table 3) and one with the quality features used in the creation of MTQE tools (Table 4).

In order to give an overview of the different concepts of translation quality in translation studies, the approaches of Reiss (1971), House (1997), Reiss and Vermeer (1984), Ammann (1990), van den Broeck (1985), Mossop et al. (2020), and Göpferich (2008) have been explained in detail and the quality criteria that they use have been highlighted. Next, it has been stated that the quality of translation is of course also an important topic in the translation industry and that quality criteria have been found in the industry too in order to assure a certain level of quality. The SAE J2450 quality model (Schütz 1999; Hui 2017) and the Multidimension Quality Metric (Lommel et al. 2015) have been presented as two of the most widely used quality models in the translation industry. Moreover, Popović's (2018) error taxonomy for manual error classification and analysis has been introduced in Chapter 6.

In terms of QE for MT, four different systems, the QUETCH model (Kreutzer et al. 2015), the Predictor-Estimator model (Kim et al. 2017), the hybrid approach by Martins et al. (2017) and the use of Referential Translation Machines (Biçici & Way 2014; Biçici 2019), have been presented and the quality features used for QE at subsentence, sentence, and document level have been listed.

10.1 Overview of the quality criteria

In order to perform the comparative analysis, first, a look is taken at the beforementioned approaches on translation quality in translation studies and the translation industry. Considering all the approaches, the criteria for translation quality can roughly be divided into textual and extra-textual features. Textual features include features concerning the language and the content of the translation, i.e., what the text itself consists of. Extra-textual features, on the other hand, involve everything that is not immediately part of the text but has an impact on it, such as

design, text type, medium, function, sender(s), intended recipient(s), the time and place of creation, and possible given specifications.

The quality of the language of a translated text is good if lexis, morphology, syntax, orthography, and style are correct. Lexis includes features such as lexical adequacy, consistency, additions, omissions, words that are left in the source language without justification, and words that have been translated to the target language even though they are not supposed to. Morphology refers to the correct formation of words by inflection, derivation, and composition. The syntax of a sentence is correct if the grammar rules are observed. Orthography refers to adequate punctuation, spelling, and capitalisation according to the rules in the target language. And style, eventually, involves the application of the correct terminology and the adequate language register.

The quality of a translation in terms of content can be measured by checking if the translation is semantically accurate, i.e., if the meaning of the target text is equivalent to the meaning of the source text. Another aspect, that can be assessed in terms of a text's content, is its comprehensibility. The comprehensibility can be evaluated by looking at the dimensions that have been brought forward by Göpferich (2008), which are concision, correctness, motivation, structure, simplicity, and perceptibility.

Apart from these two textual factors, there are also several extra-textual factors that need to be taken into account whenever a judgement on translation quality is expressed since they have a great influence on the actual quality in the concrete situation. These extra-textual factors include the design of the text, the text type, the medium in which the text is presented, the function that the text conveys, the time and place where the source text has been created, the sender, the intended recipients, and possible given specifications.

The design of a text involves its layout, the typography, the structure, as well as elements of text convention in the target language. A good design is important to make the content more understandable, especially in terms of perceptibility. The different text types have been covered by Reiss (1971), who differentiates between informative, expressive, and operative texts. Depending on which text type a translated text belongs to, the decision of which quality criteria should be treated with priority changes. The function of a translated text is one of the most important, if not the most important, quality criterion since it influences the translator's choice on almost all textual quality criteria. Depending on the function that the translation should have in the target-language culture, adequate lexis, syntax, and style have to be chosen. All the

beforementioned approaches to translation quality, as well as the MT and PE communities, agree that the function of a text needs to be preserved by all means and that the adequate linguistic and semantic tools need to be chosen in order to achieve this goal. A factor that is related closely to the function is the intended group of recipients. Depending on parameters such as their age, sex, social status, education, etc., the textual features of a translation have to be chosen. Also, the social role relationship between the recipients and the sender, the social attitude towards them, and their participation in the communicative act influence the choice of linguistic and semantic aspects. And finally, some specifications, such as style guides, controlled languages, legal requirements, etc., might be given which should be observed in the translation process. Table 3 presents an overview of all these factors.

Table 3: Overview of the most important quality criteria in translation studies

Textual factors	Language	Lexis
		Morphology
		Syntax
		Orthography
		Style/Register
	Content	Accuracy/semantic equivalence
		Comprehensibility
Extra-textual factors	Design	Layout
		Typography
		Structure/organisation
		Elements of text convention
	Text type	
	Medium	
	Purpose/function	
	Time factor	
	Place factor	
	Sender	Stance
		Provenance
		Social role relationship
		Social attitude

		Participation
	Recipients	Age
		Sex
		Stance
		Provenance
		Education
		Participation
		Social role relationship
		Social attitude
	Specifications	Style guides
		Legal requirements
		Client's specifications
		Employer's policies

The quality features that are used in the development of MTQE systems, on the other hand, differ depending on the level of QE, as has been shown in Chapter 8. On the word level, fluency, complexity, adequacy, and confidence features are used. The fluency features can further be divided into n-gram features, backoff behaviour features, dependency features, pseudo-reference features, and semantic features. On the phrase level, apart from fluency, complexity, adequacy, and confidence features, there are also punctuation, language model, alignment, and part-of-speech features. Here, the fluency features can be divided into context features and backoff behaviour features. Sentence-level QE systems also rely on fluency, complexity, adequacy, and confidence features, as well as on pseudo-reference features, back-translation features, and linguistically motivated features. QE on the document level, eventually, is based on fluency, complexity, and adequacy features as well as on discourse-aware features, word-embedding features, and pseudo-reference features. Table 4 gives an overview of all the quality criteria used on the different levels of MTQE as they have been described in Section 8.2.

Table 4: Quality criteria in MTQE at subsentence, sentence, and document level

Word-level QE	Fluency	n-gram features
		Backoff behaviour features
		Dependency features
		Pseudo-reference feature
		Semantic features
	Complexity	
	Adequacy	
	Confidence	
Phrase-level QE	Fluency	Context features
		Backoff behaviour features
	Complexity	
	Adequacy	
	Confidence	
	Punctuation	
	Language model	
	Alignment	
	Part-of-speech	
Sentence-level QE	Fluency	
	Complexity	
	Adequacy	
	Confidence	
	Pseudo-reference features	
	Back-translation features	
	Linguistically motivated features	
Document-level QE	Fluency	
	Complexity	
	Adequacy	
	Discourse-aware features	
	Word-embedding features	
	Pseudo-reference features	

10.2 Comparison of the quality criteria

By comparing the criteria shown in Table 3 and Table 4, overlaps, that is criteria that appear in both, translation studies and MTQE research, can be identified as well as differences, that is criteria that are present in translation studies but absent in MTQE research.

In terms of the textual factors that are considered essential for a high-quality translation in the translation studies and the translation industry, a lot can also be found in the features that are computed in MTQE systems. Lexical adequacy can be assured by backoff behaviour features and pseudo-reference features at the word level since they calculate the probability of the target word appearing in its context based on a language model and a pseudo reference translation. Possible additions, omissions, and untranslated words are covered by the phrase-level alignment features, which check if there are any words in the target phrase that are not aligned to a word in the source phrase and if there are any words that are aligned to more than one word. The correct morphology of words can be assessed by word-level dependency features. The correctness of the syntax is assured by adequacy features at the sentence level, which examine the similarity of the target sentence with the source sentence, as well as by dependency features, which deal with dependencies between words in a sentence. The punctuation of the translated text is covered by punctuation features at the phrase-level and semantic equivalence, or accuracy, are considered by MTQE systems at word level by semantic features. In terms of content, the correct register can be evaluated by linguistically motivated features at the sentence level. However, there are also several quality criteria at the textual level that the translation studies consider as important but that are not part of any MTQE system so far. First, MTQE systems have no way of estimating if a translation will be consistent in its lexis, syntax, and style. Moreover, the features of a text's comprehensibility cannot be taken into account by state-of-the-art MTQE systems.

In terms of extra-textual quality criteria, the situation is very different. MTQE systems are not yet able to consider any extra-textual factors that have been mentioned before. Even though the extra-textual factors strongly determine the form of the linguistic and semantic features that are chosen in the course of the translation process, they are not part of the development of MTQE tools. MTQE is therefore not capable to take into account the factors that determine whether or not a text is of high quality but that are not visible in the text, and can, therefore, not consider these factors in their prediction scores for translation quality. All these results are illustrated in Table 5 for a better overview.

Table 5: Comparison of the quality criteria in translation studies and MTQE tools

Overlaps in translation quality criteria	
Quality criteria in translation studies	**Corresponding quality criterion/criteria in MTQE tools**
Lexis (only lexical adequacy, but not consistency)	Backoff behaviour features
	Pseudo-reference features
Morphology	Dependency features
Syntax (only syntactical correctness, but not consistency)	Adequacy features
	Dependency features
Punctuation (as part of orthography)	Punctuation features
Semantic equivalence	Semantic features
Register	Linguistically motivated features
Quality criteria in translation studies that have not yet been incorporated into MTQE tools	
Comprehensibility	
Design	
Text type	
Medium	
Purpose/function	
Time factor	
Place factor	
Sender	
Recipients	
Specifications	

11 Discussion

As has been shown in the course of the literature review and the comparative analysis, the performance of MTQE systems in terms of the textual factors is quite satisfactory. Correctness at the lexical and syntactical level as well as in terms of punctuation and morphology can be assessed by considering alignments and probabilities and semantic equivalence can be assured by using lexical databases. Also, the correct style can be guaranteed by using linguistically motivated features that look at the specific linguistic aspects of the languages involved in the task at hand.

However, all the textual factors that depend on extra-linguistic factors, such as comprehensibility, cannot be assessed by MTQE systems so far and are therefore not part of the estimation scores they produce. The same applies to all the extra-textual factors that have been found and listed in Table 3 in the previous chapter. Whereas a human evaluator would also include the function of the text, the text type, the design, the medium, time and place, sender(s) and intended recipient(s), as well as given specifications to his or her judgment and check for whether or not they have been taken into account by the translator, MTQE systems fail to capture these features. Thus, in the score that they deliver, no extra-textual factors of any kind are considered. However, all the approaches on quality criteria in the translation studies that have been presented in this thesis agree that extra-textual factors play an essential role in the quality of a translated text; many approaches, such as the text typological approach by Reiss (1971), House's (1997) pragmalinguistic approach, and the functional approaches by Reiss and Vermeer (1984) and Ammann (1990) consider the function of a text as the central criterion that must, by all means, be preserved in the target text, since it influences numerous textual features. Also, the MT community who distinguishes mainly between MT for assimilation and dissemination purposes, as has been stated before, considers a text's purpose as one of the main aspects, even though MT systems still often struggle to preserve the purpose. And last but not least, for performing PE of a translated text, which is one of the main application fields of MTQE, it is important to know which purpose the translation has in the target-language community, whether it is used for gisting only or should be published. This differentiation determines the level of translation quality that is necessary.

Moreover, whereas a human evaluator is able to judge a text at all levels, MTQE systems usually estimate the quality of a translation based on an analysis of either the word, phrase, sentence, or document level but not at every level. Most of them, apart from some exceptions

such as RTMs, are simply not capable to perform QE at all levels. Yet, there are some aspects that are only covered by QE on one level but not on the others, and they might be left out of consideration because an MTQE system that does not cover this level is chosen. For example, lexical and grammatical consistency can only be considered by document-level MTQE systems since all other systems look at either single words, phrases, or sentences and are not able to assess anything that goes beyond that level. Since state-of-the-art MT systems all work at sentence level and are mostly oblivious to inter-sentence coherence and consistency, it is even more important that MTQE systems are able to take this level into account. However, as has been stated, the document level has received the least attention so far resulting in the fact that there are not as many systems that can perform document-level QE as there are systems for subsentence or sentence-level QE.

On the other hand, there are also aspects that word-level QE systems can include but sentence or document-level systems cannot, especially fine-grained aspects that concern single words only and can demonstrate where exactly an error happened. For example, lexical adequacy and semantic equivalence are assessed only at word level, whereas sentence-level systems can only recognise that there is some issue somewhere in the sentence, but they do not know where it is exactly and what type of issue occurred.

Finally, if QE is performed in order to support post-editors, sentence-level QE results are the most useful ones since post-editors usually look at the translation sentence by sentence and thus need a quality label for each sentence. Word-level QE can also give interesting information, e.g. about the exact location of the error that needs to be changed by the post-editor, whereas a document-level score is not very useful for PE purposes.

Furthermore, it can be stated that the MTQE community looks at translation quality from a rather mathematical point of view and their quality criteria are based on probabilities, frequencies, and percentages. Even the score that they produce is a numeric one, at least at sentence and document-level. This stands to reason since computers and their programmes and tools are based on numbers and any input that is given to a computer must therefore be converted to a digital form. In the case of MT and translation quality, qualitative aspects have to be expressed in a quantifiable way. This has been done for many textual features, such as lexical adequacy that can be assessed by looking at the context of a word, by calculating the probabilities of its translation suggestions and by then choosing the most probable one. However, of the extra-textual criteria that have been pointed out, not a single one is considered in state-of-the-

art MTQE systems. Yet, based on the research conducted for this thesis, it is impossible to say whether the reason for this is that the MTQE community does not yet know how to include those criteria in their systems or whether they have not been aware of their importance yet.

After this analysis, an answer can now be given to the research question that was put forward in the beginning. The whole research project was conducted in order to find out to what extent the quality criteria used for developing quality estimation tools for machine translation are different to the quality criteria prevailing in translation studies. The answer that can be given to this question is that they are very different from each other and that the MTQE community can learn a lot from the translation studies. This confirms the assumptions that were expressed in the introduction. The chosen research method has shown to be a good one to approach the research question since it provided an insight into the fields of translation quality and MTQE from various different angles. The main difficulty that came up in the course of the research process, however, was to set a limit of how many publications to consider in this thesis and where to stop. Of course, there would have been much more papers, articles, and books that could have been interesting to incorporate, however, this would have gone beyond the scope requested for this thesis. Nevertheless, the most important works on the topic have been presented and the research question can be answered adequately.

12 Conclusions

To sum up, it can be stated that this thesis has shown that MTQE systems are able to demonstrate noticeable results in terms of coverage of quality assessment criteria when it comes to estimating the linguistic and semantic features of a text that has been translated by an MT system. The results are especially noticeable considering the relatively short time since MTQE tools have come into existence. On the other hand, it has been shown that there are still many essential factors that they cannot yet take into account but that are considered as core criteria in translation studies. The most important criterion that can be found in all the approaches taken from translation studies that have been presented in this thesis is the preservation of a text's function during the translation process. The translated text should evoke the same feelings and enable the same reaction in its readers as the original does in the source-language audience. In order to achieve the same function for the target-language audience, the correct linguistic, stylistic, and semantic features have to be chosen by the translator. Therefore, the quality of a translation can never be judged isolated from its context and function, however, this is what state-of-the-art MTQE systems do.

In the discussion, the question whether extra-linguistic criteria have not been incorporated in MTQE systems yet because of the MTQE community's unawareness of the issue or because of their inability to compute those features so far has been brought forward. If the first reason is true, future research should focus on finding a way to incorporate text-external quality features in MTQE systems. This is especially the case for the text function since it influences a great many of other factors, such as terminology, syntax, etc. In case that the second reason is true, awareness should be risen for this issue inside the MTQE community in order to initiate research on how to integrate extra-textual factors into MTQE systems.

Another possible subject for further research from the point of view of translation studies could be the empirical examination of the aspects and results that have been brought forward in this thesis. This could be done, for example, by performing an experiment in which MTQE output is compared to human judgement of the same MT segments and by asking the human evaluators for the quality criteria that influenced them in their judgement. The answers could then be compared to the quality features used in the building of the MTQE system that have been used in order to point out missing quality aspects.

13 Bibliography

Allen, Jeffrey (2003). Post-Editing. In: Somers, Harold L. (ed.) *Computers and Translation: A Translator's Guide*. Amsterdam/Philadelphia: John Benjamins, 297–317.

ALPAC (1966). Language and Computers in Translation and Linguistics. A Report by the Automatic Language Processing Advisory Committee. http://www.mt-archive.info/50/ALPAC-1966.pdf (Last access: 08.07.2020).

Ammann, Margret (1990). Anmerkungen zu einer Theorie der Übersetzungskritik und ihrer praktischen Anwendung. *TEXTconTEXT* 5, 209-250.

Arnold, Doug (2003). Why Translation Is Difficult for Computers. In: Somers, Harold L. (ed.) *Computers and Translation: A Translator's Guide*. Amsterdam/Philadelphia: John Benjamins, 119–142.

Bennett, Scott & Gerber, Laurie (2003). Inside Commercial Machine Translation. In: Somers, Harold L. (ed.) *Computers and Translation: A Translator's Guide*. Amsterdam/Philadelphia: John Benjamins, 175–190.

Beuth (2015). DIN EN ISO 9001:2015-11. https://www.beuth.de/de/norm/din-en-iso-9001/235671251 (Last access: 10.01.2021).

Beuth (2016). DIN EN ISO 17100:2016-05. https://www.beuth.de/de/norm/din-en-iso-17100/246682286 (Last access: 10.01.2021).

Biçici, Ergun (2019). RTM Stacking Results for Machine Translation Performance Prediction. *Proceedings of the Fourth Conference on Machine Translation (WMT)* 3, 73-77.

Biçici, Ergun & Way, Andy (2014). Referential Translation Machines for Predicting Translation Quality. *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 313-321.

Broeck, Raymond van den (1985). Second Thoughts on Translation Criticism. A Model of its Analytic Function. In: Hermans, Theo (ed.) *The Manipulation of Literature. Studies in Literary Translation.* London/Sydney: Croom Helm, 54-62.

Castilho, Sheila; Doherty, Stephen; Gaspari, Federico & Moorkens, Joss (2018). Approaches to Human and Machine Translation Quality Assessment. In: Castilho, Sheila; Doherty, Stephen; Gaspari, Federico & Moorkens, Joss (ed.) *Translation Quality Assessment: From Principles to Practice*. Cham: Springer, 9–38.

Chatzikoumi, Eirini (2020). How to Evaluate Machine Translation: A Review of Automated and Human Metrics. *Natural Language Engineering* 26, 137–161.

Doherty, Stephen (2017). Issues in Human and Automatic Translation Quality Assessment. In: Kenny, Dorothy (ed.) *Human Issues in Translation Technology*. London: Taylor and Francis, 131–148.

Forcada, Mikel L. (2010). Machine Translation Today. In: Gambier, Yves & van Doorslaer, Luc (ed.) *Handbook of Translation Studies*. Amsterdam/Philadelphia: John Benjamins, 215–223.

Göpferich, Susanne (2008). *Textproduktion im Zeitalter der Globalisierung. Entwicklung einer Didaktik des Wissenstransfers*. Tübingen: Stauffenburg.

Hart, Chris (1998). *Doing a Literature Review: Releasing the Social Science Research Imagination*. London: Sage.

House, Juliane (1997). *Translation quality assessment: a model revisited.* Tübingen: Narr.

Hui, Liu (2017). *A study on translation quality management of automative information based on SAE J2450 translation quality metric.* Dissertation, Universität Wien.

Hutchins, W. John & Somers, Harold L. (1992). *An Introduction to Machine Translation*. London: Academic Press.

Kaindl, Klaus (2003). Übersetzungskritik. In: Snell-Hornby, Mary; Hönig, Hans G.; Kußmaul, Paul & Schmitt, Peter A. (ed.) *Handbuch Translation.* Tübingen: Stauffenburg, 373-378.

Kenny, Dorothy (2019). Machine Translation. In: Rawling, Piers & Wilson, Philip (ed.) *The Routledge Handbook of Translation and Philosophy*. New York: Routledge, 428–445.

Kim, Hyun; Lee, Jong-Hyeok & Na, Seung-Hoon (2017). Predictor-estimator using Multi-level Task Learning with Stack Propagation for Neural Quality Estimation. *Proceedings of the Conference on Machine Translation (WMT)* 2, 562-568.

Koby, Geoffrey S.; Fields, Paul; Hague, Daryl; Lommel, Arle & Melby, Alan (2014). Defining Translation Quality. *Revista Tradumàtica: tecnologies de la traducció* 12, 413–420.

Kreutzer, Julia; Schamoni, Shigehiko & Riezler, Stefan (2015). QUality Estimation from ScraTCH (QUETCH): Deep Learning for Word-level Translation Quality Estimation. *Proceedings of the Tenth Workshop on Statistical Machine Translation,* 316-322.

Läubli, Samuel; Castilho, Sheila; Neubig, Graham; Sennrich, Rico; Shen, Qinlan & Toral, Antonio (2020). A Set of Recommendations for Assessing Human-Machine Parity in Language Translation. *Journal of Artificial Intelligence Research* 67, 653-672.

Lavie, Alon & Agarwal, Abhaya (2007). METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. *Proceedings of the workshop on Statistical Machine Translation,* 228-231.

Lommel, Arle (2018). Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies. In: Castilho, Sheila; Doherty, Stephen; Gaspari, Federico & Moorkens, Joss (ed.) *Translation Quality Assessment: From Principles to Practice*. Cham: Springer, 109–127.

Lommel, Arle; Burchardt, Aljoscha & Uszkoreit, Hans (2015). Multidimensional Quality Metrics (MQM) Definition. http://www.qt21.eu/mqm-definition/definition-2015-12-30.html (Last access: 03.11.2020).

Machi, Brenda T. & McEvoy, Lawrence A. (2012). *The Literature Review. Six Steps to Success*. Thousand Oaks: Corwin.

Martins, André F. T.; Junczys-Dowmunt, Marcin; Kepler, Fabio N.; Astudillo, Ramón; Hokamp, Chris & Grundkiewicz, Roman (2017). Pushing the Limits of Translation Quality Estimation. *Transactions of the Association for Computational Linguistics* 5, 205-218.

Mossop, Brian; Hong, Jungmin & Teixeira, Carlos (2020). *Revising and editing for translators*. London/New York: Routledge.

Nida, Eugene A. & Taber, Charles R. (1982). *The Theory and Practice of Translation.* Leiden: Brill.

Nießen, Sonja; Och, Franz Josef; Leusch, Gregor & Ney, Hermann (2000). An evaluation tool for machine translation: fast evaluation for MT research. *Proceedings of the second international conference on language resources and evaluation,* 39-45.

Nitzke, Jean (2019). *Problem Solving Activities in Post-Editing and Translation from Scratch. A Multi-Method Study*. Berlin: Language Science Press.

Nord, Christiane (2003). Das Verhältnis des Zieltexts zum Ausgangstext. In: Snell-Hornby, Mary; Hönig, Hans G.; Kußmaul, Paul & Schmitt, Peter A. (ed.) *Handbuch Translation.* Tübingen: Stauffenburg, 141-144.

O'Brien, Sharon (2011). Towards Predicting Post-Editing Productivity. *Machine Translation* 25 (3), 197–215.

Papineni, Kishore; Roukos, Salim; Ward, Todd & Zhu, Wei-Jing (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 311–318.

Parra Escartín, Carla; Béchara, Hanna & Orăsan, Constantin (2017). Questing for Quality Estimation. A User Study. *The Prague Bulletin of Mathematical Linguistics* 108, 343-354.

Popović, Maja (2018). Error Classification and Analysis for Machine Translation Quality Assessment. In: Castilho, Sheila; Doherty, Stephen; Gaspari, Federico & Moorkens, Joss (ed.) *Translation Quality Assessment: From Principles to Practice*. Cham: Springer, 129–158.

Reiss, Katharina (1971). *Möglichkeiten und Grenzen der Übersetzungskritik: Kategorien und Kriterien für eine sachgerechte Beurteilung von Übersetzungen.* München: Hueber.

Reiss, Katharina (2014). *Translation Criticism – The Potentials and Limitations. Categories and Criteria for Translation Quality Assessment.* Translated from German by Erroll F. Rhodes. London/New York: Routledge.

Reiss, Katharina & Vermeer, Hans J. (1984). *Grundlegung einer allgemeinen Translationstheorie.* Tübingen: Niemeyer.

Ridley, Diana (2008). *The Literature Review. A Step-by-Step Guide for Students*. Los Angeles: SAGE.

Scarton, Carolina; Beck, Daniel; Shah, Kashif; Smith, Karin Sim & Specia, Lucia (2016). Word embeddings and discourse information for Machine Translation Quality Estimation. *Proceedings of the First Conference on Machine Translation* 2, 831-837.

Schreiber, Michael (2003). Übersetzungstypen und Übersetzungsverfahren. In: Snell-Hornby, Mary; Hönig, Hans G.; Kußmaul, Paul & Schmitt, Peter A. (ed.) *Handbuch Translation.* Tübingen: Stauffenburg, 151-154.

Schütz, Jörg (1999). Deploying the SAE J2450 Translation Quality Metric in MT Projects. *MT Summit VII*, 278-284.

Snover, Matthew; Dorr, Bonnie; Schwartz, Richard; Micciulla, Linnea & Makhoul, John (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*, 223–231.

Specia, Lucia; Raj, Dhwaj & Turchi, Marco (2010). Machine Translation Evaluation versus Quality Estimation. *Machine Translation* 24, 39–50.

Specia, Lucia; Scarton, Carolina & Paetzold, Gustavo Henrique (2018). *Quality Estimation for Machine Translation* (Synthesis Lectures on Human Language Technologies 39). San Rafael: Morgan & Claypool.

Specia, Lucia & Shah, Kashif (2018). Machine Translation Quality Estimation: Applications and Future Perspectives. In: : Castilho, Sheila; Doherty, Stephen; Gaspari, Federico & Moorkens, Joss (ed.) *Translation Quality Assessment: From Principles to Practice*. Cham: Springer, 201–235.

TAUS (2010). MT Post-Editing Guidelines. https://www.taus.net/academy/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines (Last access: 07.11.2020).

Turchi, Marco; Negri, Matteo & Federico, Marcello (2015). MT Quality Estimation for Computer-assisted Translation: Does it Really Help?. *Proceedings of the 53rd Annual Meeting of*

the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), 530-535.

Turian, Joseph P.; Shen, Luke & Melamed, I. Dan (2006). *Evaluation of Machine Translation and Its Evaluation*. New York: New York University.

Vermeer, Hans J. & Witte, Heidrun (1990). *Mögen Sie Zistrosen? Scenes & frames & channels im translatorischen Handeln.* Heidelberg: Groos.

Weaver, Warren (1949). Translation. http://www.mt-archive.info/50/Weaver-1949.pdf (Last access: 08.07.2020).

Zhou, Ming; Wang, Bo; Liu, Shujie; Li, Mu; Zhang, Dongdong & Zhao, Tiejun (2008). Diagnostic Evaluation of Machine Translation Systems Using Automatically Constructed Linguistic Check-Points. *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, 1121–1128.

Abstract (German)

Die Fortschritte im Feld der maschinellen Übersetzung der vergangenen Jahre führten zu einem Anstieg in der kommerziellen Nutzung computerbasierter Übersetzungswerkzeuge. Um die erwartbare Qualität eines maschinellen Übersetzungssystems im Vorhinein einschätzen zu können, wurden sogenannte Machine Translation Quality Estimation-Tools (MTQE-Tools) entwickelt. Da in die Entwicklung dieser Tools jedoch meist keine Übersetzer*innen oder Translationswissenschaftler*innen involviert sind, stellt sich die Frage, inwiefern Qualitätskriterien, die in der Translationswissenschaft als wichtig erachtet werden, in MTQE-Tools Berücksichtigung finden. Die vorliegende Masterarbeit beschäftigt sich daher mit den Qualitätskriterien, die im Aufbau der wichtigsten MTQE-Tools angesetzt wurden, und vergleicht diese mit den von der Translationswissenschaft angenommenen Kriterien. Dazu wird eine umfangreiche Literaturanalyse durchgeführt, deren Ergebnisse im Rahmen einer Vergleichsanalyse gegenübergestellt, verglichen und interpretiert werden. Die Ergebnisse zeigen, dass auf der textuellen Ebene ähnliche Kriterien zur Anwendung kommen, besonders auf der kontextuellen Ebene aber translationswissenschaftlich abgesicherte Elemente von Translationsqualität von den MTQE-Tools nicht berücksichtigt werden.

Abstract (English)

The improvements made in machine translation in the past decade have led to an increased commercial use of machine translation tools. In order to be able to give a prediction on how reliable a machine translation system is, quality estimation tools for machine translation (MTQE tools) have been developed. However, since usually no translators or translation scholars are involved in the creation of MTQE tools, the question comes up to what extent quality criteria prevailing in translation studies have been incorporated into those tools. This Master's thesis takes a close look at the quality criteria that have been used in the creation of the most relevant quality estimation tools to date and compares them with the quality criteria that prevail in translation studies. For this purpose, an extensive literature review is conducted, and the results are compared and interpreted in a comparative analysis. The results show that whereas similar criteria are used at the textual level, extra-textual quality features prevalent in translation studies are not (yet) part of MTQE tools.