



universität
wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„TQA done by novice translators and MA students from
the University of Vienna's Centre for Translation Studies
– A replication study using the TQA model
by Colina (2008)“

verfasst von / submitted by

Sonja Hengsberger, BA

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of
Master of Arts (MA)

Wien, 2022 / Vienna 2022

Studienkennzahl lt. Studienblatt /
degree programme code as it appears on
the student record sheet:

UA 070 331 342

Studienrichtung lt. Studienblatt /
degree programme as it appears on
the student record sheet:

Masterstudium Translation Deutsch Englisch

Betreut von / Supervisor:

Univ.-Prof. Dragoş Ioan Ciobanu, PhD

Statement of authorship – Selbstständigkeitserklärung

Ich versichere, die vorliegende Arbeit selbstständig verfasst zu haben. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt. Alle von mir für direkte und indirekte Zitate benutzten Quellen sind nach den Regeln des wissenschaftlichen Zitierens angegeben. Mir ist bekannt, dass beim Verstoß gegen diese Regeln eine positive Beurteilung der Arbeit nicht möglich ist. Ich habe die Arbeit bzw. Teile davon weder im In- noch im Ausland zur Begutachtung als Prüfungsarbeit vorgelegt.¹

Wien, im Oktober 2022



¹ In English, this paragraph reads approximately: “I assure that I have written the work in hand by myself. I have not used any sources or resources other than the ones listed in the bibliography. All sources used for direct or indirect citations are listed according to the rules of good citation practice. I acknowledge that in case of infringement of these rules, positive grading of the work in hand will not be possible. I have not submitted the work in hand nor parts thereof for grading either in Austria or abroad.” (translation S.H.)

Acknowledgements

Above all, I would like to express my gratitude to Professor Ciobanu for having been such an excellent supervisor. Not only did he show great patience, but also offered kind and very constructive feedback and was always willing to answer my many questions. Thank you!

My gratitude also goes to Professor Sonia Colina at the University of Arizona, who was kind enough to answer my e-mails and give me information on her work that was crucial for this MA project. I am very grateful!

I would further like to thank all the participants who replied to my announcements and graciously volunteered to devote half a day to helping me with my MA project. All of them showed great enthusiasm and made many interesting comments. Thank you for your help!

I also wish to extend my heartfelt thanks to Katharina Schuhmann, a friend and colleague with whom I have worked together throughout most of my studies and who was so kind as to offer a sounding board for my earliest ideas as well as do the proofreading for my MA project. I really could not have done it without her.

Last, but not least, I am also deeply grateful to my parents, who offered unfailing support over the course of all my studies. Thank you most of all!

Table of contents

List of tables	6
Introduction	7
1 Definition of <i>quality</i> and other relevant terms	11
2 Literature review: The evaluation of translation quality	20
2.1 Translation Quality Evaluation: Cornerstones, types and methods	20
2.2 TQA models and approaches	29
2.2.1 Experience-based and anecdotal approaches	30
2.2.2 Theoretical and research-based approaches	31
2.2.3 Experience-based vs. research-based approaches	37
2.2.4 Reader-response approaches	39
2.2.5 Equivalence-based approaches	40
2.2.6 Functional, textual and pragmatic approaches	44
3 TQA tool by Sonia Colina (2008)	50
3.1 Rationale of creation	50
3.2 Cornerstones and characteristics	52
3.3 Structure: Evaluation criteria	56
3.4 Empirical testing: Colina's (2008) original study	58
3.4.1 Participants	58
3.4.2 Procedure	60
3.4.3 Results and limitations	61
4 Replication study	65
4.1 Cornerstones	65
4.2 Texts	66
4.3 Brief	68
4.4 Participants	70
4.5 Procedure	74

5	Results and discussion	77
5.1	Limitations.....	77
5.2	RQ-1: Can Austrian MA-level translation studies students and recently graduated former students (novice translators) use the TQA model by Colina (after receiving training and materials) without encountering problems Colina's more experienced evaluators did not have?.....	78
5.3	RQ-2: How does the productivity (measured as rating time in minutes) of MA students and novice translators compare to Colina's groups of evaluators?	87
5.4	RQ-3: What is the attitude of MA students and novice translators towards the use of Colina's model in professional practice?.....	98
6	Conclusion	105
	Bibliography	109
	Primary sources	109
	Secondary sources	110
	Appendix I: Comparison between TQA approaches mentioned in Kingscott (2007) and Garvin (1984).....	120
	Appendix II: Disadvantages of previous TQA models overcome (or not) by Colina's (2008) model.....	122
	Appendix III: Texts	128
	Appendix IV: Correspondence.....	140
	Appendix V: Checklist.....	145
	Appendix VI: Questionnaire	148
	Appendix VII: Comparison of original study and replication study	152
	Appendix VIII: Abstracts.....	154
	English abstract.....	154
	German abstract.....	154

List of tables

Table 1: Participants' confidence in scores (partially adapted from Colina (2008: 121)).....	79
Table 2: Evaluation time per text and participant (in minutes) as well as productivity in words per hour (indicated between brackets) in the replication study.....	87
Table 3: Scores given by participants in each category for each text as well as total scores per text and per participant in the replication study	89
Table 4: Average evaluation time per text (in minutes) and per ten words (in seconds) as well as average productivity expressed as words per hour, relative to source text length in the replication study	91
Table 5: Average evaluation time for the Chinese language per text (in minutes) and per ten words (in seconds), relative to source text length in the original study (adapted from Colina (2008: 118)).....	92
Table 6: Average evaluation time for the Russian language per text (in minutes) and per ten words (in seconds), relative to source text length in the original study (adapted from Colina (2008: 118)).....	92
Table 7: Average evaluation time for the Spanish language per text (in minutes) and per ten words (in seconds), relative to source text length in the original study (adapted from Colina (2008: 120)).....	92
Table 8: Average rating time per ten words (in seconds) of professional translators in the original study compared to that of students and novice translators in the replication study....	94
Table 9: Average rating time per ten words (in seconds) of language teachers in the original study compared to that of students and novice translators in the replication study.	95
Table 10: Start and finish time of each participant for each text (given in 24-hours format).	96
Table 11: Number, length and placement of breaks between texts for each participant (in minutes).....	97

Introduction

The subject of this Master's thesis is the assessment of translation quality and, in particular, the replication of a translation quality assessment (TQA) study by Sonia Colina (2008), using a specific TQA model proposed by Colina (2008; 2009).

Quality is a topic that has always been important in Translation Studies and is still important today, even if the nature of people's interest in it has changed over time. Where the importance of quality was once mainly religious and political, nowadays it is mostly economic, legal and professional (Williams 2001: 327). In fact, over the last two decades, the translation services industry has become increasingly focused on quality and quality assessment (QA), due in part to the numerous national and international standards that have been published (Drugan 2013: 74f.), but also due to the growing volume of texts to be translated and the increasing competition between translation providers (Secară 2005: 39).

The need for the assessment of translation quality is manifold and applies to the translation industry itself as much as to clients and translation practitioners (Kingscott 2007: 317). The industry has a need to establish credibility vis-à-vis its customers and an interest in taking steps towards ensuring that translators do not provide substandard services and thus tarnish the industry's reputation. Customers need a means of ascertaining that they receive a quality product, since they wish to avoid the consequences of faulty translations (reputational damage, liability, additional costs etc.) (Lommel et al. 2014: 456). Therefore, clients expect practitioners to be able to evaluate and defend the quality of their work (Drugan 2013: 69). Professional translators themselves have a keen interest in TQA, since it enables them to protect themselves in case of litigation by proving, through the use of TQA, that due diligence was exercised and acceptance criteria were met. It also enables them to differentiate themselves from amateur translators and allows them to justify rates (Drugan 2013: 41; Kingscott 2007: 317; Lommel et al. 2014: 456; Secară 2005: 41).

Consequences that may arise from defective translations can be financial, legal or even life-threatening in nature (one need only think of ambiguously translated legal texts that lend themselves to unintended interpretation or of mistranslations in the medical sector) (Byrne 2007: 2, 5f.). For translators in particular, this can mean liability for breach of contract, for commission of a tort (i.e. a civil wrongdoing that causes damage to another), or for violating the translator's duty of care (Byrne 2007: 10f.). Of course, professional translators' associations like Universitas Austria or Germany's federal association of interpreters and translators offer liability insurance to protect their members from the consequences of

(perceived) substandard translations (BDÜ 2021; Universitas Austria 2017). It certainly appears advantageous, however, for practitioners to prevent defective translations from reaching clients altogether by discovering their lack of quality before submitting them. This is where TQA comes into play.

From what has been outlined above, it becomes clear that TQA can help professional translators protect themselves (at least in part) from the risk of legal consequences that they incur with their work, but they are not the only ones facing this risk. Since translation is a so-called unregulated profession in Austria (Universitas Austria 2019), i.e. a profession that does not require the person exercising it to provide any kind of qualification (BMDW 2021), students can already work as translators while following courses at university. This means that they are facing the same risks as professional translators, and thus TQA becomes of interest to them as well. In the field of TQA, students are often the ones whose translations are being assessed (e.g. Bittner 2020; Colina 2003; Hagemann 2007), but it is hardly ever the other way around. For my Master's thesis, students and recently graduated former students (called novice translators) were the evaluators, meaning they were the ones carrying out the translation quality assessment.

However, selecting a single TQA framework is a challenging task because of the great variety of models that allow one to assess the quality of a translation (see e.g. Bittner 2020: 21-64). In the present study, the TQA model by Sonia Colina (described in Colina (2008) and Colina (2009)) was used. Colina's model has not been given much attention in TQA literature. In fact, it is hardly ever mentioned in works on TQA published after 2009, neither in the positive nor in the negative sense². Drugan (2013: 46) even goes so far as to claim that only four detailed, reproducible TQA models are available, completely overlooking the one proposed by Colina. Since the model does not seem to have been paid much attention in academic circles, this MA project may serve to provide some more data on its use, and particularly on its use by students and novice translators. Colina's TQA model was chosen for this MA project, because it is functionalist and componential (i.e. it recognizes different aspects of quality), adaptable and tested (Colina 2008; 2009) and because the author provides both training materials (Hablamos Juntos 2009) and examples regarding the use of the model (Colina 2015).

² One notable exception to this is Hague et al. (2011: 256), who find Colina's TQA model "convincing[...]", and another is O'Brien (2012: 65), where Colina's work is wrongly cited as belonging to the educational context (which is true for Colina (2003), but not for Colina (2008; 2009)), leading to the assumption that this was a mistake.

Furthermore, Colina's model conforms to the criteria formulated by Martínez Melis & Hurtado Albir (2001: 283):

In our view, translation assessment should: 1) Use objective criteria which define error types (scales); 2) Establish the seriousness of the error on the basis of functionalist criteria [...] without ascribing fixed coefficients to the errors; 3) Also take into account the good solutions in the translation; 4) Adopt a flexible view of assessment, allowing partial assessments to be carried out as necessary.³

As will be discussed further in chapter 3, the TQA model by Colina (2008; 2009) does all of the things called for above, except that it does not define error types, since the model is based on descriptors instead of error counting (Colina 2008: 104, 106).

In order to provide further data on the use of Colina's little-mentioned model and to contribute data concerning the use of TQA models by MA-level students and novice translators, this MA project replicated the study reported on in Colina (2008), which originally served to empirically test the use of Colina's TQA model by bilingual evaluators, professional translators and language teachers. In light of all that has been said above, the thesis aimed to answer the following research questions (RQs):

1. Can Austrian MA-level translation studies students and recently graduated former students (novice translators) use the TQA model by Colina (after receiving training and materials) without encountering problems Colina's more experienced evaluators did not have?
2. How does the productivity (measured as rating time in minutes) of MA students and novice translators compare to Colina's groups of evaluators?
3. What is the attitude of MA students and novice translators towards the use of Colina's model in professional practice?

Regarding these RQs, the following hypotheses (H) were formulated. With regards to RQ-1, it was expected that students and novice translators would be able to use Colina's TQA model (H-1.1), since they received training and materials and, after her first study in 2008, the author drew the tentative conclusion that no specialized knowledge of translation theory seemed to be necessary in order to successfully use the model (Colina 2008: 121, 123). Still with regard to RQ-1, it was hypothesized that participants would encounter some problems that more experienced raters did not have, especially where the matching of descriptive statements to the evaluated text are concerned (H-1.2), since that was a point raised in Colina's study. It

³ Note that here, *partial assessment* does not refer to assessing only parts of the text (as distinguished in footnote 16), but rather to an assessment focusing on specific aspects or parts of quality, without taking other parts into account (Martínez Melis & Hurtado Albir 2001: 283).

should be noted, however, that Colina's participants did not seem to consider it a problem that impeded or seriously hampered their use of the model (Colina 2008: 123).

Concerning RQ-2, it was expected that students and novice translators would rate more slowly than Colina's professional translators did (H-2.1), due to their lack of experience. As concerns RQ-2, it was also expected that students and novice translators would rate as fast or even faster than language teachers in Colina's study did (H-2.2), since language teachers tended to rate more slowly than professional translators (Colina 2008: 118). The author hypothesized in her later study (Colina 2009: 255) that this might be because teachers spend more time reflecting on their evaluation, which would make them better raters when feedback is required. Since translation studies students and graduates are not teachers (who presumably are often required to give feedback), they should therefore rate somewhat faster (or just as fast, owing to their lack of experience).

Relating to RQ-3, it was assumed that participants might consider the use of Colina's model in professional practice, yet remain somewhat skeptical (H-3.1), since they lacked experience with which to judge its actual use in practice (either in the negative or in the positive sense). It was further assumed that students and novice translators would recognize the need to assess the quality of either their own translations or of colleagues' translations in practice as well as the need to use a structured model to do so (H-3.2), since they would presumably be aware of the fact that quality is of great importance in the industry (Drugan 2013: 69, 74f.). Lastly, it was hypothesized that participants would see the functional, descriptive nature of the model in a generally positive way (H-3.3), since at least some professional translators in Colina's study did the same, saying that it allowed them to evaluate important aspects of quality that they had not known how to capture before (Colina 2008: 124f.).

Now that the replication study and the research questions have been outlined, some vital terms such as *quality* and *assessment* will be defined in the following chapter.

1 Definition of *quality* and other relevant terms

Before discussing the subject of translation quality evaluation in detail, it is necessary, as Bittner (2020: 50) points out, to first define its constituent concepts: the terms *translation*, *quality* and *evaluation* (and/or *assessment*). Therefore, this chapter will show which different perspectives quality can be regarded from as well as the terminological difficulty surrounding the terms *evaluation* and *assessment*. However, before either of these latter concepts can be discussed, *translation* must be defined first, since every attempt at translation quality evaluation is inherently linked to one's understanding of translation (House 2016: 67). Colina (2015: [230]; 2020: 458) points out that this need not be an explicit theory (as House implies) but can also be an implicitly stated view about translation. She agrees, however, that translation should always be defined first.

A basic problem with the term *translation* is that, in English as well as other languages, it can mean both the process of translating a source text into a target text and the result of this translation process, that is to say the translated text, also called the translation product (Karoubi 2016: 256; Larose 1998: 167). Since in Colina's (2008) study, translations were assessed as products (i.e. the translated texts were evaluated), the same was therefore done in the replication study. *Translation* for the purpose of this thesis thus means translated texts as finished products, resulting from the process which will be called *translating* or *translation process*.

However, even the definition of *translation* as a product is not clear: some scholars strictly distinguish between translations and different forms of versions and adaptations (e.g. House 2016: 78), while others accept any text presented for assessment under the label of a translation as such (Karoubi 2016: 257). For the purpose of the thesis, the following definition of translation as a product, i.e. as a translated text, given in Colina (2012) (and almost verbatim in Colina (2020: 459)) within the context of TQA will be adopted:

A translated text, also known as the *target text* [TT], is defined here as a text that fulfills or attempts to fulfill a certain function in a target culture (in accordance with a set of explicit or implicit instructions, known as the *translation brief*) and that bears a translation relationship to another text in another language, the *source text* (ST); the specifics of a translation relationship (versus version, adaptation, etc.) can vary from one culture to another. (Colina 2012: [1], emphasis in the original)

Colina also makes it clear that translation as a product includes both the text and its extra-linguistic context (e.g. audience, purpose, medium) (Colina 2012: [2]) and that a translation always results from the interaction of social participants (like the writer, the translator, the target audience) and a purpose (Colina 2008: 99).

Now that the term *translation* has been defined for the purpose of this thesis, the first constituent concept of QA, *quality*, can be discussed. *Translation quality* is a term that has been much debated in Translation Studies for decades (see i.a. House 1997; Brunette 2000; Lauscher 2000; Colina 2003; Karoubi 2016: 257f.), but has so far proven elusive when it comes to definitions and standards (Bittner 2020: 50; Colina 2009: 236). Although most people have an intuitive understanding of what (translation) quality means (Harvey & Green 1993), only the quality of translation as a service has been benchmarked in the form of certain national or international standards (Gouadec 2010: 270; Thelen 2009: 203).

Examples of these standards include DIN 2345 (previously subsumed into EN 15038 (Drugan 2013: 74), which is now superseded by ISO 17100 (Schnierer 2019: 19f.)), ÖNORM D 1200 and 1201, and (in part) ISO 9000 (a standard relating mainly to business processes) (Budin 2007: 55ff.; Colina 2008: 98f.), ISO 17100 (Schnierer 2019: 44ff.) or ASTM F2575 (regarding translation projects) (Colina 2008: 98f.). None of these standards specify features of the finished translation product but rather apply to processes that (supposedly) lead to the production of a quality product (Budin 2007: 55f.; CEN s.a.: [2]; Drugan 2013: 74f.). Quality of the process, however, does not necessarily ensure quality in the finished product (Gouadec 2010: 271), and for translation product quality⁴ there are no common standards, neither within the translation industry nor within academic circles (Colina 2015: [240]). In fact, “there are perhaps as many ideas about what is *good quality* translation as there are persons involved in translation” (Thelen 2009: 204, emphasis in the original), and it is this lack of standardization of translation quality that engenders great variety when it comes to TQA (Colina 2020: 458).

In part, the difficulty in defining (translation) quality stems, once more, from a double meaning of the word: from a descriptive point of view, *quality* means a feature merely distinguishing one thing from another, whereas from a prescriptive view point, *quality* means a judgement and an attribution of value, as in ‘good’ quality or ‘bad’ quality (Bittner 2020: 51, 54; Chesterman 2016: 117).

Furthermore, views on (translation) quality depend on the user of the word and his or her situation (Gouadec 2010: 270; Harvey & Green 1993); quality is thus, like translation, context-dependent (Bittner 2020: 51; Secară 2005: 39). Gouadec (2010: 272) even goes so far as to say that the quality of a translation product is extrinsic (relating to requirements of audience, purpose, medium and other external parameters) as well as intrinsic. In addition to

⁴ This is perhaps best exemplified by the fact that the ISO 17100 standard, which takes the product-quality-through-process-quality approach, does not even define translation quality as a term (although many other related terms are defined) (ISO 2015a).

that, ideas about translation quality may change over time (Maier 2000: 140). This has, in fact, already become apparent in the notion of equivalence, which used to be the benchmark for almost all evaluations of translations, but in recent years has been regarded as more and more controversial, restrictive and unsuitable for TQA (Colina 2009: 238; Colina et al. 2017: 274; Lauscher 2000: 151, 158).

Yet another reason why *translation quality* is hard to define is a notable difference in focus between the translation industry and the academic world: according to Drugan (2013: 77), translation quality in the industry is not necessarily seen as a variable that can be measured and assessed only in a text, as most academic models presume, but rather as a continuous process that may begin well before the translation brief is ever received.

Due to the elusiveness of the concept shown above, no attempt will be made to represent all possible notions of quality in this chapter, since this would go beyond the scope of this MA project. Only some selected approaches to quality will be presented in brief, before (*translation*) *quality* will be defined for the purpose of the thesis.

The concept of quality can be approached from several different points of view, as Harvey & Green (1993) show. They distinguish between five different ways of seeing quality⁵: quality as exception, quality as perfection, quality as fitness for purpose, quality as value for money and quality as transformative. Similar categories of quality have been established by Garvin (1984), and Fields et al. (2014: 407) attest that these match the ideas of many notable scholars in the quality management discipline. For purposes of comparison, (near-)matching categories⁶ by Harvey & Green (1993) and Garvin (1984), respectively, will be given next to each other in the following discussion.

The view of **quality as exceptional** can take three different forms. Firstly, there is the notion of quality as distinction, which neither defines what quality is nor offers any benchmark with which to compare quality. It simply constitutes an intuitive understanding that something is so distinguished that it must be of high or even the highest quality (Harvey & Green 1993). This non-objective approach supposes that quality is universally recognizable and is therefore called the transcendent approach by Garvin (1984: 25; Kurz 2020: 40). In the context of translation, Nishino (2020: 156) (based on Fields et al. (2014: 406)) adds that this

⁵ The article by Harvey & Green (1993) discusses quality in higher education, but the categories established there can be useful for a translation context as well, as will be demonstrated in the following paragraphs. The article also does not feature pagination, being an HTML-document, therefore no page references can be given.

⁶ Concerning the overlaps and differences in notions of quality, Nishino (2020: 158f.) points out that approaches can also be combined and are often interrelated in any case.

approach may be adopted by specialists who have developed an intuitive understanding of what is good or bad quality through experience.

Secondly, there is the idea of quality as exceeding standards. Here, criteria for quality are defined, but in such a way that these criteria are all but unfulfillable under normal circumstances, meaning that the very best of everything is required in order to achieve a level of excellence equated with quality. This of course makes this view of quality as exceeding (the highest) standards a very elitist notion. The third notion of quality as exceptional is a more lenient version of the second variation in that it sets attainable standards. It is an inclusive idea of quality, because any product that conforms to these standards can be said to be of quality (Harvey & Green 1993); Garvin therefore calls it the product-based approach (Garvin 1984: 25f.; Kurz 2020: 40). Since it consists of measuring product characteristics that are equated with quality, this last approach is often found in translation, where aspects like accuracy and fluency are measured in a text (Fields et al. 2014: 408).

The second category established by Harvey & Green (1993) is that of **quality as perfection or consistency**. This view is focused on processes, since processes are meant to ensure ‘zero defects’ in the end product by assuring that specifications are continuously met. Called the manufacturing-based approach by Garvin (1984: 27f.; Kurz 2020: 41) and renamed the production-based approach by Fields et al. (2014: 406) to better match the translation context, this idea of quality is reflected in above-mentioned industry norms such as EN 15038 and ISO 17100 (Fields et al. 2014: 410).

Harvey & Green’s (1993) third category, **quality as value (for money)**, is also found in Garvin (1984: 28; there called the value-based approach) and is centered around the cost-benefit ratio of a product. It is therefore difficult to measure (Kurz 2020: 41). Nishino (2020: 157) points out that in Translation Studies, this approach is often adopted when talking about machine translation (e.g. “Taking the cost into account, the result of machine translation is not bad” (Nishino 2020: 157).⁷). It is also a perspective often adopted by end users (of a translation) who are interested in costs above most other aspects (Fields et al. 2014: 410).

⁷ The perspective taken by Nishino (2020: 157) is of course based on the assumption that machine translation is cheap and therefore yields results whose quality reflects the low production costs; however, while this might sometimes be the case, machine translation can, in fact, be quite costly when training of engine and acquisition of training data are taken into account. A survey by Do Campo Bayón (2020), for example, showed that at least 500,000-1,000,000 segment pairs are necessary to adequately train an MT engine. A quick search of the TAUS data marketplace (TAUS 2021) reveals that word prices of €0.013 for training data in some language pairs (Japanese-Arabic, in this case) are not so unusual, which amounts to €6.82 for just 44 segments (extrapolated that would mean €77,500 for 500,000 segments).

Harvey & Green (1993) establish a fourth category, **quality as transformation**, under which quality is seen as added value and qualitative change. There is no matching category in Garvin (1984), but in the translation industry this view of ‘enhancing’ might be applicable to transcreation (i.e. the translation and redesign of marketing materials (Risku et al. 2017: 54)).

The last and for the thesis the most important category by Harvey & Green (1993) is **quality as fitness for purpose**. Under this viewpoint, quality is defined in a functional way rather than as something exceptional, and quality becomes measurable if requirements and purpose are stated clearly (Harvey & Green 1993)⁸. Quality as fitness for purpose is an inclusive definition of the term, since “[e]very product and service has the potential to fits [sic] its purpose and thus be a quality product or service” (Harvey & Green 1993). As this statement already indicates, quality under the ‘meeting requirements’ approach is measured in the product and not in the process, so that it is well-suited for a replication study examining translation products.

In line with the five categories of quality given above, Kingscott (2007) mentions five approaches to TQA that will briefly be outlined here. For a more detailed comparison between the approaches mentioned in Kingscott (2007) and those set out by Garvin (1984) as well as a more exhaustive list of the shortcomings of each approach see Appendix I.

The first of the five approaches Kingscott (2007) discusses is the **subjective evaluation of a translation by a translator** who is more experienced than the translator of the text in question and gives his or her expert opinion (Kingscott 2007: 318f.) (reminiscent of Garvin’s (1984) transcendent approach). Secondly, the so-called **qualifications of performer approach** (Kingscott 2007: 319f.), is based on stipulating the competences of the translator (e.g. training, language proficiency etc.). It partly matches Garvin’s (1984) manufacturing-based approach as it follows the same idea that the product will be of high quality if the ‘materials’ (in this case the translator) and procedures are. Thirdly, the **application of constrained procedures** (for the selection of translators, for translation itself etc.) as exemplified in ISO 17100, is an approach which can, however, only ensure consistency, not quality (Kingscott 2007: 320f.). As a fourth approach, Kingscott also mentions the **application of the so-called theoretical translation model**. The theoretical model, according to Kingscott (2007: 321), consists of two elements: comprehension of intent (by the

⁸ Garvin mentions a partially matching category here, the user-based approach, but describes it as the highly subjective view of the customer/user as to whether the product matches his or her needs and expectations (Garvin 1984: 27; Kurz 2020: 40). From a translation perspective, this is not really helpful, since users need translation precisely because they do not speak the source language, which they would need to understand in order to judge all aspects of quality (Fields et al. 2014: 409).

translator) and equivalence of response (on the end user-side). This fourth approach thus evaluates those two parameters, partially matching Garvin's (1984) product-based and user-based approaches.

Lastly, and again most importantly, there is the approach that Kingscott (2007: 321) calls **compliance with specification**, or in other words, **fitness for purpose**. Kingscott (2007: 321f.) strongly recommends this last approach due to various shortcomings of the other approaches (e.g. ever more experienced translators are hard to find; quality of process and competence of translator do not guarantee a quality product etc.) Other scholars, such as Kurz (2020: 43), also recommend the fitness-for-purpose approach for TQA or, as Kingscott (2007: 322) calls it, the principle of PEXIS (Purpose-oriented Explicit or Implicit Specification) as a yardstick for TQA. O'Brien (2012: 56) also mentions that industry players increasingly recognize a fitness-for-purpose view of quality and grow dissatisfied with existing one-size-fits-all assessment approaches.

Therefore, and because the definition of *quality* as fitness-for-purpose resonates well with a functionalist TQA model (as the one in Colina 2008), the notion of evaluating translations against the yardstick of fitness-for-purpose and against specifications or criteria will be examined in more detail.

The notion of quality as defined by (functionalist) criteria already appears in 1983 in Sager (1983: 121), who states that "[t]here are no absolute standards of translation quality but only more or less appropriate translations for the purpose for which they are intended." This idea is reinforced with the publication of Reiß and Vermeer's *Skopos* theory, which uses the purpose of a translation as the yardstick for evaluating its quality (Reiß & Vermeer 1984: 100), an idea subsequently to be found also in Larose (1998).

In recent years, more and more scholars have evaluated translations against requirements or specifications provided by the customer or against what Karoubi (2016: 261) calls the "assessment *skopos*" (my emphasis) contained therein. A similar trend can be observed among industry players, who increasingly take a customer-defined view of quality (Koehn 2020: 42; O'Brien 2012: 56). Based on the works of scholars like Colina (2008) and Angelelli (2009), Hague et al. (2011: 258) thus conclude that the translation industry and academia both increasingly recognize the need for specifications in TQA and that academic tendencies seem to converge in that direction. Examples include the works of Lauscher (2000) and Rosenmund (2001); the work of Hague et al. (2011) (who propose a specifications-based approach to TQA); the definitions of quality given in ISO 9000 as "degree to which a set of inherent **characteristics** [...] of an **object** [...] fulfils **requirements**" (ISO 2015b: [3.6.2], emphasis in

the original); the MQM-DQF quality matrix (see chapter 2.2.2, whose definition of “quality translation” mentions compliance with negotiated specifications (Dalla-Zuanna 2018: 11; Lommel et al. 2015)); and, not least of all, Kingscott’s above-mentioned (2007: 323) recommendation.

The JTF (Japan Translation Federation) also takes a specifications-based stance, with quality being defined as “the degree to which a translation product fulfils specifications agreed by the parties” (Nishino 2020: 159, translated from Japanese). The above definition thus touches upon another notion, namely that of specifications (and therefore quality, as measured in a functionalist approach) being subject to negotiation and consensus, something that is also mentioned by Lauscher (2000: 149). This reflects the fact that relatively recent TQA proposals, such as seen in Lauscher (2000) and Colina (2008; 2009), allow for a case-specific, customer-defined and criteria-based definition of *quality* that is adaptable to various situations and purposes (Colina 2011: 47).

In light of all that has been set out above, the following definition of *quality* drawn in part from Karoubi (2016: 261) and in part from DIN EN ISO 9000 (ISO 2015b: [3.6.2]; ISO 2015b: [3.6.4]) will be adopted for the purpose of the thesis: *quality* is the extent to which the features of the translated text under evaluation meet the stated and implied requirements for the fulfillment of the assessment *skopos*.

Before discussing different methods and approaches to TQA in the next chapter, the last constituent term of TQA, i.e. *evaluation* and/or *assessment*, must be defined first. Following that, the terminological difficulty surrounding translation (quality) evaluation/assessment itself will be briefly touched upon. When reviewing relevant works regarding the terms *translation evaluation* and *translation assessment*, it becomes immediately clear that these terms are often used interchangeably or as synonyms. This is done frequently without the author stating his or her definition and use of those terms⁹, an issue that tends to engender confusion and has thus been brought up by scholars like Arango-Keeth & Koby (2003: 119), Brunette (2000: 169), Karoubi (2016: 254) and Maier (2000: 137).

Of course, some attempts at distinction between the terms have been made. Arango-Keeth & Koby (2003: 119), for example, use the term *evaluation* to indicate that translation competence is evaluated (for a didactic purpose) rather than the suitability of a translation

⁹ There are, nevertheless, some notable exceptions to this. Colina (2011: 43; 2012: [1]) and Lauscher (2000: 150), for example, explicitly state that they are using both terms as synonyms, since this seems to be common practice in Translation Studies.

product, the latter of which they call *assessment*. Kiraly (2000: 140), on the other hand, uses *evaluation* to refer to the interpretation of information gathered about students' translation competence in a collection process that he calls *assessment*. The list of diverging definitions goes on (e.g. McAlester (1999: 169) or Colina (2011: 43)) and clearly shows that no definitive consensus has been reached regarding the use of *translation quality evaluation* (TQE) versus *translation quality assessment* (TQA).

Adding to this terminological confusion, TQA (translation quality assessment) shares the same acronym¹⁰ as translation quality assurance (Drugan 2013: 36). In fact, these two terms constitute an additional source of confusion, since TQA (translation quality assessment) in the translation industry is often confused with TQA (translation quality assurance) in the localization industry, especially because of the widespread adoption of parts of the (now defunct) LISA Quality Assurance model, "which actually provided a model for quality assessment" (Lommel et al. 2015).

For the purpose of this thesis, I will clearly state here that the terms *translation quality assessment* (TQA) and *translation quality evaluation* (TQE) have been used interchangeably, since the main part of the work in hand is a replication of a study by Colina (2008), who also used them as synonyms.

Now that all constituent terms have been discussed and defined, some attention should also be given to the term *translation quality assessment/evaluation* as a whole. Translation quality evaluation is, naturally, a kind of evaluation (Williams 2009: 4), but one that can be seen from different perspectives. James Holmes (1994: 78) defines translation evaluation as part of translation criticism, a branch of applied translation studies, and Karoubi (2016: 253) describes TQA as a complex decision-making process. Yet at its core, assessment is about distinguishing acceptable translations from unacceptable translations (Bittner 2020: 2). It therefore means measuring quality and judging quality (Martínez Melis & Hurtado Albir 2001: 275), by giving a (numerical) rating or by making some kind of value judgement (Kockaert & Segers 2017: 148) or by assigning an evaluative attribute like 'pass/success' or 'fail/failure' (Colina 2012: [1]). Some scholars see TQA as comparing the target text (TT) under evaluation with an ideal (hypothetical) TT (Lauscher 2000: 162) or as comparing translation solutions in the TT to potentially better translation solutions (Bittner 2020: 56),

¹⁰ Note, however, that in the context of the thesis, the abbreviation TQA only refers to translation quality assessment. It does not mean translation quality assurance, since quality assurance refers to the process of providing assurance about the quality of a product to the customer (through continuous controls and error prevention), not to the assessment of the quality of a finished translation product (ISO 2015b: [3.3.6]; Lommel et al. 2015; Williams 2004: 163).

while yet other scholars turn to definitions of assessment from outside the field of Translation Studies in order to define TQA (Karoubi 2016: 256).

For the purpose of the thesis, a definition of TQA has been synthesized based on Colina's view of translation and the process of TQA set out for her study (Colina 2008; see also the definition of *translation* above) as well as the steps of the TQA process set out by Bittner (2020: 70f.): reading of TT, comparison with ST and evaluation criteria, counting and weighting information thus obtained, marking achievement.

The definition of TQA thus reads: TQA is the process of one or more evaluators reading a translated text (target text) and comparing this target text with the source text to which it bears a translation relationship, in order to assess whether and to what degree the target text fulfills its purpose as set out in the translation brief and to interpret the information gained in this manner so that a judgement can be given.

Although the evaluator is the only person explicitly mentioned in the above definition and plays a key role by carrying out the assessment, using the assessment tool and, perhaps most importantly, interpreting any findings and giving a judgement (Karoubi 2016: 275), it should at this point be noted that there are also other participants in the TQA process. Those other participants are, minimally: the client and/or commissioner of the assessment (who plays a vital role in providing the translation brief on which the assessment is based); and the (presumed) target audience of the translation (Karoubi 2016: 274).

Regarding making a judgement at the end of the assessment, House (2016: 80f.) differentiates between assessment as a linguistic and scientifically objective analysis and what she calls a socio-psychological evaluation, i.e. a judgement. However, not all scholars do so, and Colina (2011: 45) in particular sees no need for such a distinction.

In line with footnote 10 above regarding quality assessment and quality assurance, TQA must, for completeness' sake, also be distinguished from revision and proof-reading (Gouadec 1981: 101; Kockaert & Segers 2017: 148). Both of these latter practices are intended to either educate the translator or improve the translated text, not to judge its quality (Brunette 2000: 170-173). TQA must further be distinguished from quality control (QC), which is a decision process that focusses on the fulfillment of quality requirements and need not include the ST (Brunette 2000: 171ff.; ISO 2015b: [3.3.7]; Lommel et al. 2015).

Now that the relevant terms *translation*, *quality*, *evaluation/assessment* and *translation quality assessment* have been defined and, where necessary, differentiated from related terms, the literature pertaining to translation quality assessment can be discussed in detail.

2 Literature review: The evaluation of translation quality

2.1 Translation Quality Evaluation: Cornerstones, types and methods

The review of the relevant literature pertaining to TQA is divided into two parts: first, the review of works and scholars dealing with the cornerstones, types and general methods of TQA and, second, the review of specific approaches and models intended for the evaluation of translations.

Scholars in Translation Studies generally agree that every kind of (translation) evaluation is built around certain cornerstones. Three of the most important cornerstones are: the purpose of the evaluation (Why is the evaluation carried out? What is the aim of the evaluation?), the evaluation method (How is the object being evaluated?), and the object under evaluation (What is being evaluated?) (Larose 1998: 163f.; Martínez Melis & Hurtado Albir 2001: 275). Other cornerstones mentioned by Larose (1998: 163f.) are the evaluator(s) (Who is carrying out the evaluation?) and the evaluation parameters, which will both be discussed in chapter 3 when Colina's (2008; 2009) TQA model is described in detail.

For each of these cornerstones, especially the main three (evaluation purpose, method, and object), scholars like Larose (1998) and Colina (2012: [2]; 2015: [245]) insist that they must be defined and made explicit prior to the evaluation to ensure clarity and replicability. In order to be better able to place the replication study (for which all the above cornerstones will be defined in chapter 4) within this theoretical framework, different assessment purposes and different methods of translation evaluation will be discussed in brief here.

The **aim or purpose of an evaluation** is especially important for a componential approach like the one proposed by Colina (2008), because it is the purpose of the evaluation (as well as the purpose of the text under evaluation) that “determines the relative importance of the evaluation components” (Colina 2015: [234f.]), meaning the evaluation criteria.

There are many different reasons for carrying out a translation evaluation. The aim can be to assess the suitability of a translation product, to serve a scientific purpose, to evaluate the competence of a translator, or to aid in some kind of decision (Colina 2015: [232ff.]; Martínez Melis & Hurtado Albir 2001: 279), amongst other things. When the suitability of a translated text is being evaluated, the question is often whether that text complies with the assessment *skopos* or whether it matches certain standards (e.g. of acceptability) (Colina 2015: [233f.]; Karoubi 2016: 272). The result is often expressed numerically or as an evaluative statement (Karoubi 2016: 272). It is important to note that, when the translation product is evaluated, characteristics of the product are being examined irrespective of the

thought processes, decisions or intentions of the translator (Karoubi 2016: 275); those factors may, however, feature heavily in the assessment of translator competence, especially in an educational context (Colina 2015: [232]).

Furthermore, TQA may be conducted for informational reasons (translation criticism in the media) or advertising purposes (if the result is used to emphasize the advantages of a product, e.g. a translated book) (Martínez Melis & Hurtado Albir 2001: 279), but a more common use would be for scientific or research purposes. This is when the results or some aspects of the evaluation itself are used for Translation Studies (Gouadec 1981: 101; Martínez Melis & Hurtado Albir 2001: 279), as is the case with the work in hand. A common example of research-focused TQA can be found in the sector of machine translation (MT), where developers of MT systems need to assess how changes to the systems under development affect the output's quality (Lommel et al. 2014: 456).

The results of TQA can also be intended as a decision aid in an employment context, to help decide whether a translator should be recruited for a job, to determine that person's potential (Colina 2015: [234]); Martínez Melis & Hurtado Albir 2001: 279) or to indicate fitness for promotion (Gouadec 1981: 101). This is what Martínez Melis & Hurtado Albir (2001: 277, 279) call *diagnostic* or *pedagogical function* in the context of translator training, where TQA may serve as a placement or selection tool or as an entrance test.

In that connection, a distinction is commonly made between formative and summative assessment (purposes). Formative translation evaluation features almost exclusively in a translator training context, since its main function is to provide feedback to the translator, so that he or she can understand his or her mistakes and learn from them (Colina 2015: [232]; Gouadec 1981: 100; Martínez Melis & Hurtado Albir 2001: 277). Summative evaluation, on the other hand, does not usually include feedback¹¹ and is meant to constitute an overall assessment, i.e. provide an overview of translation quality with regard to a translation product (Colina 2015: [232, 235, 246]). Apart from the distinction between formative and summative evaluation, there are also other types of assessment that will be briefly explained in the following paragraphs.

¹¹ Note, however, that while summative evaluation is encountered more commonly in the professional context, it can also be used in an educational environment, where it usually stands at the end of the learning process and is used to determine whether or not the required skills have been acquired (e.g. end-of-term examinations) (Martínez Melis & Hurtado Albir 2001: 277).

Different types of (translation) quality assessment, constituting the second cornerstone, the **evaluation method**, can best be expressed in contrastive pairs: analytical vs. holistic assessment, criterion-referenced vs. norm-referenced assessment, positive vs. negative evaluation, as well as other types that will be discussed below.

Perhaps the biggest discussion in TQA literature concerns the dichotomy **analytical vs. holistic assessment**. Analytical assessment is focused on finding specific errors in a translation and sorting these errors into fixed error categories (Dalla-Zuanna 2018: 12). Analytical evaluation thus makes use of error analysis grids, scales or matrices in which a value is assigned to each error (Kockaert & Segers 2017: 149), such as the MQM-DQF, the Framework for Standardized Error Marking (by the American Translators Association) (Kockaert & Segers 2017: 150), the SAE J2450 Translation Quality Metric (SAE International 2021), or the (now defunct) LISA TQE model, on which many matrices in the industry are still based (O'Brien 2012: 56). There is also a subtype of analytical assessment that is exclusively focused on error impact, not penalizing any error that does not prevent the intended use of the translation (fatal error) (Gouadec 1989: 43).

Holistic assessment, by contrast, is focused on the text as a whole, not merely on specific errors. Therefore, errors are not sorted into detailed error categories, but the evaluation is based on the entirety of the text (Dalla-Zuanna 2018: 12; Kockaert & Segers 2017: 149). This can make holistic assessment both faster and more adaptable to different assessment needs when compared to analytical evaluation.

While analytical assessment can be partially or entirely automated and is commonly said to be highly objective and reliable, it is also time- and work-intensive and not very flexible (Dalla-Zuanna 2018: 12; Martínez Melis & Hurtado Albir 2001: 283). In addition to that, it poses several other problems commonly associated with error counting. One of the most striking of these problems is that a translation may contain no obvious errors, yet still not be of good quality, whereas another translation may contain quite a number of errors without being of bad quality (Garant 2009: 12).

In spite of this, a survey among TAUS (Translation Automation User Society) members from different industry sectors conducted by O'Brien in 2011 found that of eleven TQA models currently in use, ten were error-counting models, while the last one was a process-focused model (O'Brien 2012: 57). The survey did not show any holistic models (i.e. not based on error counting) to be in use. This finding reflects the continuation of assessment practices, which traditionally have been based almost exclusively on error counting (Conde 2011: 70), although O'Brien (2012: 55f.) explicitly mentions that TAUS had found its

members to be increasingly discontent with static (error-counting) approaches that give no or little consideration to factors such as communicative function, context and end user requirements.

It thus becomes clear that both the analytical and the holistic method have their advantages and disadvantages. Although it is often seen as a quick and easy way of conducting TQA (Kurz 2020: 45), the analytical method tends to take more time to apply than the holistic one (take for example the analytical assessment approach by Gouadec (1981: 109), which allows for differentiation between no less than 675 error types). In recent decades, translation service providers have thus become more and more dissatisfied with these time-intensive approaches, since they generate high costs (O'Brien 2012: 56). Compared to the holistic method, analytical assessment also has the disadvantage that evaluators focus only on small parts of the text (specific errors) and do not necessarily have a good overview of the text as a whole (Kockaert & Segers 2017: 150). In addition to that, analytical assessment schemes typically assign a fixed coefficient to each error category or error type (Martínez Melis & Hurtado Albir 2001: 283), which makes it difficult to take functionalist ideas of weighting errors in relation to the *skopos* into account.

There are, nonetheless, typically two arguments made in favor of analytical assessment: it is supposed to be more reliable than holistic assessment (i.e. it supposedly provides for higher inter-rater agreement), and it is seen as more objective. As one of the few empirical studies on this subject, Waddington (2004) examines four different TQA methods for reliability¹² ((1) error-based without consideration of error impact, (2) error-based with consideration of impact, (3) holistic and (4) hybrid (combination of error-based and holistic)) and indeed comes to the conclusion that the error analysis methods are more reliable than the holistic one (reliability coefficient (.93 vs. .84) (Waddington 2004: 34). However, it must be noted that the holistic method used in Waddington's study was based only on three criteria (accuracy of content transfer, quality of target language (TL) expression and task completion), which were not weighted in accordance with the text in question¹³ (Waddington 2004: 30).

¹² Waddington (2001: 313-316; 324) also examined the same four methods for validity but found no significant differences between the four methods.

¹³ Additionally, the fact that the hybrid method scored higher for reliability than both the holistic and the error-counting models does not seem to be very meaningful, given that the hybrid method was a combination of the scores from the error-counting and the holistic model, where the former counted for 70% of the combined score and the latter for 30%, without any explanation as to why this ratio would be appropriate (Waddington 2004: 34).

Another empirical study by Eyckmans et al. (2009) compared the reliability of an analytical, a holistic, and the CDI-method, a so-called calibration TQA method in which specific elements of the ST (and corresponding elements of the TT) are selected for evaluation according to their discriminating power (i.e. their contribution to measuring differences in translation quality) (Eyckmans et al. 2009: 75) (see chapter 2.2.2 for a more detailed description of a calibration method). The results showed the highest reliability for the CDI-method and lowest for the holistic method, with the analytical method in between (Eyckmans et al. 2009: 81). However, this result was to be expected since the CDI-method selects elements (significantly reducing the number of elements evaluators can disagree on) and the holistic method did not use any kind of defined scoring criteria. According to rater instructions, the holistic assessment was to be based solely on a grade the rater “feel[s] corresponds to the impression [he/she] obtained from the translation as a whole” (Eyckmans et al. 2009: 91). Furthermore, the analytical method had fixed error scores, without any possibility for weighting or adaptation to the context (Eyckmans et al. 2009: 92f.). It would thus seem that there still remains some question as to the reliability of both analytical and holistic methods.

As mentioned above, the analytical assessment method is also commonly seen as the less subjective of the two, with adjectives such as “impression-based” and “imprecise” assigned to holistic assessment methods (Gouadec 1989: 42; Kurz 2020: 43; Martínez Melis & Hurtado Albir 2001: 283). One common misconception in this regard is that holistic assessment is based solely on the gut feeling of the evaluator and does not employ any kind of defined assessment criteria, thus resulting in meaningless ratings like thumbs-up or thumbs-down (Kurz 2020: 43). Interpretability of ratings is certainly an important point, since any score is useless, if no one knows what it means (as is the case with the BLEU metric for MT evaluation, for example) (Koehn 2020: 64). As Garant (2009: 10) points out, however, “‘holistic’ [...] does not mean ‘unsystematic’. Rather, it refers to a systematic way in which the [assessor] arrives at an overall impression of the text as opposed to relying on a discrete points-based scale.” If nothing else, Colina’s model (Colina 2008; 2009; 2015), which will be described in chapter 3, clearly demonstrates that a holistic model can employ defined assessment criteria and that ratings need not be meaningless if they are set in relation to the translation brief.¹⁴

¹⁴ As will be set out in detail in chapter 3, Colina’s (2008) model is composed of four different evaluation components: Target Language, Functional and Textual Adequacy, Non-Specialized Content (Meaning) and Specialized Content and Terminology (Colina 2008: 104). These components are weighted for each assessment in accordance with the text under evaluation, so that ratings assigned to each of these components reflect aspects of quality relative to their importance for the situation in question (Colina 2008: 107; 2015: [269f.]).

Many scholars also point out that an analytical evaluation in no way guarantees objectivity (e.g. Kockaert & Segers 2017: 150), and even Waddington (2004: 31) admits that analytical methods are not as objective as might be thought, given that human evaluators invariably constitute a subjective element. As Karoubi (2016: 273) so aptly puts it, no evaluation can be entirely objective, be it holistic or analytical, since no TQA model can by itself interpret the results and make decisions; that must, by its nature, be a human (and therefore subjective) action. The issue of subjectivity will be further discussed chapter 4.4.

A much less controversial dichotomy of assessment methods is **criterion-referenced vs. norm-referenced** evaluation. In criterion-referenced assessment, translation quality is examined in relation to a previously established criterion (such as the translation brief in the case of Colina's model). Norm-referenced assessment, on the contrary, compares the quality of a given translation with a norm, such as the average quality or the highest/lowest quality level in a given group, or with one or more specific other translation(s) (Colina 2015: [235; 243]).

Another distinction may be made between **positive and negative** evaluation, referring to whether only errors are taken into account (negative evaluation) or if achievements, i.e. positive aspects of a translation such as exceptionally good translation solutions, are also recognized in the assessment (positive evaluation) (Gouadec 1989: 46). Bittner (2020: 230) points out that several combinations of these two types are possible: focusing entirely on errors; focusing entirely on achievements as the sole indicator of quality; counting both errors and achievements and offsetting them against each other; and reaching a conclusion without referring either to specific errors or achievements. Neither type of evaluation is superior in and of itself, but Garant (2009) mentions that informants in his study found positive grading (i.e. counting both achievements and errors) better than exclusively negative grading, since the latter could easily lead to "error avoidance translation strategies" (Garant 2009: 12).

Gouadec (1989: 41) further distinguishes between evaluations whose **outcome has some kind of effect** on or consequence for the translator (as might be the case for an entrance evaluation) and evaluations whose **outcome does not affect the translator** (e.g. TQA for scientific purposes).

Another distinction might be made between evaluations carried out **by humans** and evaluations conducted (partially) **by computers** (Colina 2011: 44). There are automated evaluation programs that can compare ST and TT and thus find formal and terminological (but rarely stylistic) mistakes (House 2016: 82). These programs usually focus on error detection only, without attempting to provide an overall quality score (Lommel et al.

2014: 458). Some programs, like translationQ, constitute a middle way between automated and human assessment in that they support human evaluators with artificial intelligence (televic education 2022).

Returning to purely human-conducted assessment, Gouadec (1989: 42) mentions one further type of TQA that is based solely on revision time¹⁵, i.e. where the time necessary to transform a given translation into a translation of acceptable quality is taken as an indicator of quality. It is an approach reminiscent of what Koehn (2020: 51f.) describes in the context of machine translation, where quality is often measured in post-editing time. The problem with this type of assessment is its dependence on the competence and experience of the revisor. This method would also fall under what Martínez Melis & Hurtado Albir (2001: 283) call partial assessment¹⁶, i.e. an evaluation that only examines one aspect of a translation out of a sum of relevant factors (e.g. evaluating a translator's performance solely on his time efficiency). It thus constitutes somewhat of an opposite to the componential approach by Colina (2008: 100), which aims at recognizing several different aspects of translation quality.

Larose (1998: 174) mentions three further types of QA: firstly, having an expert read the translation (and possibly the ST) and then using that expert's opinion as an indicator of quality (see also Kingscott (2007) for comparison); secondly, giving the TT to target audience members to read and then assessing quality based on their answers to questions that were drawn up with reference to the ST only; and thirdly, comparing the translation to a translation of established (high) quality, which serves as a gold standard¹⁷. The first method would seem very subjective indeed, barring the use of defined assessment criteria, and the third method may not always be meaningful in all contexts, given that the gold standard translation would have to be comparable to the translation under evaluation in all aspects relevant to the translation brief.

The third cornerstone of TQA is the **object of assessment**, i.e. that which is being evaluated. Translation quality evaluation can have many different objects: a target text (i.e. the translation product); a series of target texts that together constitute a translator's work; the

¹⁵ This TQA method mentioned by Gouadec (1989: 42) is not to be confused with a method that will be used in this replication study: evaluators will be asked to make note of the time needed for their evaluation, but only so that it may be compared to the time measurements Colina (2008) obtained, NOT as an indicator of quality in itself.

¹⁶ Partial assessment is not to be confused with a practice commonly found in professional TQA that is called sampling and refers to evaluating only parts of a translated text (samples) due to time or financial constraints, e.g. in the case of very long texts (Brunette 2000: 171).

¹⁷ Regarding the relation between machine translation and TQA it is interesting to note that, in the field of MT, human translations are practically always seen as the gold standard. MT systems aspire to reach human quality levels, while different levels of quality between these human-translated texts are rarely distinguished at all (Drugan 2013: 69).

translator and his/her competence; the translation service; the translation process as well as a (machine) translation system; a translation project or even the evaluation of a translation¹⁸ (Hague et al. 2011: 257f.; Karoubi 2016: 255; Kockaert & Segers 2017: 148; Sager 1983: 126). In the case of the target text being evaluated, the TT may be a student translation¹⁹ (for educational purposes), a professional translation²⁰ (for publication outside a training context) (Colina 2015: [231]) or what Martínez Melis & Hurtado Albir (2001: 273) call a published translation (e.g. literary or biblical texts).

In either case, when conducting an evaluation, it is very important to make sure that the object under evaluation aligns with the aim of the evaluation, since not all objects can serve to fulfill all assessment purposes. If, for example, the quality of a translation product is to be examined (purpose), it cannot be done by merely having a look at the process (object that does not align with purpose), because the quality of the process does not necessarily ensure quality in the product (Gouadec 2010: 271). This is why standards like ISO 9000 or ISO 17100 can only specify good practices regarding the organization of translation projects and the translation process, but they do not in and of themselves guarantee good quality in the finished product (Colina 2008: 98f.; 2011: 44; Drugan 2013: 74f.; Williams 2001: 329).

Most scholars therefore agree that to ascertain product quality, the product needs to be examined. However, not all scholars share this opinion, and some, like al-Qinai (2000: 497) or Drugan (2013: 63), call for the inclusion of the translation process or even the decision-making process of the translator in TQA, referring to what Chesterman (2016: 133f.) calls introspective assessment (i.e. taking into account the translator's thoughts and decisions). In that connection, it seems important to again keep in mind, however, that the translator's decisions, for example, cannot be evaluated by means of the target text, because, as Larose (1998: 171) so fittingly points out, only that which can be perceived can be examined.

¹⁸ In Bittner (2020: 224-264), the object that is being evaluated is not a translation product but rather the evaluations of translation products that had already been carried out by other persons. Thus it is the evaluation itself (in the form of a list of mistakes and a statement/protocol) that is being examined.

¹⁹ Drugan (2013: 39) specifically criticizes the use of student translations when testing TQA models, since student translations do not reflect the reality which the models should be able to hold up against. It is worth noting that Colina's studies (Colina 2008; 2009) did not use student translations, therefore neither will the replication study.

²⁰ Professional translations of course include the translation of audiovisual texts such as subtitles. However, in the context of the thesis, TQA models that deal with the QA of subtitles (e.g. Pedersen 2008) will not be discussed further, since the object of the replication study are target texts intended to be viewed without images or sound. Furthermore, because of the various temporal and spatial constraints it is subject to, subtitling and the translation of subtitles is by its nature different from the object of the thesis (Díaz Cintas 2013: 274ff.).

Or in other words: what evaluators receive for evaluation is the target text, not the thoughts of the translator²¹. This does not mean that the translation product may not in some cases serve to give an indication of the ability of the translator who translated it (Bittner 2020: 55; Karoubi 2016: 255f.; Martínez Melis & Hurtado Albir 2001: 274). It merely means that the evaluation of the product can never fully equal the evaluation of aspects of translator ability not apparent in the product, if the aim is to assess translator competence (Gouadec 1989: 38), and vice-versa. Furthermore, the quality of a translation may be due to factors unrelated to (lack of) translator competence (e.g. time constraints, availability of research tools, changes made by editors etc.), so that to evaluate the translator through the text alone or the text solely on the basis of the translator's competence would be making assumptions on both accounts. Lauscher (2000: 161f.) also points out that while a comparison of ST and TT can yield differences between the two, these differences cannot be attributed to any one cause as long as the translation process is not known in its entirety (which seems hardly possible in practice).

It becomes clear that different evaluative aims require different objects, and different objects require different methods and means of evaluation (Kockaert & Segers 2017: 148). The translation process, for example, may be examined through means like keystroke logging, screen recording, eye tracking or think-aloud protocols (Colina 2015: [231f.]; Kockaert & Segers 2017: 155), whereas the assessment of translator competence may involve competence-specific assessment criteria, specialized correcting and grading scales and aptitude tests, in addition to translated texts (Martínez Melis & Hurtado Albir 2001: 279). Translation as a service can be evaluated indirectly by verifying the service provider's adherence to standards (Kockaert & Segers 2017: 155f.) or more directly, one might imagine, through customer surveys and feedback.

When it comes to the evaluation of translation as a product, however, the means of assessment may take many different and differentiated forms, as will be shown in the second part of the literature review. This second part will discuss different categories of TQA models and some of their characteristics and particularities and also go into a little more detail on some of their advantages and drawbacks. The latter is important, because it will then serve to

²¹ Some CAT-tools do, of course, allow the translator to make comments on some of his or her decisions or make note of parts where they think it important for the evaluator to know why they made specific decisions; however, it seems unlikely that a translator would make detailed comments on all translational decisions in any given text, since explaining each decision in an exhaustive manner would probably take longer than the translation itself. Furthermore, in order to make such explanatory comments, the translator must be aware of the fact that his or her translation will be submitted to evaluation later on and that the evaluator will have access to any comments made. Chesterman (2016: 134) concedes this point, stating that introspective assessment is better suited for descriptive than for evaluative assessment.

highlight in later chapters how Colina attempts to overcome several shortcomings of previous approaches with the model she proposes in Colina (2008: 103f., 128ff.).

2.2 TQA models and approaches

When it comes to evaluating translation products, a great variety of different approaches and TQA models can be observed. This lack of agreement about how translations should be evaluated for quality (Drugan 2013: 36) is due, in part, to differences in the understanding of quality and also due to differences in the understanding of translation (Colina 2012: [1]; Colina 2020: 458). To take a historical example: for many years, the essential TQ evaluation criterion was the distinction between literal and free translation, but later on it was the (now controversial) notion of equivalence (Colina 2009: 238; Lauscher 2000: 151; Martínez Melis & Hurtado Albir 2001: 272). More recently, a shift has again taken place towards function-oriented and specifications-referenced assessment (Hague et al. 2011: 253, 257).

Due to these conceptual differences, it would be difficult to give a chronological overview of TQA models, given that many different approaches have existed at any given point in time. There are, of course, certain tendencies to be observed over time: the tendency towards evaluating quality on more refined, point-based scales where judgements used to be ‘pass’ or ‘fail’ only (Garant 2009: 9f.); the introduction of explicit evaluation criteria and analytical (error) categories, where subjectivity had prevailed before (Horguelin & Pharand 2009: 7f.; Martínez Melis & Hurtado Albir 2001: 273; Secară 2005: 39); a trend towards adaptable models and away from static approaches (Colina 2011: 47; Rodríguez Rodríguez 2007: 19; Williams 2004: 149); as well as a tendency towards holistic models and away from pure error counting (Garant 2009: 11; Williams 2001: 328; 2004: 28). However, an overview according to categories of TQA models seems to be the more structured approach.

It must be noted here that there are many classifications of existing TQA approaches (to be found e.g. in Chesterman (2016: 121-138); House (2009: 222ff.; 2016: 67-71); Lauscher (2000: 151); and Williams (2004: 3-19)). For this MA project, the categorization by Colina (2008; 2009; 2015; 2020) was chosen and will be supplemented by other classifications (e.g. Chesterman 2016: 121-138; Drugan 2013: 45ff.) where appropriate. Colina (2008; 2009; 2015; 2020) identifies five different categories of TQA approaches: experience-based and anecdotal approaches; research-based (academic) approaches; reader-response approaches; equivalence-based approaches; and textual and pragmatic approaches²².

²² Furthermore, Colina (2015: [236]) mentions procedural approaches, meaning models that see the translation process as a prerequisite for quality and do not necessarily evaluate the translation product itself. An example

Each of the following sub-chapters will discuss one of the above-listed categories of TQA approaches.

2.2.1 Experience-based and anecdotal approaches

The first broad category established by Colina (2008: 99f.; 2009: 237) is that of experience-based and anecdotal approaches, also called experiential approaches. As might be guessed from the name, these TQA models are based on practical experience gathered within an organization, rather than on empirical data or theoretical models. Thus, they are usually developed for and fitted to the context of one particular organization or industry only, and are therefore difficult to adapt for other purposes or contexts. Or in other words, the lack of theoretical framework limits their generalizability and transferability to other environments (Colina 2015: [238]; 2020: 458). Experiential models tend to be comprised of error scales and are often sentence-based, giving less regard to the text as a whole (Colina 2011: 44; 2015: [237]; O'Brien 2012: 58). Furthermore, these models do not always clearly define what constitutes an error, so that this decision is left to the assessor, thus limiting inter-rater reliability (i.e. aptness for use by more than one rater, when the goal is to obtain similar results) and replicability (i.e. validity of use in other contexts and situations) (Colina 2008: 100; 2015: [238]).

A classic example of an experience-based TQA model is **SICAL (Système canadien d'appréciation de la qualité linguistique / Canadian Language Quality Measurement System)**, created by the Canadian government's Translation Bureau in the 1970s. SICAL is one of the first TQA models that is based on the categorization of errors into fixed error categories and that allows errors to be labeled according to impact (major or minor) (Secară 2005: 39; Williams 1989: 33). As is often seen in error-counting models, SICAL operates at the word and sentence level only, ignoring everything relating to relations between sentences and the text as a whole (Larose 1998: 175; Secară 2005: 40; Williams 2001: 330). One interesting thing to be noted is that SICAL already showed some rudimentary functionalist ideas, since the category 'major error' was defined in relation to the purpose and the usefulness of the TT (i.e. if an essential element of the text's message was mistranslated, the entire TT was regarded as undeliverable) (Williams 2001: 330). What exactly is meant by 'essential element', however, remained undefined. Apart from some early functionalistic traits in TQA, SICAL is also an example of the above-mentioned tendencies observed in the field

would be the approach proposed by Bolaños Cuellar (2007: [ivff.]) in his doctoral thesis, which integrates both product- and process-oriented aspects. Since the quality assessment of the translation process is not the topic of this MA project, procedural approaches will not be discussed in much detail.

of TQA: the model has gone through multiple versions over the years, changing focus with each, from TQA conducted after delivery of the TT to continuous evaluation over the entire translation process to TQA conducted before delivery to the client, and from absolute quality judgement to degrees of quality (Larose 1998: 166f.).

Examples of other experience-based models include the SAE J2450 Translation Quality Metric, developed for the automotive industry (SAE International 2021), or various entrance tests for professional organizations, like the certification exam scale of the CTTIC (Canadian Translators, Terminologists and Interpreters Council) (CTTIC 2005).

2.2.2 Theoretical and research-based approaches

Somewhat opposed to experience-based (i.e. practical) approaches are the theoretical and research-based (i.e. academic) approaches. Theoretical approaches are built not on experience, but on a theoretical or empirical framework and explicit assumptions regarding the nature of translation (Colina 2008: 100; 2015: [238]). While usually more generalizable than experience-based approaches, academic TQA models tend to be difficult to apply in professional and/or training situations (Colina 2008: 100). The following categories (reader-response approaches, equivalence-based approaches, and functional and textual approaches) along with the models featured in them all fall under the overarching category of theoretical approaches, being based on empirical or theoretical research (Colina 2015: [239f.]).

A typical example of a research-based, analytical model is the **TQA approach by Gouadec (1981)**. Consisting mainly of an error typology featuring two categories (lexical and syntactic errors) as well as a list with correction marks (Gouadec 1981: 103ff.), the model's evaluative process is simple: first finding and categorizing errors in a translated text, followed by calculating the accumulated error points (Gouadec 1981: 110f.). It should be pointed out that Gouadec's (1981) model is already an 'advanced' model in that it tries to mitigate some of the well-known problems of error counting: the model is, for example, not exclusively word- or sentence-based, but can also take entire paragraphs into account, depending on the type of error (Gouadec 1981: 105, 108). Furthermore, Gouadec's (1981: 116) model allows the evaluator to assign fixed error points associated with each error (depending on its severity relative to the importance for the text 'message' as well as the 'circumstances'), thus already showing some functionalist thoughts (Gouadec 1981: 104, 111).

However, the model is, in line with its nature, based on the counting of errors and therefore on the definition of what constitutes an error, its seriousness and its nature (Gouadec 1981: 104). This seemingly simple question still challenges scholars to this day (Martínez

Melis & Hurtado Albir 2001: 281f.; Williams 2001: 328), as shown by the fact that there are many diverging definitions and classifications of errors (e.g. by Gouadec (1989: 37ff.) or Pym (1992: 282)), which are hardly ever reconcilable with each other. Even Gouadec (1989: 37) himself says that every error with its particular impact and context is always a “cas d’espèce”, i.e. a singular and special case, showing that categorizing errors is difficult. Perhaps the best example of this is Gouadec’s (1981: 113) list of correction marks and error categories, which allows for 675 different combinations in marking and classifying an error.

Taking into account further problems associated with error counting (e.g. that analytical models seldom allow for positive evaluation or error marking without penalization and that the focus on errors alone may cause the evaluator to miss other relevant aspects of quality (Gouadec 1989: 36; Nishino 2020: 159; Williams 2001: 333), this explains why many scholars and practitioners in more recent times argue against error-counting models (e.g. Hagemann 2007; Muzii 2016; Williams 2001). Nonetheless, the fact remains that, in spite of the discontent with static error-based models found by O’Brien (2012: 55), the most widespread TQA method in the industry still seems to be error-based (O’Brien 2012: 74).

However, a clear development away from static approaches may be discerned when comparing classic error-counting models with more recent approaches in that department, such as the well-known **MQM-DQF**²³. MQM is a dynamic framework that consists of ten dimensions (criteria), allowing the user to build user-specific quality metrics for the functionally oriented evaluation of verifiable aspects²⁴ of TTs and/or STs (Dalla-Zuanna 2018: 11f.; Lommel et al. 2014: 456; Lommel et al. 2015).

MQM (as a flexible and customizable superset of existing quality metrics (like the LISA QA model, SAE J2450 or the retracted ISO 14080 standard²⁵)) was created partly as a reaction to the above-mentioned criticism leveled at static evaluation models (Dalla-Zuanna 2018: 11). Interestingly, another (seemingly contrary) reason for its creation was to introduce

²³ MQM-DQF is an error-based typology that was created by TAUS (the Translation Automation User Society) as a recognized subset of MQM (Multidimensional Quality Metrics) (Lommel et al. 2015). The MQM-DQF can be found at TAUS (2022). Definitions of the MQM and each of its dimensions as well as information on its use can be found in Lommel et al. (2015).

²⁴ Verifiable aspects as opposed to entirely subjective criteria such as artistry or elegance (Lommel et al. 2015), thus explicitly excluding what House (2016: 67f.) calls psycho-social approaches or mentalist views (TQA approaches that are based entirely on intuition and favor aesthetic and artistic aspects)

²⁵ Although MQM is based in part on models like the LISA QA model and on industry practice (as contributed by TAUS members) (Lommel et al. 2014: 458; Lommel et al. 2015), several factors allow the author of this thesis to argue its inclusion in the category of theoretical and research-based approaches: the MQM initiative was partially funded as part of the EU projects Horizon 2020 and QTLaunchPad (allowing the assumption that it was a structured and theoretically sound construct), the developers of MQM are scientists working i.a. for the German Research Center for Artificial Intelligence (DFKI), and many of the issue types mentioned in MQM match the ones found in scholarly works such as House (1997) (Lommel et al. 2014: 458; Lommel et al. 2015).

some consistency and interchangeability amongst too many user-specific metrics, since the popular pre-existing LISA model had been adapted by so many industry players that each company or industry branch was using a different model (Lommel et al. 2014: 457). MQM thus aims to “avoid the problems of both a one-size-fits all model and the fragmentation of unlimited customization” (Lommel et al. 2014: 458).

The TAUS DQF (Dynamic Quality Framework), being a specific error typology constructed as a subset of MQM, is remarkable because it includes also three additional features not normally found in error matrices (Lommel et al. 2015). Each of these features constitutes an error category that can be marked using the severity level ‘neutral’, indicating that these issues ought to be pointed out without penalizing the translator (Lommel et al. 2015; TAUS 2022)²⁶: ‘Client edit’ means some change requested by the client and ‘Query implementation’ marks changes made after queries to the client. The third feature is ‘Kudos’, a positive severity level that can be used to point out positive aspects in a translation and allow credit to the translator. However, whether and how Kudos impact a DQF-calculated error score remains unclear (Lommel et al. 2015; TAUS 2022).

Another research-based TQA model that (similar to error-counting approaches) shows some of the difficulties involved in concentrating on specific parts of the text is the so-called **PIE (Preselected Items Evaluation) method** by Kockaert & Segers (2017), which is based on the CDI (Calibration of Dichotomous Items) method by Eyckmans et al. (2009)²⁷. Both the CDI and the PIE method are calibration methods. That means that the model is calibrated (adjusted) to the text under evaluation (Kockaert & Segers 2017: 150f.) by using a representative sample of the ST and identifying certain items that have discriminating power, i.e. that can serve to distinguish translation ability and/or performance (Eyckmans et al. 2009: 75). In the case of the PIE method, the selection of elements (items) in the ST is carried out by statistical means and relative to how important the item is with reference to the translation brief (Kockaert & Segers 2017: 152).

Once the preselection is complete, correct and incorrect translation solutions for each preselected item are defined, to be compared later to the translation solutions proposed by the translator(s) (Kockaert & Segers 2017: 152). Contrary to the CDI method, in which only the

²⁶ This possibility of marking errors without penalizing them is a somewhat rare feature in error-based models, though not unique to the MQM-DQF. First ideas towards it can also be found in Gouadec (1981: 109).

²⁷ Note that both approaches are intended only for high-stake summative evaluation (e.g. job selection testing), where many translators are being evaluated and decisions must be made with as little subjective influence as possible (Eyckmans et al. 2009: 76).

correct solutions listed beforehand count as being correct²⁸, the PIE method also recognizes correct solutions that are unexpectedly proposed by the translator(s) (Kockaert & Segers 2017: 153). Both methods have in common, however, that they are dichotomous methods (meaning they only distinguish between correct and incorrect, not between levels of error (Kockaert & Segers 2017: 150f.)) and that they are preselection methods (meaning only the preselected items are examined during the assessment, with all other parts of the text being disregarded as having no discriminating power (Kockaert & Segers 2017: 153f.)).

The PIE method was created to ensure transparency in grading one translation in relation to all other evaluated translations and reduce subjectivity to zero (Kockaert & Segers 2017: 150ff.), although it might be argued that a subjective element is still present in the selection of correct solutions for each item. Although the number of items is limited in the PIE method (contrary to the CDI method), the PIE method still appears fairly time-intensive and complex (Eyckmans et al. 2009: 76; Kockaert & Segers 2017: 150). Both the PIE and CDI methods are also very heavily abstracted, leaving little or no room for professional (human) judgement, as shown by the fact that a simple statistical recalculation can change a translator's score from 5/10 to 1/5 (Kockaert & Segers 2017: 160).

The preselection of items also seems to hold other potential pitfalls. Since only a selected number of items are being evaluated out of the entirety of the items in the ST, the evaluation does not reflect the quality of the translation, but the quality of isolated parts of it, leaving large grey areas. If, for example, a translator translates all the items not preselected correctly, but fails to correctly translate the preselected items, he or she will have failed the assessment with a score of 0, although 85% (assuming that the preselected items make up 15% of all elements in the text) were translated correctly (Kockaert & Segers 2017: 157). Inversely, a translator can achieve a correct translation of all preselected items, but make severe mistakes in other parts of the text (making these impossible to understand or factually wrong); yet this would not show up in the evaluation²⁹ (Kockaert & Segers 2017: 157f.).

²⁸ The fact that only the one correct solution counts towards a positive score was criticized in the CDI method, since scholars and practitioners argue that good translation quality can often be achieved in more than one way and that more than one solution may be entirely acceptable in a given context (Bittner 2020: 54; Muzii 2016).

²⁹ An example can be taken from a case study in which an excerpt of a legal text, translated from French into Dutch, was evaluated using the PIE method (Kockaert & Segers 2017: 156ff.). Only 10 items were preselected (representing only 25% of words in the ST), meaning that some sentences contained no preselected items at all. The following is an example sentence taken from the ST: “Maître ZZZ, conseil [PI 8] du prévenu, a été entendu [PI 9] en sa plaidoirie.” (Kockaert & Segers 2017: 157; emphasis in the original). An approximate translation into English might read like this: “Mr./Ms ZZZ, Esquire, counsel for the defendant, was heard with his/her arguments.” (The translation for both preselected items, underlined in the sentence, was provided in Kockaert & Segers (2017: 158)). As seen above, only “conseil (counsel)” and “a été entendu (was heard)” were selected as items with discriminating power, although the translation brief said to translate “for a legal audience” (Kockaert

Yet another clearly research-based model was proposed by **Fan (1990)**, who calls it a **mathematical model** (Fan 1990: 43). Fan's approach is based on the principles of the fuzzy subset theory and relies heavily on statistical calculations (Fan 1990: 48). Although Fan (1990: 46f.) proposes evaluation criteria as well (faithfulness of the structural message, the affective (accompanying) message and the central message), her model is mainly concerned with resolving the fuzziness, or in other words, the ambiguity of ratings such as 'Excellent', 'Good', 'Fair' or 'Poor', which can mean any range of numerical scores (e.g. 'Excellent' can mean a score of 100/100, but a score of 95/100 can also be 'Excellent', only to a different degree) (Fan 1990: 44f.). To quantify these different degrees, Fan (1990: 45) uses fuzzy subsets of categories like 'Excellent'. The workflow of her model is as follows: human judges assign ratings like 'Excellent', 'Good', 'Poor' etc. to a given translation using her evaluation criteria; the ratings are then fed as input to a computer that assigns weights through mathematical operations, calculates the results and gives a data output (Fan 1990: 50, 52). The results are still ratings like 'Good' or 'Fair', but Fan labels these resulting ratings as less fuzzy, because they were attributed a degree of membership to a subset through mathematical means, stating the extent to which the rating in question belongs to a rating like 'Good' (more or less depending on the coefficient obtained) (Fan 1990: 54).

Fan's model has the advantage of being largely computer-based, which ensures the objective handling of weights and values; nevertheless, the criteria themselves and the assignment of weights are still subjective decisions made by the evaluator (Bittner 2020: 30). When it comes to practical application, however, Fan (1990: 54) admits that her method may not be applicable in practice and is more of theoretical interest. It does serve to illustrate, however, a point already mentioned in chapter 2.1: the importance of tying ratings to some kind of quantifiable meaning, whether that meaning be linked to percentages or be expressed in relation to the translation brief.

Also worth mentioning among the category of research-based models are approaches based on concepts taken from academic fields other than Translation Studies. Examples of such fields that have informed TQA approaches are discourse analysis (Bensoussan & Rosenhouse 1990) or argumentation theory (Williams 2004; 2009).

& Segers 2017: 157). In French, the word "Maître" can mean many things (e.g. teacher, master of something or someone, including a dog), but in this context is obviously used as a French honorary title for a lawyer (which is why "Esquire" is given in the working translation, since this is an honorific often attributed to lawyers in the US-context). Since the item 'Maître' was not preselected, it might have been mistranslated as the Dutch equivalent of 'Master' or something similar, yet would not have garnered any negative points, even though the mistranslation would presumably seem very peculiar to any law practitioner.

A further example of such a model would be the **Argumentation Approach**³⁰ by **Bittner (2020)**, which is based on a theory of logic (Bittner 2020: 171). At its core, the model consists of comparing one translation solution to another translation solution (this need not be an actual TT but can be a translation solution thought up by the evaluator) by means of arguments (Bittner 2020: 172). ‘Argument’ in this context means reasons and justifications why or why not one translation solution is (un)satisfactory in relation to another (Bittner 2020: 175). Bittner (2020: 175) argues that such an approach is advantageous, because the translator and evaluator both should always be able to justify and argue their decisions, and evaluation should be based on whose arguments are more convincing (Bittner 2020: 176). This, however, assumes that the evaluator is always able to reconstruct the translator’s decisions between one translation solution and the other (Bittner 2020: 176, 187, 198), a process that appears inevitably subjective and prone to involve guessing. It also begs the question of how to decide which argument is more convincing, the translator’s or the evaluator’s (given that the evaluator is the one to decide, one might assume a certain bias towards the evaluator’s arguments). Furthermore, to fully realize the potential of the argumentation approach, the evaluator needs to be familiar with dialectical structures and a specific hermeneutical method that serves to reconstruct arguments (Bittner 2020: 188f.; 194), the use of which is evidently time-consuming, as Bittner (2020: 197) admits. In addition to these problems of subjectivity and limited applicability, Bittner also never describes the use of his model for entire texts and leaves the impression that the argumentation approach is to be understood as a somewhat theoretical model that serves as more of a precursor to Bittner’s (2020) study of commented translations than as a stand-alone TQA model.

Another example of a model that is based on a concept from outside of Translation Studies is the one drawn up by **Gutt (2000/2014)**. It is neither based on any specific theory of translation (because the author doubts that such a thing is possible (Gutt 2000/2014: vii, 5ff., 67f.)) nor on the notion of equivalence (which Gutt (2000/2014: 14) deems inadequate for TQA). Instead, the model is based on relevance theory, a communication theory that focuses on maximum benefit for minimum effort (Bittner 2020: 23; Gutt 2000/2014: 212f.). In the context of TQA this means that the target audience should not have to make too great a processing effort to comprehend the translation (Gutt 2000/2014: 101ff., 107). Therefore, Gutt’s (2000/2014: 108ff.) model consists of judging (parts of) translations by how much cognitive effort they require from the audience in relation to the benefit the audience gets

³⁰ Not to be confused with the argumentation-centered approach by Williams (2004; 2009), which is described under textual approaches in chapter 2.2.6.

from making the effort to understand. As Bittner (2000: 25f.) points out, however, Gutt's model is neither intended nor apt for being applied in practice, for several reasons: Gutt (2000/2014) does not describe any method of quantifying the results of the analysis; the model does not seem to be easily applicable to entire texts of any significant length; and the judgement seems to be highly subjective. One might also add that the effort-benefit approach does not appear to be very adaptable, since it offers no parameters that could be adjusted to fit different evaluation contexts or needs.

Now that some examples of research-based approaches have been given, it appears important to go into detail on some views regarding experience-based and research-based TQA approaches, before the discussion of the remaining categories resumes in the following chapters.

2.2.3 Experience-based vs. research-based approaches

The above-mentioned opposition between experience-based and research-based approaches, between theory and practice, is one along which a great divide has opened up. Scholars like Drugan (2013) and Lauscher (2000) point to the fact that academic approaches to TQA are rarely helpful for the professional practice of assessing translation quality, because they do not take the "reality of translating and translations" (Lauscher 2000: 158) into account. Quah (2006: 27) remarks that this is a shortcoming of translation theory (which should solve problems faced by practitioners not only as a matter of interest), which leads to academic approaches being mostly ignored or even rejected by the industry (Lauscher 2000: 149).

The rift between theoretical models and practical TQA is perhaps also due to a fundamental difference in focus: the academic world seeks to determine when a translation is 'good', whereas the professional world rather asks for when a translation is 'good enough' (Drugan 2013:41f.; Koehn 2020: 41). In practice this means that many academic models focus on finding and categorizing errors, whereas professionally used models place an emphasis on sorting adequate and acceptable translations from unacceptable translations (Drugan 2013: 54). That is only one aspect of the distinction between what Drugan (2013) calls academic and applied (or professional) approaches.

Drugan's (2013: 45ff.) distinction goes beyond the mere differentiation between research- and experienced-based approaches, though both categorizations have some points in common with each other. When Drugan (2013: 49) speaks about applied TQA approaches, she does not merely mean models that have been tested on some select cases to measure that model's strengths and weaknesses and/or its applicability (like Colina did in her 2008 and

2009 studies). ‘Applied’ in Drugan’s (2013: 47) sense means that the TQA model in question is “applied on a regular basis in the real world”³¹. Consequently, this means that these applied approaches are “designed, adopted and refined based on ongoing applied professional experience” (Drugan 2013: 49). Furthermore, applied approaches focus heavily on production processes and standards, and the underlying notion is always the one of translation as a service (Drugan 2013: 47). As a result, these professional, applied models not only evaluate the translation product (TT), but also include the translation process³² (including, e.g., supplier recruitment, tools, post-translation checks, post-project evaluation and ongoing quality as well as project management) (Drugan 2013: 127ff.).

Drugan (2013: 80) further distinguishes between two types of real-world (applied) approaches: top-down models and bottom-up models. Top-down models are often one-size-fits-all models and divert many resources to management processes, being focused on structures, workflows and processes, and levels of quality. They are therefore rather inflexible (Drugan 2013: 126, 155f.). ‘Top-down’ refers to the fact that they are hierarchical systems in which TQA is needed for selection and attribution of rewards and penalties and carried out by means of matrices (Drugan 2013: 125f.)³³.

Bottom-up models, on the other hand, are applied to user-generated, fast-changing content (like fan-content or crowdsourced content) (Drugan 2013: 159f.). ‘Bottom-up’ refers to the fact that these approaches emerged to meet the needs of contemporary translation with respect to contexts and users that have been overlooked by the industry (Drugan 2013: 176). They focus more on technology than on processes and standards and are very flexible, given that they rely heavily on users’ decisions and input, as well as feedback (Drugan 2013: 174ff.). Bottom-up models are applied to crowdsourced translations that would often not have been produced otherwise, meaning that bottom-up models allow translations of poor quality rather than having no translation at all (something that, according to Drugan (2013: 180), TSPs often cannot bring themselves to do). An example of a bottom-up TQA model would be the one used by Facebook, where translations are generated by users (with no

³¹ Drugan (2013) does not sort models according to whether or not they would be applicable in a professional context, but according to whether or not they are applied in a broad manner. Therefore, models like the one proposed by O’Brien (2012: 57), which recognizes many practical aspects of translation (such as time limitations, potential impact(s) of mistranslations on the customer etc.) would still fall under academic models according to Drugan’s (2013: 80) categorization.

³² They thus partially match the briefly-mentioned category of procedural approaches by Colina (2015).

³³ A concrete example of top-down approaches cannot be provided, since due to confidentiality reasons, Drugan (2013) does not mention any names of LSPs (language service providers) or TSPs (translation service providers) or any specific matrices. For examples of the different (translation project) stages (pre-translation, supplier selection etc.) in which top-down models operate and what aspects they concern in detail, refer to Drugan (2013: 127-157). Giving even a summary of one such example here is not possible due to spatial reasons.

recognized translation qualification) and approved by public vote (Drugan 2013: 165ff., 174), thus being an example of the ‘good-enough’-focus mentioned above.

After this brief excursus on the division between applied and academic approaches, the following chapters will discuss the remaining categories of TQA approaches by Colina (2008; 2009; 2015).

2.2.4 Reader-response approaches

One particular branch of theoretical models that constitute the third category of TQA models are reader-response approaches (Colina 2008: 100). These models evaluate translation quality based on the effect that a translation has on its target audience: if readers of the TT respond to the text in a similar way that readers of the ST would respond to the original, then the TT is considered to be of high quality (Colina 2008: 100; 2009: 237). This approach, however, presents several problems.

First of all, reader response is difficult to measure, and it is even more difficult to determine whether responses to TT and ST are equivalent, as even monolingual texts do not always garner the same response from different readers (Colina 2008: 101). Additionally, reader response is not equally important for the quality of all types of text, since some translations (like legal texts) may not be reader-oriented at all and thus their quality does not depend much on their effect on the audience (Colina 2015: [239]). Another important issue, especially from a functionalist point of view, is that the translation brief may, in fact, require a different target audience response, or the TT may address a different audience than the ST (Colina 2008: 100f.; 2015: [239]). The reader-response approach is also fairly time-consuming and not easily applied in practice (Colina 2009: 238), as works by scholars like Nida (1964)³⁴ and Carroll (1966) show³⁵.

In connection with reader response, it is interesting to note that Brunette (2000: 176f.) divides the quality criterion *purpose* into *intention* and *effect* (of a TT on its audience), much like the traditional-model approach that Kingscott (2007: 321) mentions. Brunette (2000: 176) thus subsumes reader response under purpose, which is judged by the evaluator based on his or her own reaction to the text. While perhaps somewhat questionable in its practical

³⁴ Perhaps the most well-known example of reader-response-focused TQA is the collection of studies by Nida and Taber (1982: 169-173), wherein they also propose quality tests that do not even require the use of the ST, e.g. by having readers fill in gaps in the TT or having them read the TT aloud and noting where they hesitate or stumble.

³⁵ Aspects of the reader-response approach have been implemented on websites, however, where users are invited to e.g. evaluate automatic translations at the bottom of the page.

applicability, this example makes clear that effect on readers is only one aspect of quality that needs to be assessed alongside and in relation to others (purpose, intention, etc.).

A fairly early TQA approach that might be classified as being based on reader response was proposed by **Carroll (1966)** and is valid for the evaluation of both human and machine translations. The approach consists of dividing a translation into sentences and then letting raters read and rate each sentence on a nine-point scale for intelligibility (of the TT) and later for informativeness (of the ST relative to the TT) (Carroll 1966: 55, 57f.). The time it takes the rater to read and rate each sentence is recorded, providing a supposedly objective correlate against which to measure intelligibility³⁶ (Carroll 1966: 59).

While certainly not an approach compatible with functionalist ideas (the translated sentences are presented in random order to the raters and they never see the entire TT or the ST as a whole), there is one interesting thing that must be noted regarding Carroll's (1966: 58) rating scales: next to a numerical rating they also contain descriptive statements that clearly describe to the rater what each value means and how it can be recognized in the text (the descriptor of value 9 on the intelligibility scale says, for example: "Perfectly clear and intelligible. Reads like ordinary text." (Carroll 1966: 58)). This is a rare feature in a TQA model, as testified by Colina (2008: 106), who herself uses descriptive statements in her model and writes that, to the best of her knowledge, such descriptors had never before been used for the evaluation of translations³⁷.

One notable difference between the descriptive statements in Carroll (1966: 58) and Colina (2008: 128ff.) is that some of Carroll's descriptors reference each other, with value 7 on the informativeness scale being somewhat tautologically described as "Between 6 and 8" (Carroll 1966: 58). This cross-referencing of descriptions seems to somewhat defeat the purpose of a descriptor, which should explain what a numerical value stands for instead of merely stating that it lies between the next-higher and next-lower value. Without a description to match the value to the text under evaluation, evaluators are free to interpret what this in-between rating stands for, which increases the subjectivity of the rating.

2.2.5 Equivalence-based approaches

The fourth category of TQA approaches mentioned by Colina (2020: 459) are equivalence-based approaches. As their name already suggests, these models rely on an a priori notion of equivalence to measure quality, a practice that Colina (2009: 238) calls "problematic and

³⁶ Note that no such supposedly objective correlate seems to exist for the criterion informativeness.

³⁷ Carroll's (1966) model is the only exception that could be found by the author of this thesis.

limiting in descriptive as well as explanatory value”. It must be clarified here that equivalence-based models can, but need not be reader-response approaches (the third category described in the previous chapter 2.2.4). There are also other equivalence-based models, for example some textual and pragmatic approaches or those employing back-translation.

Back-translation (BT) is a translation evaluation method first suggested by Brislin (1969) that is rarely mentioned in Translation Studies (Colina 2020: 459), but often employed for TQA in cross-cultural research³⁸, especially in social research and the health-care context³⁹ (Colina et al. 2017: 268; Tyupa 2011: 36). Back-translation consists of taking a TT and translating this TT back into the source language from which it was originally translated, and then comparing the original ST with the back-translation (Behr 2017: 573; Colina 2020: 459). In theory, this allows the evaluator to spot errors in the translation (errors consequently showing up in the back-translation), but the BT method presents quite a few problems (Colina et al. 2017: 273), which is why Behr (2017: 573) calls it controversial.

Apart from the problematic nature of equivalence itself (addressed in this context e.g. by Tyupa (2011: 38) and Colina et al. (2017: 272, 274f.)), BT is also deceptive, suggesting that the TT is equivalent to the ST, although several factors may lead to seeming equivalence⁴⁰ without the translation being, in fact, of good quality (Brislin 1969: 17). The back-translator might, for example, succeed in disentangling a barely comprehensible translation (having access to the ST) and thus gloss over problems in the translation (Behr 2017: 574; Brislin 1969: 17). Moreover, the more literal the translation and the BT are, the closer the match between BT and ST will be, without saying much about the actual quality of the translation, especially with reader-oriented texts that require natural use of the target language (Behr 2017: 574; Colina et al. 2017: 272, 276). Relating to that, aspects like audience and purpose of the text or TL norms cannot be taken into account with BT, since any alteration in the TT justified by these aspects would appear as a mismatch when compared to the ST (Colina et al. 2017: 273; Txabbarriaga 2009: 3). This might very well result in the

³⁸ A survey conducted by Maneesriwongul & Dixon (2004) examined 47 studies on the translation of research instruments from English into other languages. The survey revealed that 38 of the 47 studies used BT to assess the quality of the translation, and 13 studies even used BT only, without otherwise testing the research instruments (Maneesriwongul & Dixon 2004: 175).

³⁹ BT is notably a step in the linguistic validation of Clinical Outcome Assessments (questionnaires given to patients, which need to be translated and cross-culturally adapted when new medication undergoes clinical trials) (Bond 2020; Wild et al. 2005: 96f.). The aim is normally to attain equivalence between validated English questionnaires and the translated and adapted versions in order to obtain the same results regardless of where the questionnaire was applied (Colina et al. 2017: 269).

⁴⁰ Empirical testing by Behr (2017: 582) shows that BT does not guarantee the equivalence of ST and TT. Behr (2017: 576f.) also found that BT presents very low inter-rater reliability.

communicative failure⁴¹ of the research instrument (Colina et al. 2017: 277). Furthermore, BT constitutes a one-size-fits-all approach that Colina et al. (2017: 281) find problematic (reflecting the increasing tendency towards adaptable as opposed to static evaluation models). Due to these limitations of BT as a TQA tool, IMIA (the International Medical Interpreters Association) advises against BT as a TQA method in their guide on medical translation (Txabarriaga 2009: 3). Colina et al. (2017: 270, 275) also argue that BT is not suitable for evaluating translations, although they admit that it is useful for spotting purely referential errors (e.g. the word-for-word transfer of proper names or chemical formulas).

It should perhaps be mentioned at this point that equivalence-based TQA models need not be reserved for specialized technical or scientific translations. They can also focus on literary translations, as is the case with the **model proposed by van den Broeck** (1985/2014: 57, 59f.), which takes factual equivalence between ST and TT as the yardstick for quality. Van den Broeck's (1985/2014: 57f.) model also makes use of an interesting concept, the 'Adequate Translation', a hypothetical construct that serves as a basis of comparison for optional and obligatory shifts occurring between the ST and TT. This last aspect, however, makes it difficult to apply the model in practice, since it requires the evaluator to take the thoughts and decisions of the translator into account, which of course, is often hardly possible and therefore is not described in detail by van den Broeck (Lauscher 2000: 156).

Another problem with the model is that, while based at least in part on functionalist ideas (i.e. requiring all textual elements of functional relevance to be taken into account), function and how to operationalize it for use in the model remains undefined (Lauscher 2000: 156; van den Broeck 1985/2014: 56ff., 61). Since the Adequate Translation (as constructed by the evaluator) is used as a gold-standard for comparison, a large opening is also left to subjectivity, given that the evaluator alone decides, without any defined reference, what constitutes an optional or obligatory shift. It thus becomes apparent that van den Broeck's (1985/2014) model suffers from limited applicability (a problem common to many functionally oriented models) and furthermore also presents some difficulties arising from the equivalence-approach, apart from lack of information on how to quantify the judgement resulting from the analysis (van den Broeck 1985/2014: 58). In fact, it might be said that van

⁴¹ Fourie & Feinauer (2005: 351f.) reached a similar conclusion when they compared two sets of translations of a quality-of-life questionnaire. One set had been produced using a literal translation method whereas the second was translated using a functional method. Fourie & Feinauer (2005: 358f.) found that target group members rated the second set as superior and as easier to comprehend. On a related note, Douglas & Nijssen (2003: 629, 632) found that a questionnaire translated using back-translation only was not as effective as a translated questionnaire that had also been adapted to the target audience and culture.

den Broeck's model is not necessarily an evaluation approach but is more suited for what van den Broeck (1985/2014: 55, 58f.) calls translation criticism and description.

Yet other examples would be the equivalence-based models by Gerzymisch-Arbogast (2001) as well as the often-mentioned TQA model by House (1977; 1997; 2001; 2015), which the author calls a functional-pragmatic model (and which is discussed in more detail in the following chapter 2.2.6 under functional models).

At the end of this chapter it should be pointed out once again that the concept of equivalence has become increasingly controversial in Translation Studies and TQA models tend to use it less. An example of this would be **al-Qinai's (2000) model** which is based on the approach by House (1997) but distances itself from the yardstick of equivalence (al-Qinai, 2000: 498). Instead, al-Qinai (2000: 499f.) uses seven parameters to evaluate the adequacy of a translation, which take into account both textual and functionalist aspects. Each parameter serves as a category in which (positive and negative) correspondences and differences between ST and TT are noted (al-Qinai 2000: 500-515). Following that, a holistic summary of similarities and divergences⁴² between ST and TT is presented for each parameter; however, the model does not provide any means for quantifying these results or assessing the adequacy (or lack thereof) of the TT (al-Qinai 2000: 516). In fact, al-Qinai (2000) does not even describe how to weigh each parameter against the other or how to counterbalance negative findings with positive ones (Drugan 2013: 59). Furthermore, Drugan (2013: 59) remarks that the analysis is too detailed to be workable in practice. A curious point in that regard is that al-Qinai (2000: 517) seems to question the need for the application of his own model, insisting on the necessity of conducting market research among target audience members in order to evaluate a translation. While this shows a consideration for practical issues (market research being, after all, common in many industries), it also begs the question as to why his model is then purely text-based and why evaluation commissioners should use it instead of market surveys. This might serve as an illustration of the previously mentioned idea that scholars who create a TQA model should also consider its application and use in practice, or risk limiting their model to academic usefulness only.

⁴² Given this purely descriptive outcome of the analysis, one might assume that the entire model is descriptive in nature and not intended for actual TQA, but al-Qinai (2000: 499) explicitly states that his model is, in fact, intended for the evaluation of finished TTs.

2.2.6 Functional, textual and pragmatic approaches

It becomes clear from what has been set out above that Colina's (2008; 2009; 2015) categories overlap in places and that some functional and pragmatic approaches rely on the notion of equivalence in some way or another⁴³. Contrary to some equivalence-based models, however, textual and pragmatic models tend to assess texts as a whole (Colina 2008: 101). They compare textual features of the ST and TT and evaluate quality relative to text type and text function, or to put it otherwise, they assess quality with respect to functional equivalence between the ST and the translation (Colina 2012: [3]; 2015: [239]). In addition to this controversial reliance on equivalence, some functional-textual approaches have been criticized (Colina 2009: 238), because they focus almost exclusively⁴⁴ on either the ST (like the model by Reiß (1986: 23, 93)) or the TT (like the *Skopos* theory developed by Reiß & Vermeer (1984)). Furthermore, as has been mentioned before, many functional models are also difficult to apply in practice and therefore still not widely adopted (Colina 2009: 238; 2015: [239]).

An example of this would be **House's (1977; 1997; 2014) functional-pragmatic⁴⁵ TQA model**, which is often discussed in academic circles, but presents some problems when it comes to practical application. The model consists of a comparative analysis between the linguistic and situational features of the ST and the TT, followed by an assessment determining their match (Colina 2008: 101; House 2009: 224), but only the theoretical part (the analysis) is described in House's work, not what she calls the social judgement, i.e. the assessment itself and how to reach a conclusion or a score (Colina 2020: 460). This lack of consideration for applicability in practical TQA is a feature common to quite a few textual and pragmatic approaches, according to Colina (2008: 102). Furthermore, House's model relies on the matching function (and textual profile) of the ST and TT, which is problematic

⁴³ Not all textual and/or pragmatic approaches do so, however. Notable exceptions include the *Skopos* theory by Reiß & Vermeer (1984), which is based on TT function, the model by Bowker (2001) described below and the specifications-based approach by Hague et al. (2011).

⁴⁴ Thus constituting the purest forms of what Chesterman (2016: 121ff.; 125ff.) terms *retrospective assessment* (focused on the relationship between ST and TT) and *prospective assessment* (focused on the TT and the effects it is intended to have).

⁴⁵ House's (1977; 1997; 2015) model combines both functionalist and equivalence-based features and might thus be placed in both categories. In fact, the author sees equivalence as the basic criterion for all TQA (House 2016: 72); however, since House (2014: 248) labels her model as 'functional-pragmatic', it was included in the category of functional and textual-pragmatic approaches. House's (1977: 107; 2016: 75ff.) model presents many interesting aspects that have been often discussed in the field of TQA at large (e.g. the distinction between overt and covert translation or between translations and versions, or the concept of a cultural filter), but discussing them in detail would go beyond the scope of this MA project.

because the function of a TT might be different from the function of the ST, which House's model does not allow for (Colina 2008: 101).

Another model that may serve as an example of a functionalist approach suffering from applicability problems is the **TQA model by Ammann (1990)**, which is intended for the evaluation of literary translations. Ammann's model (1990: 211) is based on *Skopos* theory as well as Vermeer's⁴⁶ model of translation criticism. Translation quality is evaluated in terms of text function and coherent interpretation by the target audience and by means of the scenes-and-frames sematic (Ammann 1990: 209). The evaluation is conducted by first determining the parameters of function and coherence of the ST and TT, respectively, before comparing the coherence of ST and TT with each other (Ammann 1990: 212). Like many functionalist approaches, however, Ammann's (1990) model suffers from the above-mentioned applicability problems due to how vague the scenes-and-frames approach is and because the author never describes how to determine whether a scene has been adequately expressed by a frame (Lauscher 2000: 157). Furthermore, the results of the analysis are not quantifiable in any way, merely summarized by statements that are, as Ammann (1990: 245f.) admits, more or less subjective. Her model thus lacks the means of reaching some kind of score or quality rating that could hold meaning in relation to the translation brief and the function of the TT under evaluation.

Before this, one of the first systematic and perhaps also one of the best-known text-based TQA approaches was proposed by Reiß in 1971. **Reiß's textual approach** largely relies on textual analysis of ST and TT and subsequent comparison of the results, but uses the vague concept of 'optimum equivalence' as yardstick for the evaluation (Lauscher 2000: 152; Reiß 1986: 23). The model also makes use of Reiß's well-known typology of text types (Reiß 1986: 31) to determine which translations would be considered adequate for the type of text in question. Reiß (1986: 69ff.) already shows consideration for the non-linguistic context of translations (captured in her category 'extralinguistic determinants') and explains that evaluation must take place above the word-level, thus paving the way for holistic evaluation approaches (Bittner 2020: 28; Lauscher 2000: 152). However, Reiß's model suffers the distinct disadvantage of not being applicable to all target texts, since her approach is not suited for the evaluation of translations that do not address the same target audience as the corresponding ST (Lauscher 2000: 152; Reiß 1986: 93).

⁴⁶ This model of translation criticism by Vermeer is described in "Beiträge zu einer Theorie der Übersetzungskritik", a manuscript by Vermeer cited in Ammann (1990), which does not seem to be publically available and thus could not be included in the bibliography of this thesis.

Based on the TQA approach by Reiß, Nord (2005) proposed a **model of translation-relevant text analysis** for the purpose of translation evaluation. This text-based model consists of a very thorough analysis of the ST (using 76 questions) to determine the features and function of the ST (Drugan 2013: 66; Nord 2005: 41ff.). The TT is then analyzed and compared to the analysis of the ST, with the focus above sentence-level and the purpose and effect of the translation in comparison to the ST being the yardstick of the evaluation (Nord 2005: 183). Nord's (2005) approach is very detailed, so much so, in fact, that scholars have questioned the exhaustive and seemingly over-detailed nature of her analytical parameters and features (Pym 1993: 186) (somewhat reminiscent of the 675-error-types model by Gouadec (1981: 109, 113)). Another drawback to Nord's model of translation analysis is that she exclusively describes how to analyze features of a ST and compare specific parts of it with the TT equivalents, without going into detail on how one would then arrive at some kind of overall quality assessment⁴⁷ (Williams 2004: 13). Furthermore, the strong focus on the ST apparent in Nord's (2005) model is somewhat questionable in TQA, since the relationship between ST and TT is only accessible to a small part of the target audience (those who actually understand the source language and have access to the ST), whereas the TT is the most important element for the majority of readers in the target culture (Karoubi 2016: 269).

A somewhat different type of textual TQA model is the **argumentation-centered TQA model (called ARTRAQ model)** developed by Williams (2004; 2009), which applies argumentation theory to TQA. The approach is based on the assumption that an argument⁴⁸ (macro-)structure is present in every text, even in texts that are not argumentative in nature (Colina 2020: 460; Williams 2004: XVIIIff.; 2009: 11, 13). The reason behind Williams' (2004; 2009) application of argumentation theory (which focuses on macrostructures instead of microtextual aspects) to TQA is that in the author's opinion, evaluators are less likely to disagree about the argument structure in a text than about microtextual issues (Bittner 2020: 14; Williams 2004: 150).

⁴⁷ It must be noted, however, that use in a professional TQA environment does not appear to have been Nord's (2005) main goal, as she never mentions any kind of scoring or marking grid, although she writes extensively about the application of her model in translation training (Nord 2005: 155).

⁴⁸ It is important to note here that the term *argumentation* in Williams' (2004; 2009) works does not only mean the sense of the word that is commonly understood, but refers to Toulmin et al. (1984) and their analysis of argument structure (Williams 2004: 23). Williams (2001: 335) defines argumentation as "reasoned discourse [that] [...] embraces the techniques of rhetoric as means of persuading the audience or readership". According to this, the argument structure inherently present in every text consists of six elements: claim, grounds, warrant, backing, qualifier, and rebuttal (Williams 2004: 24). Each argument is characterized by one or more organizational relations (such as problem-solution), by connectives (conjunctives), by propositional functions, by argument types (e.g. comparison), by figures of speech, and narrative strategy (e.g. depersonalization) (Bittner 2020: 14 referring to Williams 2004: 24). For a concrete example of how to apply these elements in TQA, see Williams (2004: 28ff.).

According to Williams (2009: 11), it is the argument macrostructure that must be preserved when a ST is translated into a TT. Consequently, the ARTRAQ model consists of an analysis of the argument structure in the ST and TT, followed by a comparative assessment of both argument structures (and especially the transfer of argument to the TT) that results in an overall quality statement (Drugan 2013: 60; Williams 2004: 73, 138). When it comes to argument transfer, Williams distinguishes between different error severities defined in relation to the effect the error has on the text's central reasoning (and thus its usability) (Williams 2004: 133; 2009: 13). Furthermore, the evaluation parameters (argumentation structure elements) are weighted according to text type and purpose of the TT, thus incorporating functionalist aspects into the ARTRAQ model (Williams 2004: 154f.). Additionally, Williams' model is adaptable to different contexts and client needs, including the option to add grades or (minimum) scores as necessary (Williams 2004: 149; 2009: 18f.). The final result can be interpreted through a set of 'standards' of acceptability defined by Williams (2004: 156; 2009: 19).

Unlike some TQA approaches, Williams' ARTRAQ model is designed for application in practical TQA and thus both quantifiable and applicable (Bittner 2020: 15), but it does appear very complex and difficult to learn (see footnote 48 for a very brief summary of elements and their relations) and also has some other drawbacks. Williams (2004: 73f.; 2009: 14) himself admits that an ARTRAQ analysis is time-consuming and that in order not to miss something crucial, all passages containing key arguments should be examined. This does not allow for sampling (see below) (or only on the basis of identifying key areas of the argument structure, which requires prior reading of the entire text) (Williams: 2004: 134, 151). Regarding the practical application of the ARTRAQ model, Drugan (2013: 63) therefore points out that an argumentation-centered approach would be difficult to implement in practice, where evaluators often work in separate teams and only on samples of texts.

The last, and perhaps major question asked in relation with Williams' model is whether the argument structure really is equally important and viable for all texts and text types (Colina 2020: 461; Drugan 2013: 63; House 2009: 223). This question Williams (2004: 69) preemptively tried to answer by testing his model on texts that do not obviously rely on argumentation in order to see whether the ARTRAQ model works for them. While his study shows that the model is, in principle, applicable to such texts, the results also cast some doubt as to the applicability and perhaps the reliability of the approach: all four texts assessed with the ARTRAQ model were found to be substandard, although in a prior test run with a not

yet refined model, two out of four texts were found to be satisfactory (Williams 2004: 147, 124).

One further TQA approach that shows strong functionalist tendencies is **Larose's (1994) assessment model**, which others have called a teleological⁴⁹ model (Drugan 2013: 48). This is a fitting description, given that Larose's central criterion for his multi-criteria evaluation is the translation purpose (Larose 1994: 362). In accordance with that, Larose's model places great emphasis on the translation brief, with all criteria being weighted relative to their importance in the brief (Larose 1994: 366f.), a characteristic also to be found in Colina's (2008; 2009) model (see chapter 3). However, like several models mentioned above, Larose's (1994) model is described only broadly and no specific criteria are defined, which would probably make the model difficult to apply in practice.

Also included in the category of textual approaches **are corpus-based approaches**, like the one proposed **by Bowker (2001)** for TQA in a translation training context. Being aptly named, this approach is based on the comparison of features of a TT to features retrieved from a corpus (i.e. an electronic collection of naturally occurring texts selected according to a special purpose) (Bowker 2001: 345f., 349). It does not rely on textual equivalence as an evaluation parameter (Colina 2011: 46), but neither does it provide instructions as to how a translation should be graded after the analysis is complete (Bowker 2001: 346).

Bowker (2001: 346) holds the opinion that TQA should not rely on intuitive judgements or anecdotal evidence, but rather on "empirical analyses of larger bodies of authentic text" (Bowker 2001: 346) (compare the tendency towards empirical and reasoned approaches mentioned at the beginning of chapter 2.2). Combining quantitative and qualitative assessment (provided by the computer and the assessor, respectively (Bowker 2001: 346)), Bowker's approach is nevertheless time-consuming and rather complex (Colina 2020: 461), since it requires the evaluator to construct an Evaluation Corpus, comprised of four different subcorpora⁵⁰. The approach presents the unquestionable advantages of allowing evaluators to judge a text even if they are not experts in the subject field in question and of relying partially on computers (who present data consistently and reliably without getting

⁴⁹ Teleological assessment is defined as the "[d]etermination of quality with reference to the consequences, outcomes, or goals of the subject of the assessment" (Williams 2004: 165).

⁵⁰ These are the Comparable Source Corpus (containing STs similar to the ST in question, allowing the evaluator to judge how representative the ST in question is), the Quality Corpus (a small corpus of high-quality, reasonably current texts written by experts that allow the evaluator to compare conceptual content), the Quantity Corpus (a larger corpus of representative texts, enabling the evaluator to discover patterns and make decisions based on concrete evidence) and the Inappropriate Corpus (comprised of inappropriate texts in relation to the subject matter, so that poor translation solutions may be recognized) (Bowker 2001: 350-354).

distracted) (Bowker 2001: 345f.), but still remains difficult to apply in professional environments (Colina 2008: 102f.; 2009: 239), where subject fields of texts may change frequently and time constraints may not allow the construction of an Evaluation Corpus.

One thing that all of the models in all the categories described above have in common is that they all focus on partial aspects of translation quality (e.g. reader response, textual or pragmatic aspects, equivalence with the ST etc.) (Barghout 1990⁵¹; Colina 2011: 46; 2015: [240]), whereas a “comprehensive approach to evaluation would address multiple components of quality simultaneously and would allow for prioritizing the components according to user-defined views of quality, user needs and the purpose of the translation” (Colina 2015: [241]). However, in 2001, Martínez Melis & Hurtado Albir (2001: 274) still noted that there was “no operative model of analysis for translation evaluation that is capable of bringing together textual, contextual and functionalist criteria, and that has been sufficiently validated by means of empirical-experimental research.” Now, one might argue, this gap has been filled by the TQA model proposed by Colina (2008; 2009)⁵², which will be described in detail in the following chapter.

⁵¹ In the abstract to his doctoral thesis, Barghout (1990: iif.) remarks that existing TQA models focus only on one aspect of quality each, and, like Colina (2008; 2009), he seeks to overcome this in his rhetorical TQA model, which focuses on shifts of meaning and consists mostly of ST and TT analysis (Barghout 1990: 108; 154). For more information, refer to Barghout (1990: 108ff., 154ff.), since a detailed comparison between Barghout’s (1990) and Colina’s (2008; 2009) models would go beyond the scope of this MA project.

⁵² It remains unclear what Martínez Melis & Hurtado Albir (2001: 274) would see as ‘sufficient validation’, but given that Colina (2008; 2009) conducted two empirical scientific studies on the usability of her model, it can be argued that the above criterion demanded by Martínez Melis & Hurtado Albir has been fulfilled.

3 TQA tool by Sonia Colina (2008)

3.1 Rationale of creation

Sonia Colina's TQA model (Colina 2008; 2009) was developed under the **Hablamos Juntos** grant program, a US national initiative funded by the Robert Wood Johnson Foundation (Colina 2015: [241]). The aim of the program was to study and mitigate language barriers in US health care for patients with little or no English proficiency (Hablamos Juntos 2009: 12). One subcomponent of the program concerned the **quality of translations**, which had been found to often be inadequate in the US health care context; therefore a tool was developed to assess translation quality for a wide range of texts (Colina 2009: 236; 2015: [241]).

In doing so, it was Colina's (2008: 99) stated goal to "attempt[...] to overcome some of the deficiencies of existing [TQA] models" (as pointed out in chapter 2.2). Colina's (2008; 2009) work corresponds to a demand that has also been uttered by Translation Studies scholars like House (2009: 224), who called for "a model which attempts to transcend anecdotalism, reductionism [...] and intuitively implausible one-sided considerations of the source or target text alone".

As stated in chapter 2.2.1, anecdotal approaches lack a theoretical framework, something which Colina (2008: 103f.) clearly has remedied in her model: it is based on functionalism and textual considerations, explicitly states its purpose, object and other cornerstones, and the empirical testing in Colina (2008; 2009) included testable research questions and hypotheses. Reductionism in the above statement refers to the fact that many existing TQA models are based on one aspect of quality alone (e.g. reader response), a fact that Colina (2008: 103) has criticized heavily and attempted to overcome with her componential approach. 'Componential' refers to the fact that the model recognizes several qualitative aspects (Colina 2008: 104), as further laid out in the following chapter. Concerning the last part of House's (2009) demand, the TQA model by Colina (2008: 128ff.; Hablamos Juntos 2009: 5) instructs evaluators to read both the ST and the TT so as to be able to judge the TT's quality relative to the ST and the translation brief (contrary to some text-based models).

Further shortcomings of previous models, pointed out by Colina (2008: 103), include the lack of practical applicability that many research-based approaches exhibit (due to over-complexity of the models in addition to them being time-consuming) as well as the lack of empirical testing. The latter was of course implemented in Colina (2008: 121; 2009), where it

was found that evaluators had almost no trouble applying the model and found it easy to learn its use.

Yet more shortcomings can be extracted from chapter 2.2, where they were pointed out in existing TQA models (for a comprehensive overview of all shortcomings and how and to what extent they are (or are not) overcome by Colina's (2008) model, see the table in Appendix II). Quite a few of those models do not allow for the results of the evaluation to be quantified in a meaningful way, which is made possible by the score sheet in Colina's (2008: 130) model⁵³. Other models, particularly experience- and error-based ones, adopt a one-size-fits-all approach with fixed categories and fixed error scores, whereas the model by Colina (2008: 107) is adaptable and allows for weighting of categories. Furthermore, most previously existing TQA models allow for negative evaluation (error marking) only, with few providing possibilities to note positive aspects and mark them accordingly. The nature of Colina's (2008: 128ff.) model, however, overcomes this difficulty, since descriptive statements associated with each criterion allow for a continuum of grading both negative and positive aspects.

As mentioned above, Colina's (2008) model aims for a comprehensive assessment including both the TT and the ST associated with it. This instruction for the raters to evaluate the entire text, however, constitutes a somewhat idealized process that, while recommended by scholars wherever possible, cannot always be observed in practice due to time constraints (Drugan 2013: 63; Kurz 2018: 20; Larose 1998: 180), which sometimes allow only for the evaluation of selected portions of a text. This practice, known as sampling, has the distinct advantage of being less time-consuming than the evaluation of texts in their entirety, but also presents several disadvantages: it is difficult to ensure that the chosen sample(s) is/are representative of the whole text; evaluators may not gain an understanding of the whole text; they may get an overly negative impression of the text because errors or changes made in the sample under examination are compensated for in other parts of the text; or the exact opposite may be true and a fatal error may lurk just one sentence outside the portion of text chosen as the evaluation sample (Gouadec 1989: 50; Kurz 2020: 62f.; Williams 2001: 328, 335).

⁵³ In Colina's second study (Colina 2009), the model was adapted slightly from the earlier version used in 2008, and a cover sheet was included that allows evaluators to further relate the evaluation results to the brief by giving one of the following recommendations: "Publish and/or use as is [;] Minor edits needed before publishing [;] Major revision needed before publishing [;] Redo translation [;] Translation will not be an effective communication strategy for this text. Explore other options (e.g. create new target language materials)" (Colina 2009: 258).

This problem of sampling vs. time-intensive evaluation is a difficulty that all TQA approaches (barring, perhaps, automatic assessment) face and that Colina (2008; 2009) has not managed to overcome in her TQA model. In fact, evaluators should ideally examine a translation four times (once for each assessment criterion), or at least once (without reference to the ST) for the criteria Target Language and Functional and Textual Adequacy and a second time (referring also to the ST) for the other two criteria (Colina 2015: [266]; Hablamos Juntos 2009: 5). It remains unclear whether Colina's (2008; 2009) model allows for sampling, since the author does not mention this specifically. It might be assumed, however, that sampling is possible (given the functionalist nature of the model) if not ideal, entailing all the risks mentioned above.

3.2 Cornerstones and characteristics

As has been mentioned in chapter 2.1, any TQA attempt should first describe its cornerstones and theoretical assumptions. They will now be described very briefly for Colina's (2008; 2009) model and later be mentioned in chapter 4.1 for the replication study.

Colina's (2008: 99, 105) model is intended for the summative assessment of translation products, whereas the translation process is not part of the evaluation. It is further intended solely for the use by human evaluators (Hablamos Juntos 2009: 10f.). As will become apparent in this chapter, the model provides for criterion-referenced assessment that takes both positive and negative aspects into account.

Colina's (2008: 107f.) model is furthermore intended to be applicable across languages, as shown by the fact that testing (Colina 2008; 2009) took place using the languages English, Russian, Chinese, and Spanish. Through the present MA project, the model was also tested on the German language. However, as long as descriptive statements and instructions are provided in a language understandable to the evaluators, there appears to be no reason why Colina's (2008; 2009) model should not be applicable to most, if not all, languages (see also the paragraph on 'adaptable' in this chapter).

In accordance with the author's attempt to overcome previously existing shortcomings, the TQA model by Colina (2008; 2009) has some very specific characteristics: it is functionalist, "componential, descriptive (and user-defined), theoretically explicit and testable [...] [and] also easy to apply" (Colina 2015: [242]). Each of these characteristics will now be briefly discussed in the following paragraphs.

‘Functionalist’ refers to the fact that the model is based on functional approaches (like *Skopos* theory by Reiß & Vermeer (1984)) or textual approaches (as in Holz-Mänttari (1984)) (Colina 2008: 104). In a TQA context that means that a) function is the factor that determines the relationship between the ST and the TT; b) function is captured in the translation brief; c) criteria for translation evaluation are defined in accordance with the translation brief; and d) textual approaches come to bear because text type and text function play a major role in evaluating how well the TT fulfills the specifications asked for in the brief (Colina 2008: 104). It follows, therefore, that evaluation must always take place relative to and with regard to the translation brief (Colina 2008: 104; Gouadec 1989: 39; Karoubi 2016: 261), which is “a set of instructions that define the purpose and use of a source [and/or target] text and establish priorities to guide the translators” (Hablamos Juntos 2009: 8). Or to put it otherwise, the translation brief serves as reference point and determines the degree to which the TT under evaluation must comply with the (four) components of quality that are the criteria of Colina’s (2015: [243]) model (see the following chapter). For more detailed information on how the brief is captured and operationalized in the model, see chapter 4.3.

‘Componential’ means that to operationalize the notion of quality, it is evaluated in the form of separate components (Colina 2015: [242]). This is done because no single criterion (e.g. translator qualification, certification to a standard, quality of process etc.) can in itself be a measure of translation quality (Kurz 2009: 148f.). In order to achieve a comprehensive evaluation that is suitable and adaptable to different evaluation contexts and not reduced to one single criterion, more than one aspect of quality needs to be recognized, which is just what the componential approach does (Colina 2008: 103; 2015: [242]; Larose 1994: 367). The componential nature of Colina’s (2008: 105) model may be seen as a consequence of the author’s attempt to integrate different approaches to translation quality (functional, pragmatic, textual, reader response etc.) that previously existing models addressed separately (see chapter 2.2 and Appendix II).

The four components featured in Colina’s (2008: 104) tool are “linguistic form: spelling, grammar, lexicon, etc; functional adequacy; meaning[;] and specialized content”, corresponding to the assessment criteria Target Language; Functional and Textual Adequacy, Non-Specialized Content (Meaning); and Specialized Content and Terminology (Colina 2008: 128f.). These non-technical assessment categories not only reflect the functionalist nature of the model, but are also easy to understand for customers (who can decide on evaluation priorities by weighting some criteria higher than others and easily understand the results) (Colina 2008: 124).

One point that perhaps merits some further discussion is the number of assessment criteria: four. As Reinart (2014: 375) correctly states, the reality of TQA is very complex and cannot be broken down into merely some selected criteria; yet in order to be applicable in practice, that is what any TQA model must do. A great number of criteria may make for a comprehensive and convincing evaluation, but will ultimately render the model inapplicable due to over-complexity and time restraints (compare to the model by Gouadec (1981) featuring 675 error classifications) (Bittner 2020: 53; Muzii 2016). Consequently, any model is by necessity somewhat reductionist (Williams 2001: 333), which is why Colina's (2008; 2009) TQA model features only four, carefully selected components (although they can be added to at need). It might be said therefore that Colina's (2008) componential approach corresponds well to the demand that criteria should be "easy to understand, practical, limited in number, and verifiable" (Brunette 2000: 174).

As mentioned before, Colina's (2008: 106) model uses descriptive statements to mark a translation's degree of compliance with each of the four evaluation criteria; the model is therefore **descriptive** in nature. This is a somewhat rare feature in a TQA model, which has not often been incorporated in TQA approaches before⁵⁴, although it can be more frequently observed in the language proficiency context, such as the European Language Portfolio (Abuja et al. 2014: 8f.). In practice, a descriptive model means that, instead of finding errors and deducting points from a total, evaluators compare the TT to the descriptive statements associated with each evaluation criterion and then choose the descriptor that best matches the text under evaluation (Colina 2008: 106; 2015: [242]). This allows evaluators to assess the textual and functional adequacy of translation as opposed to only being able to mark errors at the word or sentence level (Colina 2008: 107). Furthermore, descriptors can be adapted to reflect priorities of a specific evaluation context (Colina 2015: [242]).

In fact, Colina's (2008: 107) model in its entirety is **user-defined**, i.e. **adaptable**, because the combination between brief-referenced evaluation and multiple criteria allows for weighting of the evaluation criteria according to the needs of the requester, something that many scholars recommend (Bittner 2020: 10; Kurz 2018: 17; Sager 1983: 127) and some have even incorporated into their own models (e.g. Larose (1994: 366f.) or the JTF (Nishino 2020: 162)). Weighting criteria in accordance with the context is necessary because "quality requirements are variable, and one and the same target text can be acceptable in one translation situation and unacceptable in another" (Bittner 2020: 52).

⁵⁴ Notable exceptions are Carroll (1966) (see chapter 2.2) and Colina (2003), but the latter is a TQA model for the translator training context only.

As will be further discussed in chapter 4.3, each of the four evaluation criteria in Colina's (2008: 107) model is assigned a total numerical weight (e.g. 20), which represents the highest possible rating a text can obtain for that criterion and is therefore attributed to the descriptor reflecting the highest point of the quality range. The other descriptors associated with the same criterion are consequently attributed lower weights (e.g. 15 for the second highest descriptor, 10 for the one below that etc.). By modifying the total weight associated with a criterion or by attributing especially high or low weights to certain descriptors, the model can be adapted to suit the needs of a specific evaluation context (Colina 2015: [269f.]). Beyond the adaption of criteria-weighting, the model also allows for the modification (and/or addition and omission) of quality components (i.e. criteria) or descriptors⁵⁵ as needed, or for a summary evaluation without any numerical results (by simply not attributing any weights to the descriptors) (Colina 2015: 270; Colina 2020: 461).

Last, but not least, Colina's (2015: [243]; 2020: 461) model is **intended for practical application** (contrary to many functionalist models) and therefore **testable** in a wide variety of contexts. It has often been criticized that most scholars do not pay enough attention to practical concerns (such as recognizing good translation choices as well as errors, summarizing results in a way that clients can understand, or providing flexibility in the face of different texts) (Drugan 2013: 62f.), but rather "devise models from theoretical principles [and] then cherry-pick cases to test them" (Drugan 2013: 39). This is often done using student translations in what Drugan (2013: 156) calls an artificial situation, i.e. not in the everyday professional use of a LSP. However, even Drugan (2013: 36) herself admits that avoiding artificial situations is a difficult undertaking, since LSPs are mostly reluctant to try academic models. Either way, the fact remains that a TQA model, once created, needs to be tested in order to convince both theorists and practitioners. In the case of Colina's (2008; 2009) TQA model, the present MA project can be seen as (a small) part of that testing and made use of real-world, professional translations as test objects (see chapter 4.2).

A further characteristic that showed up in previous testing (Colina 2008; 2009) is the **ease of application** of the model. All raters found they had no major problems using the tool (Colina 2015: [245]), setting it apart from those models mentioned in chapter 2.2 that suffer from limited applicability due to how difficult they are to apply. In addition to that, potential

⁵⁵ This was put into practice in the Colina (2009: 244) study, where descriptive statements were added and/or deleted until there were four descriptors associated with each component, reflecting a four-point quality range for each criterion.

evaluators can easily learn the use of the tool by using training materials⁵⁶ developed along with the model.

3.3 Structure: Evaluation criteria

Now that the essential characteristics of Colina's (2008; 2009) model have been explained, a little more detail should be provided on the four assessment criteria. As set out above, the four criteria are components of quality that are examined in a translation, constituting a componential approach to TQA. The four assessment criteria in Colina's (2008: 128f.; 2009: 247ff.) model, and therefore in the replication study, are: Target Language (TL); Functional and Textual Adequacy (FTA); Non-Specialized Content (Meaning) (MEAN); and Specialized Content and Terminology (TERM).

The criterion **Target Language** refers to "the degree of adequacy of the language in the target text as an example of target language in use" (Colina 2015: [248]). It therefore concerns aspects such as spelling and grammar, clarity and comprehensibility of language use, whether or not the text sounds natural and whether it would be linguistically acceptable in the target language (Colina 2015: [248f.]). TL is one of two components which can and should be evaluated without referring to the ST (Colina 2015: [249]; *Hablamos Juntos* 2009: 5). Additionally, in order for the evaluator to gain as similar an impression of the TT as possible to a target reader, this criterion should be evaluated before the other three (*Hablamos Juntos* 2009: 5). It is also important to note that, like all evaluation criteria, TL must be examined relative to the brief, which might justify seemingly defect translation aspects like awkward-sounding language, an overly literal translation, or less than perfect grammar, depending on the function and context (Colina 2015: [249]).

The second component that can be evaluated by examining the TT alone is **Functional and Textual Adequacy** (Colina 2015: [249]; *Hablamos Juntos* 2009: 5). Functional Adequacy obviously means the degree to which the TT corresponds to the function as given in the brief, while Textual Adequacy refers to "the conventions (graphic, organizational and linguistic) that readers use to identify text types and genres" (Colina 2015: [253]). Aspects examined under this component include the text type and how well the TT exemplifies it, the

⁵⁶ In her article, Colina (2008: 106) points to the website www.hablamosjuntos.org for access to the training material; however, that website does not seem to exist any longer in 2022. According to the Research Institute of United States Spanish (RIUSS 2021), this is due to the cancellation of the *Hablamos Juntos* grant program. However, a copy of 'Part 6 of the More Than Words-Toolkit' that was developed to explain the nature and use of Colina's TQA model could be obtained from RIUSS. It remains unclear, however, whether this document is the same document or similar to the training materials that Colina (2008: 106) gave to her raters.

level of formality of language and the acceptability of the TT (for its specific function) in the target culture (Colina 2015: [253]).

The criterion **Non-Specialized Content (Meaning)** evaluates “how well the translated text conveys the meaning of the source text with respect to the translation brief” (Colina 2015: [263]). When examining the TT for MEAN, evaluators should pay special attention to omissions, additions or changes and distortions in meaning, where it is not connected to or follows immediately from specialized terminology that is incorrectly translated (Colina 2015: [263]). It should also be mentioned that MEAN should always be considered relative to the function and the target audience, which might justify certain additions, omissions or changes (Colina 2015: [263]).

The last criterion, **Specialized Content and Terminology**, obviously only applies to texts that feature some kind and amount of TERM (Colina 2015: [265]). The word *Content* should be highlighted in particular, because it expresses that the text is examined for specialization not only on the lexical level (terminology), but also on a textual level, “examining how it is captured in the target text, relative to the content of the original” (Colina 2008: 105). This is important, because deficiency in the TERM area may not only express itself as inadequate terminology, but also cause meaning and coherence problems, because TERM was not understood correctly in the ST or used wrongly in the TT (Colina 2015: [265f.]).

As mentioned before, each of the four criteria boasts a certain number of descriptive statements associated with it⁵⁷. Since evaluators are instructed to choose one descriptor from each component and use the score (numerical weight) associated with it for the quantification of results, there are four partial scores that are added up to one final score per evaluated text (Colina 2008: 107). As each criterion is assigned a total numerical weight, which might differ between criteria (see chapter 4.3), descriptive statements do not have the same numerical value across categories (Colina 2008: 107). Furthermore, the numerical scores are not given next to the descriptive statements, since pre-pilot testing by Colina (2008: 107) found that evaluators tended to judge according to numerical values rather than choosing descriptive statements, as they are supposed to.

⁵⁷ While in the modified version of the model used in Colina (2009: 244), each group of descriptors reflects a four-point range of quality for one criterion, it is not yet so in Colina (2008: 107), where some components have a greater number of descriptors than others. This is due to the fact that the quality continuum reflected in the descriptors is based on the quality continuum found in existing translations, which may differ from component to component (Colina 2008: 107). This was later changed in Colina (2009: 244) to even out the number of descriptors, just as the total weights (maximum score) was adjusted to add up to 100 (whereas in Colina (2008: 128ff.) the total maximum score was 85).

3.4 Empirical testing: Colina's (2008) original study

Two separate studies were conducted to empirically test the TQA model by Colina (2008; 2009). Only the 2008 study will be described in detail, since that is the study which served as the original study that was replicated in this MA project. Similarities and differences to the second study (Colina 2009) will be mentioned where they appear relevant.

The goal of Colina's (2008) study (referred to in this chapter as "the Study" for simplicity's sake) was to empirically test Colina's (2008) TQA model and to answer the following research questions:

- Can an evaluation tool, designed according to the theoretical principles stated above, produce reliable evaluation of translated documents?
- Are there differences in ratings that could be attributed to the qualifications of the raters?
- What are the qualifications of the ideal rater?
- Can this tool be used with ease by raters who are not experts in translation theory? Is meta-linguistic awareness a requirement for the use of this tool or others? Is training necessary and to what extent? (Colina 2008: 105)

In order to answer these questions, an experiment was conducted using the tool (to be found in Colina (2008: 128ff.)) described above. The evaluation conducted was criterion-referenced and summative in nature and had translated medical (health-education) texts as objects (Colina 2008: 104, 106). The Study featured three target **languages** (Spanish, Russian and Chinese), with all the STs being written in English (Colina 2008: 107).

3.4.1 Participants

The Study had 40 **participants** in total, who were each paid for one day of work (Colina 2008: 106) and of which three were cultural and language experts (called consultants), who provided advice on certain issues, such as the use of language varieties in Chinese (Colina 2008: 107ff.), and later also took part in the rating (Colina 2008: 118ff.). The other 37 participants were evaluators (called raters), who were recruited in three⁵⁸ groups: translators, teachers, and bilinguals (Colina 2008: 108ff.). Note, however, that bilingual raters were only available for the Spanish language, due to difficulties in finding sufficient numbers of bilinguals (matching the below-described criteria) for Russian and Chinese (Colina 2008: 110).

⁵⁸ As can already be guessed from the uneven number, neither participants nor groups were equally distributed across languages. The composition was: Chinese – 10 participants (5 translators, 5 teachers); Russian – 8 participants (5 translators, 3 teachers); Spanish – 19 participants (6 translators, 8 teachers, 5 bilinguals) (Colina 2008: 110).

Irrespective of the group, all raters had to possess native or near-native language proficiency in both the source language (SL) and the target language (TL) (Colina 2008: 108). Most raters also held a type of university or college degree, although this was not a requirement for two of the three groups and many participants held their degrees in fields other than translation or even language-related fields (Colina 2008: 110). Furthermore, few raters were competent in the medical domain and only two were medical interpreters (Colina 2008: 110). Presumably, this was acceptable, because the texts chosen for the Study did not feature much domain-specific terminology or content (see chapter 4.2). In addition to their language proficiency, the participants of each group had to correspond to certain criteria to justify their inclusion in the group. In the following paragraph these criteria will be briefly described, since they will serve to contrast the characteristics of students and novice translators as evaluators (discussed in chapter 4.4).

Participants in the **translator** group had to be “language professionals whose income comes primarily from providing translation services” (Colina 2008: 109). They also had to have some kind of degree or education in translation and/or a relevant field as well as significant professional experience (at least 5 years; in most cases more) and hold a membership in a professional organization (Colina 2008: 109, 112). Participants assigned to the group of **language teachers** were “individuals whose main occupation was teaching language courses, at the university or other educational institutions” (Colina 2008: 109). Some of them had some amount of translation experience, but did not rely on translation as a source of income (Colina 2008: 109). In contrast to translators and teachers, no formal (writing) proficiency in the TL was required for the group of **bilingual raters**, and they often had little or no formal schooling in the TL (Colina 2008: 109). Additionally, the group also included community interpreters (though without any formal training and translation experience) (Colina 2008: 109).

All 37 raters of the three groups were recruited to briefly learn the use of the TQA tool by Colina (2008; 2009) and then employ it to evaluate the chosen TTs. These were three texts (and their translations into each of the three TLs) classified as health education material (Colina 2008: 111) (for detail on text characteristics see chapter 4.2). Texts were purposefully not selected according to pre-determined (high or low) quality⁵⁹ in order to better test the capabilities and use of the tool (Colina 2008: 112).

⁵⁹ This was done later in Colina (2009: 243), where the rating session was preceded by a benchmark-testing session, for which half of the selected texts were very good on first reading, whereas the other half was of very

3.4.2 Procedure

The Study was conducted in the following steps (Colina 2008: 105f., 108):

- pre-pilot testing session
- training session
- rating session
- questionnaire

The **pre-pilot testing session** took place two weeks before the actual rating session and can presumably be seen as a test run, since the numerical results do not seem to have been analyzed (Colina 2008: 105f.). However, based on the feedback from the pre-pilot session, slight changes were made to the training materials and the tool itself: more examples were added to the training materials and numerical values were separated from the descriptive statements and moved to a separate score sheet (Colina 2008: 105f.). The latter was done in order for evaluators to judge a text solely by the best-fitting descriptor instead of by the numbers, while still allowing them to convert their results into numerical values at the end of the evaluation (Colina 2008: 106).

The **training session** consisted mainly of an internet/phone conference between the author and the raters and included a PowerPoint presentation as well as a question-and-answer session (Colina 2008: 106). PDF materials were also delivered to the raters via e-mail and were presumably available to them during their evaluation (Colina 2008: 106). Training materials covered the rationale for the creation of the TQA tool and a description of the components and their use, including examples of the descriptive statements (Colina 2008: 106).

The **rating session** itself started with the delivery of the STs and TTs as well as the tool worksheets to the respective raters via e-mail (Colina 2008: 106). This presumably took place around midday, since raters were allowed one afternoon and one night for their evaluation and were instructed to return the filled-in materials the next morning (Colina 2008: 106). They thus had approximately 20 hours to evaluate the three translations using Colina's (2008) TQA tool and record their partial and final scores and rating time⁶⁰. It is not

bad quality. This was done in order to find texts on which the raters agreed most in their evaluation, which were then used for the actual rating sessions.

⁶⁰ In the second study (Colina 2009: 244), raters had to complete one more step before turning in their materials: it was required that they complete the relevant sections of the cover sheet (not yet present in the 2008 version of the model) and tick or write an appropriate recommendation to the requester, another point that was added after the Study in 2008. It is also noteworthy that the first experiment had raters work with an 85-point total (possible)

known whether, during that time, raters had any way of contacting the author if the necessity presented itself.

Once their evaluation was completed, raters were asked to fill in a two-page **questionnaire**, which consisted of 19 questions (11 multiple-choice questions plus 8 open-ended questions) (Colina 2008: 131f.). The questions covered aspects such as how prepared raters felt to use the tool; which training materials had been the most helpful for them and what could have been improved; how confident raters felt about the score given for each text; whether they were able to use each criterion to rate the texts or whether they had difficulty distinguishing between descriptors and the quality levels they represent; whether they felt that important aspects of quality had been overlooked in the tool; what they liked about the tool and its underlying TQA approach; and how they felt about participating in such a study (Colina 2008: 131f.).

As will be explained in more detail in chapter 4.5, this procedure (minus the pre-pilot session) was also followed in the replication study.

3.4.3 Results and limitations

Qualitative data from the questionnaire and quantitative data from the rating results and time recordings were submitted to (statistical) analysis to answer each of the research questions given above⁶¹. The results of the Study will now be described in summarized form for each of the research questions.

“Can an evaluation tool, designed according to the theoretical principles stated above, produce reliable evaluation of translated documents?” (Colina 2008: 105; my emphasis) It would appear that the answer is yes, since inter-rater reliability (by text, TL and rater group) was between 0.79 and 0.88, which Colina (2008: 113) judges to be “good reliability”. She also mentions that a more focused study (fewer participants, more restrictive criteria) would likely produce still higher inter-rater reliability results (Colina 2008: 113f.).

“Are there differences in ratings that could be attributed to the qualifications of the raters?” (Colina 2008: 105; my emphasis) According to Colina’s (2008: 118) results, the answer would appear to be no, since no clear correspondence between rater qualifications and their respective ratings could be discerned. There were, however, certain tendencies among

score, whereas the weightings were adapted to allow for a maximum score of 100 in the second study (Colina 2009: 244), possibly to make the results more easily understandable for requesters.

⁶¹ Although the research questions were of course different, it is interesting to note that the results of the second study (Colina 2009: 235) were largely in line with those of the first Study, concerning points like the usability of the tool, the reliability and the aptness of raters. For more details see Colina (2009) as an in-depth comparison would go beyond the scope of this MA project.

rater groups that deserve to be mentioned. The most obvious trend is that teacher ratings and translator ratings were more similar to each other than to the ratings of the bilingual group, who (on average) gave the highest ratings and spent the least amount of time doing so (Colina 2008: 116ff.). An interesting fact is also that the one single teacher to not have a language degree performed similarly to the bilingual group, suggesting to Colina (2008: 118) that raters without formal education in the language field may appreciate the complexity of the task somewhat less. Concerning the ratings of teachers and translators, it can be said that teachers tended to rate somewhat higher than translators and also spent more time on their ratings⁶², which might be due to the formative nature of the grading that teachers normally engage in (Colina 2008: 118).⁶³

“What are the qualifications of the ideal rater?” (Colina 2008: 105; my emphasis) This being a very complex question, Colina (2008: 120) recognizes that the Study was not apt to find an answer to it, especially since the previous research question (qualification-attributable rating differences) did not provide a clear answer. However, the results suggest that, since bilinguals gave the highest ratings and also spent the least amount of time on their task, they might have less of an appreciation for the complexity of TQA and therefore might not make as good raters as members of the other two groups (Colina 2008: 120f.). This cannot be said with certainty, however, given that a bilingual group was only available for one TL (Spanish) (Colina 2008: 118), so that their rating behavior could perhaps be due to some cultural specificity rather than their qualifications.

“[1] Can this tool be used with ease by raters who are not experts in translation theory? [2] Is meta-linguistic awareness a requirement for the use of this tool or others? [3] Is training necessary and to what extent?” (Colina 2008: 105; my emphasis) The answer to question [1] appears to be yes, and consequently the answer to question [2] seems to be no, since questionnaire results showed that all raters (few of whom had training in linguistics or modern translation theory) understood what was expected of them and more than 95% felt confident in the ratings given (Colina 2008: 121f.). Raters also said in the questionnaire that they did not feel meta-linguistic awareness was a requirement for using the tool (Colina 2008: 122f.). However, Colina (2008: 122) admits that further research is

⁶² The Chinese teachers were an exception in this regard, since they took considerably less time to rate (on average). An explanation, according to Colina (2008: 117f.), could be that the one teacher whose ratings were more similar to the bilingual group (very short rating time compared to other teachers) was a Chinese teacher, which would have affected the average.

⁶³ This trend could also be discerned in the results from Colina (2009: 255), leading the author to suggest that both professional translators and language teachers would be equally suited to be evaluators and that their different rating behavior (slow rating by teachers, faster rating by translators) might make teachers suitable for formative evaluation (where feedback is required) and translators more suitable in time-sensitive contexts.

necessary to conclusively answer question [2], which also ties in closely with the unanswered ideal-rater-qualifications question, as one might point out.

Regarding question [3], it might be said that this also appears to be difficult to answer, since there was no control group without training whose results could be compared to the three rater groups who did receive training. Nevertheless, results showed that raters felt that the materials sent via e-mail as well as the training session via internet/phone conference were helpful and sufficient, for the most part, to prepare them for the use of the TQA tool (97.5% felt very prepared or somewhat prepared) and only 20% encountered unforeseen challenges while using the tool (Colina 2008: 122). These challenges were related either to the technology or to the division between descriptors (Colina 2008: 122). It is not mentioned in Colina (2008) whether any of the raters had previous experience with other TQA tools; however, at least for the language teachers and bilinguals it can be reasonably assumed that they did not, since TQA would not normally be part of their profession.

As has been hinted at before, the Study and the TQA tool both present some **limitations** that should now be mentioned to put the results in their proper context. Concerning the Study, the most important limitation concerns recruitment, which did not allow for equal numbers of raters across groups or for bilingual groups for Russian and Chinese (only Spanish⁶⁴) (Colina 2008: 112). One further limitation was the size of the experiment (37 raters, 3 consultants, 3 texts), which is why the second experiment in Colina (2009: 241, 243) included more raters (in the sense of raters who had not taken part in the first Study) and more texts (13 plus their respective translations). Another limitation that might be pointed out concerns the research questions, two of which do not seem to have been answerable through the methods employed in the Study (see above).

Furthermore, in their feedback through the questionnaire, raters pointed out one shortcoming of the tool itself: that descriptive statements represent fixed points in the quality continuum (Colina 2008: 123f.). Consequently, the tool does not allow evaluators to discriminate between fine points of quality, and raters found it sometimes hard to decide between two descriptors (Colina 2008: 123f.). Colina (2008: 123f.) suggests that, for future studies, more descriptors could be introduced to provide finer points of quality, a suggestion that was partially put into practice in her second study in 2009 (Colina 2008: 124; 2009: 259f.). Additionally, it is also possible to allow raters to add comments in places where

⁶⁴ This is the reason why the bilingual rater group was eliminated in Colina (2009: 242), since bilinguals would once again have been available for Spanish only.

they found a decision between two descriptors to be difficult, or to assign in-between numerical values when converting their results (Colina 2008: 123).

However, none of these suggestions was put into practice in the replication study in order for it to be as close as possible to the original Study (Colina 2008). The details of the replication study as well as any differences to the original Study will be laid out in the following chapter.

4 Replication study

4.1 Cornerstones

As mentioned in chapter 2.1, the cornerstones (purpose, method (type), object etc.) of every quality assessment study should be defined beforehand. This will now be done for the replication study, whose cornerstones largely match those of Colina's (2008) original study.

The evaluation **objects** were translation products, i.e. target texts (Colina 2008: 99; 2015: [246]), whose characteristics are described in detail in chapter 4.2. The **purpose** of the evaluation was twofold and depended on user types: from the translation evaluators' point of view, the aim was to evaluate the quality of translated texts, whereas from the perspective of the scientific researcher, the aim was ultimately to test the use as well as the usability of the TQA tool.

The assessment was carried out by **means** of a theoretically explicit, functionalist TQA model that consists of four evaluation components (criteria), namely:

1. Target Language
2. Functional and Textual Adequacy
3. Non-Specialized Content (Meaning), and
4. Specialized Content and Terminology (Colina 2008: 128ff.).

The evaluation itself was summative, in that the assessment results were not intended for the translator(s) and they did not receive any feedback (as opposed to formative assessment, where feedback is more common; see chapter 2.1). The results of the replication study did not affect the translator(s) of the selected texts, since their identity is unknown. The evaluation was also criterion-referenced, the criterion being the extrapolated translation brief (Colina 2015: [243]). Moreover, the assessment was carried out by human evaluators, i.e. neither the assessment nor any parts of it were automated. Neither did the assessment concern machine-translated texts (Colina 2011: 44)⁶⁵ – at least not to the knowledge of the translation commissioners, as detailed in Appendix IV.

Summarizing what has been set out here as well as in chapter 2.1, the replication study was carried out by means of a holistic (as opposed to error-based), summative, criterion-referenced, and human-conducted assessment that is a mixture of positive and negative

⁶⁵ Since nothing to the contrary seems to be stated in her work, it might be assumed that Colina's TQA model is intended for the assessment of human translations only, contrary to TQA tools like the J2450 metric (SAE International 2021) or the model by Carroll (1966). However, this point ultimately remains unclear.

evaluation (taking into account both positive and negative aspects) (see chapters 3.1 and 3.2 for details on this last point).

4.2 Texts

Colina's (2008: 107, 133) original study featured three English STs as well as their respective translations into the languages Chinese, Spanish and Russian. Accordingly, the replication study also featured three STs, which are written in German, as well as a translation into English of each ST. The language combination German-English was chosen because it did not feature in Colina's (2008) study (thus providing new data on the use of the model for this combination) as well as for ease of evaluator recruitment and due to language proficiency in both German and English of the author of this thesis (for text selection, correspondence etc.).

The three texts (ST plus corresponding TT) for the replication study were selected to be as close as possible in nature to the texts of Colina's (2008) study. Colina (2008: 111) originally chose her STs in accordance with the aims of the *Hablamos Juntos* project (texts from the medical field, translated for the general public etc.); and texts for the replication study were selected to be as similar as possible in order to reduce divergence in the results that could be attributable to differences in text type, text difficulty etc. Therefore, texts were chosen according to the following characteristics (extracted from Colina's (2008: 110ff.) description of texts used in her study):

- **Prerequisite:** Texts are available in the SL along with a translation into the TL.
- **Length:** Texts are between 270 and 1,200 words in length.⁶⁶
- **Text type:** Texts are public health and health education materials, i.e. reader-oriented texts intended for public consumption.
- **Topics:** Texts deal with medical topics⁶⁷.
- **Difficulty:** Texts are written for the general public, therefore featuring limited medical terminology (either explained in the text or terminology that can be reasonably assumed to be known to lay persons), simple syntax, relatively short sentences and occasional bullet-point lists.
- **Availability:** Texts are available online through the websites of government agencies and/or public organizations (called *institutions* after this point).

⁶⁶ The three STs chosen for the original study were respectively 1,200, 270 and 367 words long (Colina 2022). Texts matching these exact word counts as well as the above-named characteristics could not be found, therefore the word count in the replication study diverged somewhat from the original study (see Appendix VII).

⁶⁷ In Colina's (2008: 111) study, topics were the Polio Vaccine, Hepatitis B and Medicare's coverage for drugs, whereas all texts chosen for the replication study are about a Covid-19-related topic. This was not intentional, but rather a result of Covid-19 having been the most prevalent health care topic at the time of text selection.

Due to this last characteristic, texts for the replication study were selected from the Internet. It is important to note here that texts were not selected according to their quality, but exclusively on the basis of the characteristics listed above. Now that the characteristics applying to all selected texts have been laid out, each of the three texts and their translations⁶⁸ will briefly be described in detail.

Text 1: The first selected text originated from the website of the Austrian Integration Fund (ÖIF 2022a; 2022b), which offers information on topics surrounding Covid-19 (vaccination, safety measures etc.) and points to further sources of official information (websites, hotlines, information sheets etc.). According to correspondence with the ÖIF (2021), the website aims to provide information on Covid-19 for members of the general public with little or no German proficiency and was translated by a professional TSP. An excerpt made up of two consecutive sub-headings featuring information on the Covid-19 vaccination and compulsory vaccination was chosen as Text 1 (see Appendix III for the selected texts as they were presented to the evaluators). The German ST is 428 words long and the English translation is 531 words in length⁶⁹.

Text 2: The second text originated from the website of the Vienna Tourist Board (Wien Tourismus 2022a; 2022b). Amongst other things, this website offers tourists information about Covid-19 testing options in Austria and what to do in case of a Covid-19 infection. According to correspondence with Wien Tourismus (2022), the text in question was translated by professional external translators so that travelers can access Covid-related information in their own language(s). The part of the website dealing with the above-mentioned topics was chosen as Text 2, amounting to 608 words for the German ST and 735 words for the English translation.

Text 3: The third text, being a PDF-document, is available for download on the website of the Austrian Federal Ministry for Health and Social Affairs (BMSGPK 2021a; 2021b). The PDF document is an information sheet on how to correctly use FFP2 masks, directed at the general public with the aim of educating citizens and travelers from foreign countries. According to correspondence with the BMSGPK (2022), the information sheet was translated by professional translators. The German (ST) version of the sheet is 280 words long, and the English translation amounts to 308 words.

⁶⁸ The respective institutions behind the websites from which the selected texts have been taken were contacted to ensure that the English texts in question are indeed translations and not merely parallel texts (see Appendix IV).

⁶⁹ All words counts according to Microsoft Word.

4.3 Brief

As explained in chapter 3.2, Colina's (2008; 2009) model is functionalist, meaning that the evaluation is always conducted relative to specifications found in the translation brief. Concerning Colina's (2008: 111) first study (the original study), it remains unclear in what form the raters received the brief (e.g. as a text, as a list). For the second study, however, the tool was modified slightly to include a cover sheet (Colina 2009: 240, 258), on which the brief (to be filled in by the requester of the evaluation) is captured as a list. This list is to be filled in for the ST and TT (Hablamos Juntos 2009: 8), respectively, and includes the following points: Text Title, Text Type, Target Audience, and Purpose of Document. For the replication study, it was therefore decided to design a similar cover section on the first page of the tool sheet for each text, containing the above-listed points as well as the source of the text to serve as the translation brief for the evaluators to refer to during their evaluation.

Of course, the brief was not provided along with the texts on the websites where they were found, but had to be reconstructed⁷⁰ based on information inferred from the texts themselves (their topics and whom they address, the nature of the institutions on whose websites they were found) and from correspondence with the respective institutions (see Appendix IV). While not ideal, this approach is actually similar to what is done in practice, seeing that not all requesters will provide a brief for every translation (Kingscott 2007: 322; Williams 2001: 334), meaning that points like target audience or purpose have to be either deduced or obtained by contacting the client (Hague et al. 2011: 251; Williams 2001: 334).

Based on the Hablamos Juntos (2009: 8) training material, the text type for all selected German STs and English TTs is "Health education material". The Target Audience for the three STs and TTs, along with the Purpose of each Document is taken from correspondence as given in Appendix IV. The table on the next page presents a brief summary⁷¹.

⁷⁰ Colina (2008: 111) followed a similar procedure of reconstructing a brief, except that she relied on the experience of a professional translator familiar with the medical context to extrapolate a brief, rather than contact the institutions who created the texts.

⁷¹ Only the brief for the TTs was presented to the evaluators in the replication study, since it is the TTs being evaluated and there are no important differences beyond the language itself (German- vs. English-speakers).

	Text Title	Text Type	Target Audience	Purpose of Document
ST1	Informationen zur Corona-Schutzimpfung	Health ed. mat. (website content)	all German speakers in Austria	to inform/educate about Covid-19 topics in Austria
TT1	Information on the coronavirus vaccination	Health ed. mat. (website content)	all English speakers in Austria, especially migrants	to inform/educate about Covid-19 topics in Austria
ST2	Testmöglichkeiten und To-dos bei einer Infektion	Health ed. mat. (website content)	all German-speaking travelers in Austria	to inform about Covid-19 measures in Austria
TT2	Testing options and to-dos in the event of an infection	Health ed. mat. (website content)	all English-speaking travelers in Austria	to inform about Covid-19 measures in Austria
ST3	Coronavirus: So trage und verwende ich meine FFP2-Maske richtig	Health ed. mat. (information sheet)	all German speakers in Austria (citizens and travelers)	to inform about a Covid-19 measure in Austria (how to wear FFP2 masks)
TT3	Coronavirus: How to wear and use your FFP2 mask	Health ed. mat. (information sheet)	all citizens and travelers who do not speak German sufficiently well for the German version	to inform about a Covid-19 measure in Austria (how to wear FFP2 masks)

In close connection with the above-listed points that form the translation brief, the tool by Colina (2008: 107) also allows for weighting of the four constituent evaluation criteria (see chapter 3.3). In order for the requester to be able to prioritize certain aspects of quality and thus coordinate the tool with the specifications in the brief, the modified model (Colina 2009: 258; Hague et al. 2011: 256) even includes a criterion ranking option (1-4) on the cover sheet. The unmodified model used in Colina (2008: 128ff.), however, does not yet have a cover sheet. Nevertheless, Colina (2008: 107⁷²) describes how criteria in her 2008 study were weighted according to the translation brief.

On an 85-point scale, the criteria Target Language and Non-Specialized Content (Meaning) were each assigned a weight of 25 points, while the criterion Functional and Textual Adequacy was weighted with 20 points, and Specialized Content with only 15 points. This was done, because health education texts like the ones described in the previous chapter do not rely heavily on specialized terminology or phraseology, given that they address a lay

⁷² Note that the points assigned to the criteria at this point in Colina's (2008) article only add up to 80 points instead of 85, because the criterion Specialized Content is said to only feature 10 points. This is clearly a typing error, since the points in the tool itself (as shown in Colina (2008: 128ff.)) amount to 85 (Specialized Content having 15 points) and Colina (2008: 116) herself later mentions that rating took place on an 85-point scale.

audience (Colina 2008: 107). The non-specialized meaning of the text, however, is of great importance in this context, as is the expression in the target language, given that readers are expected to be able to act on their understanding of the text (Colina 2008: 107). In that regard, the functional adequacy of the TT is certainly of equal importance, but this criterion was assigned slightly less weight “to avoid duplication, as some deficiencies in meaning and target language norms will indirectly affect this category” (Colina 2008: 107).

It has been noted that, ideally, the weighting for each text type should be determined through empirical means (Hablamos Juntos 2009: 9), but since this goes both beyond the scope of the current MA project and runs contrary to its replicative nature, the above-described weighting found in Colina (2008: 130) was used for the replication study.

4.4 Participants

A total of four participants took part in the replication study. They were not paid for their participation. Contrary to Colina’s (2008) original study, participants were recruited as one single group⁷³ and there were no consultants. Recruitment took place via Facebook, via an announcement in the monthly newsletter of the students’ representatives at the University of Vienna’s Centre for Translation Studies as well as via direct e-mails.

All participants signed a consent form, agreeing that they would not be paid and that their answers to the questionnaires, their scores and time recordings as well as any comments on their part may be used for this MA project. As stated in the consent form, participants’ names were to remain unpublished and their contributions were anonymized.

Participants were also asked to fill in a pre-study questionnaire in order to determine some characteristics relevant to the replication study. These characteristics will now be described in the following paragraphs. As in Colina’s (2008) study, all participants claimed to either be native speakers of the SL or the TL or to have near-native **language proficiency**. Three participants were native speakers of German (SL) and one had acquired near-native proficiency of German. For the English language (TL), all participants claimed to have near-native language proficiency, with English being one of their working languages. Note, however, that it was necessary to rely on the self-assessment⁷⁴ of the participants regarding their language proficiency, since asking for or conducting language tests would have gone beyond the scope of this MA project.

⁷³ Given that Colina (2008: 109f.) differentiated her professional translator rater group from other rater groups mainly through the length of their professional experience (min. 5 years), no differentiation is necessary between students and recent graduates, since the level of their professional experience is almost the same (see below).

⁷⁴ How Colina (2008) determined the language proficiency of her participants remains unclear.

Another characteristic differentiating replication study participants from Colina's (2008) raters is that all participants had at least one language-related **university or college degree**, and therefore a formal education in translation and/or the SL and TL. Specifically, two participants held a BA in Transcultural Communication, of whom one held only the BA in Transcultural Communication and the other one also held a Master's degree in Translation and Interpreting. The other two participants respectively held a BA in English and American studies, and a BA in English Studies as well as Intercultural German Studies. As mentioned above, one participant had already graduated from the University of Vienna's MA "Translation and Interpreting" program at the time of the replication study, while the other three participants were at this time still enrolled in this same program.

At this point, it bears repeating that students are not to be seen as learners in an educational context, but as translators in training. Conversely, recently graduated (former) students are to be seen as novice translators with very little or no professional experience. Therefore, the preliminary questionnaire also asked how far along in their MA studies participants were and/or what professional experience they had. Results from the questionnaire showed that of the three participants still **enrolled in the MA Translation and Interpreting** at the time of the replication study, one was in his/her second, another in his/her fifth and the last one in his/her twelfth **semester**⁷⁵ (the minimum study duration being 4 semesters (UniVie 2017: 2)).

The characteristic '**experience**' was differentiated on three levels:

1. professional translation experience and membership in a professional organization,
2. experience in TQA, and
3. experience in medicine.

Only one participant claimed to have any professional translation experience (amounting to less than one year as a professional translator), while the other three participants did not have any professional experience. However, two participants had completed an unpaid internship in a translation-related context of two and three months of length, respectively⁷⁶. None of the participants held any membership (junior or full) in any professional translators' organization.

⁷⁵ Further inquiry revealed that this high semester count is due to the participant simultaneously studying in the MA program English Language and Linguistics.

⁷⁶ The preliminary questionnaire did not ask for the participants' income (compared to Colina's (2008) translator rater group, whose income came primarily from offering translation services), since this did not appear relevant to the replication study. Colina (2008) presumably included it in her rater characteristics mainly to differentiate translators from language teachers.

Neither could any participant claim previous experience in the field of TQA⁷⁷. Similar to Colina's (2008) raters, the replication study participants did not have any academic or practical experience in the field of medicine, which was not problematic, given that the texts to be evaluated are low in specialized content and medical terminology.

On the topic of experience it must also be pointed out that, as students and novice translators, the participants presumably do not (and cannot) have a higher or even equivalent degree of experience compared to the translators who translated the selected texts, as scholars like Kurz (2020: 52f.) frequently demand that evaluators should. Apart from the question whether this demand reflects reality, one might also ask the question whether this means that students and novice translators should never evaluate either their own work or their colleagues' work, simply because they are not (yet) experienced. One might further raise the point of how evaluators should ever gain experience, if even at the beginning of their career they are only supposed to evaluate translations by less experienced translators. Furthermore, Kurz' (2020: 53) demand begs the question of when the "significantly higher level of experience and expertise" stops: If a translator has 40 years of working experience, does the evaluator then need 50 years of experience? Assuming that this evaluator had to complete a relevant education, he or she would then likely be 75 years or older, and therefore long since retired. Furthermore, Colina's (2008) results seem to discourage the assumption that a higher level of experience is needed, since teachers (who did not have the same 5 years minimum experience as the translators) did just as well in the use of the tool. Whether translators would see an evaluation by someone less experienced than themselves as acceptable or fair is another matter that goes beyond the scope of this MA project.

One other factor that must also be mentioned when discussing evaluators is subjectivity. Subjectivity in an evaluation means that the evaluation results are influenced by the evaluators themselves and their personal and professional characteristics: age, gender, personality, language(s), skills and abilities, knowledge (domain-specific, translation-specific and general) and experience (Larose 1998: 163; Karoubi 2016: 272; Pöchhacker 1994: 123).

Subjectivity can be a problem in TQA, because evaluators may overlook errors, classify errors as non-errors or vice-versa, misjudge the severity and consequences of errors, or generally make incorrect assumptions about the text under evaluation (e.g. regarding features necessary for or detrimental to the attainment of its function) (Behr 2017: 579;

⁷⁷ Except that one participant had at one time engaged in proofreading within the context of a university course (without any specific assessment method the participant could recall), and that one other participant had conducted (but not participated in) an experiment for his/her own MA thesis, which involved the evaluation of post-edited translations by means of error categories.

Karoubi 2016: 272; Thelen 2009: 204; Williams 1989: 13). Evaluators will also have different quality expectations, meaning that some will show a tendency towards higher and some towards lower scores, and will differ in how much of the available rating scale they use (some evaluators never assign the highest possible or the lowest possible rating) (Koehn 2020: 49).

Nevertheless, scholars agree that subjectivity is an inherent fact of TQA and thus is practically unavoidable (Bittner 2020: 10; Karoubi 2016: 272; Lauscher 2000: 163; Secară 2005: 39). As Bittner (2020: 67) so aptly puts it: “Translation quality assessment is essentially a human affair. And since human beings are intrinsically fallible, there is no possibility for TQA to eliminate [...] judgemental variance.” However, to entirely eliminate subjectivity may not be altogether desirable in any case, since it is human judgement that makes TQA models different from rigid and unthinking, automated evaluation (Larose 1998: 181). In fact, automated evaluation results (e.g. from matrices such as BLEU) are usually judged against human evaluation results, establishing human subjectivity as the yardstick against which (presumably objective) automated evaluations must measure up (Koehn 2020: 60). Incidentally, an analysis mentioned in Koehn (2020: 60ff.) shows that there is often a high correlation between human evaluation results and automated results (at least on restricted domains and language pairs), showing that the subjectivity problem need not always generate great discrepancy.

Furthermore, saying that subjectivity is unavoidable in TQA does not mean that it cannot be reduced or, conversely, that objectivity cannot be increased (Bittner 2020: 73f.). Some scholars seek to further objectivity by basing their TQA approach on empiric data (e.g. corpus-based approaches like (Bowker 2001) or the CDI method (Eyckmans et al. 2009)), others are partial to the controversial idea that quantification (i.e. the use of number-based error-counting metrics) can curb subjectivity (Kingscott 2007: 323, Williams 2001: 335), while many more scholars agree that a systematic approach and well-defined evaluation criteria contribute to a significantly less subjective assessment (Bittner 2020: 73f.; Rodríguez Rodríguez 2007: 6). Yet another way to increase objectivity, pointed out by Bittner (2020: 70, 74) and Kurz (2020: 51), is to render the assessment interindividual, i.e. to have the assessment done by more than one person and to (potentially) make it as closely reproducible as possible to further intersubjective agreement (inter-rater reliability). This is presumably why Colina (2008) analyzed her results for inter-rater reliability.

4.5 Procedure

The procedure of the replication study was designed to follow as closely as possible the procedure of Colina's (2008) original study (see chapter 3.4.2) and consisted of three steps: a training session, a rating session, and the answering of a questionnaire, all of which took place on the same day in order not to impose unduly on the participants. Contrary to Colina's (2008) study, the replication study did not include a pre-pilot session, because Colina (2008) did not provide sufficient information (What did the pre-pilot session entail?; Who participated?; What were the goals and what data was obtained?) to properly replicate this step. The three steps of the replication study will now be described in the following; for the recruitment procedure preceding them, see the previous chapter (4.4).

The **training session** consisted of a Zoom meeting (audio and video, ~40 minutes) between the author of the thesis and the participants. It involved a PowerPoint presentation followed by the opportunity for questions and answers. The PowerPoint presentation as well as a checklist⁷⁸ were later made available to the participants to refer to during their evaluation. The PowerPoint presentation covered a brief summary of the replication and the original study, the tool and its most essential features and criteria as well as the evaluation corridor⁷⁹. At the end of the presentation, an example was included for participants to try out the tool on a very short sample text with deliberately inserted errors (as recommended by Kurz (2020: 55)). The sample text was a PDF document taken from the ÖIF (s. a.a; s. a.b) and presented with a translation evaluation brief inferred from the context (not from correspondence).

The **rating session** took place on the afternoon of the same day as the training session and was designed to obtain the data necessary to answer research questions 2 and 3 (and partly 1). All necessary materials, including texts, tool sheets plus score sheets, the checklist (see Appendix V) as well as the training presentation for reference and the post-study questionnaire, were sent to the participants via e-mail. They were instructed to do the following: use Colina's (2008) tool to evaluate three translated texts for quality (RQ-1);

⁷⁸ The checklist can be found in Appendix V. It was given to participants to follow during the rating session and included all necessary steps for them to take (which materials to use in what order, when and where to record start and finish times, how to calculate scores, which documents to return at the end etc.). The instructions contained in the checklist were drawn up based on information found in Colina (2008; 2015: 5). In Colina's original study (2008: 128), rater instructions were given in brief directly on the tool sheet. For the replication study, these instructions were removed to avoid confusion among the raters, since the original instructions did not fit the replication study (English having been the SL in Colina's study, but being the TL in the replication study). All non-language-specific instructions given in Colina (2008: 128ff.) were included in the checklist.

⁷⁹ *Evaluation corridor* means "all the actions, processes and the very specific method of *how* to conduct the evaluation" (Kurz 2020: 51; emphasis in the original).

record their partial and final scores as well as the time necessary to evaluate each text (RQ-2); and to return their filled-in materials by 10 a.m. the next day. Within this time limit, participants could choose to do the evaluation in their own time. During the evaluation, participants were allowed to contact the author of the thesis (but not each other) if the necessity presented itself, although none of them did. Participants were further allowed to make comments, to use any source (online or hard-copy dictionary, search engine for parallel texts, websites etc.), and to choose whether they wished to evaluate all three texts in a row or take breaks in between the evaluation of each text. That being said, participants were discouraged from taking breaks during the evaluation of any one text and instructed to make a note of substantial interruptions (in order to not falsify the time recording).

After the end of the rating session, participants were asked to fill in a three-page **questionnaire** consisting of 13 questions. The questionnaire (to be found in Appendix VI) is the research instrument designed to answer RQ-1 and RQ-3, i.e. the research questions concerning the ability of MA students and novice translators to use the tool, problems they might have encountered, and the attitude of MA students and novice translators towards the application of Colina's (2008) model. The questionnaire was constructed using guidelines found in Dörnyei (2010)⁸⁰ and was to be answered in English, so that no translation issues would arise when analyzing and using the results. It is loosely based on the questionnaire given to Colina's raters in the original study (Colina 2008: 131f.), but includes many new questions to answer RQ-1 (questions 1-8) and RQ-3 (questions 9-12), while ignoring some of Colina's (2008: 131f.) questions (e.g. regarding the *Hablamos Juntos* grant program, finer points on the training materials etc.). Like Colina's (2008: 131f.) questionnaire, it includes both multiple-choice and open-ended questions (4 pure MC-questions, 1 purely open question and 8 combined questions that ask participants to choose one answer and then explain why they chose it or fill in other text⁸¹).

The questionnaire was sent to the participants along with the evaluation documents in the form of a Microsoft Word document, which had fill-in-only protection on it (meaning boxes could be ticked and indicated text fields filled in, but neither the questions nor the answer texts could be altered by accident). Participants were instructed to fill in the

⁸⁰ Although Dörnyei (2010) is intended to aid with the construction of questionnaires in second language research, it is a very comprehensive work applicable to many kinds of questionnaires and contains many helpful guidelines regarding question format, questionnaire length and possible pitfalls in questionnaire design.

⁸¹ For comparison, Colina's (2008: 131f.) questionnaire included 19 questions (11 MC- and 8 open questions), if each question is counted only once. The number of questions in the questionnaire is actually 23, since the last four questions feature twice due to what appears to be a copy-and-paste error.

questionnaire only after completing the evaluation and to return it (either as a filled-in Word document or in the form of a photo or scan) within the set time limit.

This concludes the description of the replication study and its procedure. For a side-by-side comparison between the original study by Colina (2008) and the replication study, see Appendix VII. The results of the replication study as well as limitations and the answers to the research questions will be presented in the next chapter.

5 Results and discussion

5.1 Limitations

Before the results obtained from the replication study are given below, the limitations that the replication study is subject to should be mentioned in brief. One basic limitation is the artificial situation the replication study took place in (compare chapter 3.2), but this is due to its replicative nature, since the same is true for Colina's (2008) original study.

That being said, the replication study was not a perfect replication of Colina's (2008) original study, since some aspects of the original study were not described in detail (e.g. pre-pilot session, see chapter 4.5) and some materials could not be obtained (most notably the original training materials). The replication study was also much more limited in scope than the original study, partly because only one language pair (German-English) was examined, but primarily due to limitations on recruitment. Most importantly, only four participants took part in the replication study (compared to 37 raters in Colina's (2008) study), and they could not be paid for their participation. Consequently, text length and questionnaire length had to be limited in order to keep the participants' efforts within reasonable bounds. Originally, the replication study included a fourth research question involving inter-rater reliability (since inter-rater reliability also featured in the original study), but this research question had to be dropped, since it would not have yielded any meaningful results with only four participants. Moreover, as detailed in chapter 4.4, all participants were students or former students of the Master's program at the University of Vienna's Centre for Translation Studies (and therefore Austrian students); when carried out in another country with a different translator education, a replication study might yield different results.

Some limitations also apply to the research instruments described in chapter 4.5, especially the questionnaire. Answers given in the questionnaire are subject to bias, whether this be intentional (desire to present oneself in a better light) or unintentional (self-deception) (Dörnyei 2010: 8f.). Possible phenomena in that regard include acquiescence bias (tendency to say yes if one is unsure), the Halo effect (tendency for a generally positive or negative impression to overshadow any contrary aspects) and the fatigue effect (participants' answers becoming more inaccurate towards the end of the questionnaire) (Dörnyei 2010: 8f.). The last point was probably mitigated somewhat by the shortness of the questionnaire (13 questions, ~3 pages), but cannot be entirely disregarded nonetheless. Now that the relevant limitations have been mentioned, the results of the replication study will be given and discussed in detail in the following chapters.

5.2 RQ-1: Can Austrian MA-level translation studies students and recently graduated former students (novice translators) use the TQA model by Colina (after receiving training and materials) without encountering problems Colina’s more experienced evaluators did not have?

H-1.1: Students and novice translators will be able to use Colina’s TQA model to evaluate the quality of translated texts, since they will receive training and materials and, presumably, no specialized knowledge of translation theory is necessary in order to use the model.

H-1.2: Replication study participants will encounter some problems that more experienced translators in the original study did not have, especially concerning the matching of descriptive statements to the evaluated texts.

All four evaluators who took part in the replication study were able to complete the evaluation, i.e. to provide partial scores and total scores for all categories, for all texts. Apart from confirming that the model can, in principle, be used by novice translators, this also suggests that the TQA model can be successfully applied to the language combination German-English (which had not been tested before). All participants further stated in the questionnaire that they felt entirely prepared to use Colina’s (2008) model for the evaluation of translated texts after the end of the training session, a sentiment similar to what Colina’s (2008: 122) raters expressed, 97.5% of whom felt very prepared or somewhat prepared. One evaluator in the replication study even specifically mentioned the ease of application that he/she experienced when using the model.

Questionnaire responses concerning the evaluators’ confidence in the scores they gave to each of the texts further support what has been said above: three or more out of four participants felt either very confident or somewhat confident in their score for each text, and only one participant did not feel confident regarding one of the texts (see Table 1). Compared to that, 95% of Colina’s (2008: 121) participants claimed to feel very confident or somewhat confident and 5% did not feel confident in their scores⁸².

⁸² As shown in Table 1, one participant also failed to answer the question regarding Text 1 in the original study.

Table 1: Participants' confidence in scores (partially adapted from Colina (2008: 121))

	Replication study (n=4)			Original study (n=40)			
	Very confident	Somewhat confident	Not confident	Very confident	Somewhat confident	Not confident	No answer
Text 1	50%	50%	–	65%	32.5%	–	2.5%
Text 2	25%	50%	25%	62.5%	37.5%	–	–
Text 3	25%	75%	–	67.5%	27.5%	5%	–

Results from the replication study further support the conclusion that Colina (2008: 122) drew after her study in 2008: that no specialized translation or linguistic knowledge seemed necessary in order to use Colina's (2008) model. When asked about problems they had and what reasons they believed these problems to be due to, none of the replication study participants checked the option "insufficient specialized knowledge of translation theory", nor did they make any comment in that regard.

This data seems to confirm H-1.1: students and former students (novice translators) were able to use Colina's (2008) tool to evaluate the quality of translated texts after they had received training and materials, and none of the four participants felt that specialized translation knowledge was necessary in order to use the model.

However, while two participants claimed to have encountered no problems at all during the evaluation, the other two participants stated that they had indeed had problems. One participant stated that he/she had trouble "finding mistakes or problems that could account for different descriptive statements other than respectively the best one" (participant 1) in any of the four categories, attributing this to his/her lack of professional translation experience and lack of experience with TQA. The participant further commented that he/she often felt insecure about his/her quality judgements of texts, and this was indeed the same participant who had stated that he/she only felt somewhat confident of his/her scores for two of the three texts and not confident of the scores given to the third text. This was the lowest confidence-rating among all the participants, since two of the three other participants claimed to be very confident of the scores for at least one text and somewhat confident of the other two scores, while one participant even felt very confident of two scores and somewhat confident of the third.

Participant 1's insecurity may, perhaps, be explained by the fact that he/she was only in his/her second semester of the Master's program, whereas the other participants were in their fifth and twelfth semester or had already graduated. In this respect, H-1.2 seems to have

been accurate, since more experienced evaluators would likely be more secure in their judgement and therefore not encounter this problem. It should also be noted that the participant did not tick the option that his/her problem had been due to “the texts to be evaluated”; however, this might, in fact, apply, since the texts, although chosen solely according to the criteria laid down in chapter 4.2 and not according to a particularly high level of quality, seem to have been relatively high-quality texts (no score was lower than 65 out of 85). Had other (lower-quality) texts been chosen as evaluation objects, then the participant may not have encountered this problem. Furthermore, it seems doubtful whether this constitutes a real problem, since the evaluation of a high-quality text would perforce involve ticking the best or some of the best descriptors. This participant’s troubles therefore seem to have been mainly due to his/her own insecurity, without actually having a negative influence on the evaluation itself, and even the participant him/herself judged the problem to be minor in nature.

By contrast, the other participant of the two participants with problems claimed to have encountered problems that he/she categorized as major, feeling that these problems prevented him/her from effectively using the tool to evaluate texts. While the problems apparently did not prevent the participant (participant 2) from completing the evaluation, they might account for the fact that the participant took much longer to evaluate the three texts than all the other participants did (see RQ-2). As comments made by the participant show, he/she had two main problems.

The first problem was related to the categories Functional and Textual Adequacy and Non-Specialized Content (Meaning), since the participant apparently had difficulty distinguishing between them. The participant stated that he/she thinks these two categories “are inherently intertwined, especially for informative texts like the ones we had to evaluate. [...] That’s why I found it strange and counter-intuitive to evaluate meaning and fulfilment of function separately.” (participant 2) The participant further commented that he/she could see little sense in evaluating the categories in question one after the other and would, presumably, have preferred some kind of combined or overarching category. Whether this would be the case for other types of texts as well (not ‘informative texts’, as the participant calls them) and what this combined category should look like in the participant’s opinion remains unclear.

Regarding the division between categories, Colina (2008: 107) herself recognized them as not being absolute, stating for example that deficiencies in the categories Meaning and Target Language will indirectly affect the category Functional and Textual Adequacy. This is why the latter was assigned slightly less weight in the scoring table, to avoid

duplication due to this indirect influence (see chapter 4.3). What Colina thereby recognizes, is that no category can ever be completely separate from the other evaluation categories, since they all reflect aspects inherent in a text. To take an example, the category Specialized Content and Terminology (one that participant 2 did not seem to have had any problems with) could equally have an influence on Meaning, since incorrect terminology would undoubtedly convey an erroneous meaning to the reader. One might further add that a distinction between Functional and Textual Adequacy and Meaning would be justified for any number of errors such as a culturally unacceptable image in the TT, which would prevent readers in the target culture from reading the text (thereby hindering the fulfilment of its function), while having nothing to do with the meaning of the text at all. It would therefore seem as though this problem encountered by participant 2 is due, in some measure, to the texts chosen for the evaluation, since texts containing problems in other categories (e.g. Specialized Content and Terminology) might have made it clear to the participant that a division between these categories can indeed be useful. Nevertheless, in spite of the participant's claims, it remains unclear in what way exactly this problem could be considered major (and prevent an effective evaluation), since the participant was obviously able to choose descriptors and give scores, having submitted a complete score sheet.

It should also be added that participant 2 generally seems to have had trouble deciding which aspects are relevant for which evaluation category, as some of the comments reveal. The participant made a comment, for example, regarding one instance of Text 1, where the same sentence was unnecessarily repeated in the ST, and the TT corrected this problem by only mentioning the sentence once⁸³. The participant made his/her comment in the column of the category Non-Specialized Content (Meaning), which was the correct category for this type of issue, given that the instance described above clearly constitutes an omission, which falls under the category Meaning (Colina 2008: 129; 2015: [263]). This was unambiguously described in the training session, and since participants had access to the training materials, the participant could have looked this up; instead, he/she seemed to think that the instance described above should have been evaluated under the category Functional and Textual Adequacy. This shows that a problem existed regarding the attribution of phenomena observed in the text to the relevant evaluation category, a problem the participant attributed to

⁸³ This may also serve as an excellent example of positive evaluation: the participant clearly observed a divergence between ST and TT and understood that it was an omission of a part of the ST material, but did not judge this to be an error, since the TT in fact corrected a shortcoming of the ST. Instead, the participant recognized the TT as superior to the ST in this respect, and therefore made a comment on it that would highlight an additional positive aspect to commissioners.

the tool itself or a part thereof. To counteract this, future studies might develop an adapted version of the model that mentions the most relevant aspects to fall under a category right next to the category in the scoring sheets (e.g. for Non-Specialized Content (Meaning) the aspects mentioned could be: conveyance of ST meaning, additions, omission and distortions of meaning). This way, commissioners might also place special emphasis on some aspect that they want particular attention paid to in the evaluation.

This leads more or less directly to the second major problem that participant 2 encountered. He/she complained of the fact that raters were instructed to evaluate the first two categories (Target Language (1) and Functional and Textual Adequacy (2)) without looking at the ST and to only use the ST for the other two categories (Non-Specialized Content (Meaning) (3) and Specialized Content and Terminology (4)), as recommended in the workflow for the model in Colina (2015: [266]). The participant observed that he/she found multiple mistakes belonging in the category FTA only while evaluating the category Meaning, because he/she then first compared the TT to the ST. The participant further commented that it made little sense to evaluate these first two categories without reading the ST, since some apparent mistakes in the TT falling under one of these categories (e.g. ungrammatical use of language by a character in a novel) might be explained and justified by the ST, but be regarded as an error when evaluating e.g. Target Language without the ST. Similarly, the participant voiced the opinion that, if a translation is supposed to mirror the function of the ST (as was the case for the texts under evaluation, which participants could deduce from the translation brief), “[the function of the translation] is tied closely to this ST and how it implements this function.” (participant 2) The participant therefore repeatedly mentioned that he/she had problems with evaluating the first two categories and also attributed his/her troubles specifically to these two categories in the questionnaire (whereas the other participant with problems did not attribute them to any category in particular).

Compounding what has been described above, the participant seems to have failed to pay attention to the part of the training session where it was mentioned that an evaluator could, of course, switch back and forth between categories, if he/she found a relevant aspect belonging to a previously examined category later on in the evaluation. The participant frequently made comments such as “[T]his probably also belongs to ‘Functional and textual adequacy’, but I detected this with the ST only.” It was certainly mentioned during the training session that evaluators were free to adjust any evaluation of any category at a later time during their evaluation of this same text. It remains unclear whether this point should have been further highlighted in the checklist (which stated which categories to use only the

TT or both the ST and TT for), whether it was somehow ambiguously explained in the training session, or whether the participant simply failed to hear this or forgot. Since none of the other participants made any mention on this in any of their comments, one might assume that the problem rather lay with participant 2 than with the training material, but this cannot be said with certainty. Whatever the cause, participant 2 (rightfully, from his/her point of view) complained that he/she was stuck with the rating for a category once given (believing that evaluators were not allowed to switch back and forth between categories) and that, if he/she had found some errors when evaluating an earlier category, his/her rating for this earlier category might have been different.

It is understandable that the participant classified this as a major problem that would definitely have had an influence on his/her rating and might even have prevented an efficient evaluation process, since thinking about this would have taken the participant some time that could have been spent otherwise. It is also a problem that none of Colina's raters seem to have encountered, although the same instructions (to not read the ST until evaluating the last two categories) were given to them in the original study (Colina 2008: 128).

Apart from these major issues that participants mentioned in the questionnaire when asked about problems they had, the comments also reveal some minor troubles, mainly related to deciding between different descriptors. Three of the four participants mentioned at least once that they were unsure in their decision between one descriptor and another in some category, although no pattern could be discerned as to any particular category or any particular text. It is interesting to note that this is a problem shared by some of Colina's (2008: 122) raters, 20% of whom encountered unforeseen challenges (an unknown percentage of these having to do with the divisions between descriptors). Even though the uncertainty in deciding between which descriptor best fits a given text therefore seems not unique to students and novice translators, it is also clear that they had decidedly more trouble with this, since three out of four replication study participants mentioned such a problem at least once, whereas less than 20% of Colina's raters⁸⁴ encountered the same challenge.

It would therefore seem as though H-1.2 could be confirmed, since students and novice translators did indeed encounter problems that their more experienced counterparts in Colina's (2008) study did not have. However, it is unclear how great an influence professional

⁸⁴ It is unknown how many raters exactly had trouble with this in the original study, since Colina (2008: 122) only mentions that "20% encountered unforeseen challenges (some having to do with the technology, others with the division between descriptors)". It is also not specified in the article what problems exactly were related to technology, and whether this meant that raters somehow had trouble using e.g. a PDF-reader (since materials were sent to them in PDF form) or whether technical problems occurred that lay beyond the control of the raters and was unrelated to their capabilities.

experience as a characteristic has, since at least some of the problems encountered by replication study participants (namely not hearing or forgetting that switching between categories was possible and that the first two categories could be re-evaluated with new aspects gleaned from the comparison between ST and TT) might have happened to any evaluator, regardless of professional experience. Furthermore, no clear relation between problems encountered and (non-professional) experience of the replication study participants could be discerned either. Both the least experienced participant (two semesters in the Master's program) and the participant who had already graduated (and thus had, arguably, the greatest experience) were the ones to encounter problems, whereas the two participants of middling experience (five and twelve semester in the Master's program) did not have any problems. This might be due to the fact that the participant who had graduated judged according to different standards (classifying problems as major that the other participants would have seen as minor) or that experience (of the non-professional nature) did not have much of a correlation with problems the evaluators had.

Furthermore, the part of the hypothesis (H-1.2) where it was assumed that problems encountered by students and novice translators would be related to the matching of descriptors to the evaluated texts, could not be entirely confirmed. While participants seem to have indeed had some trouble in this area, it was in no way a major problem, nor was it the most serious of the problems encountered. It does, however, seem to have been the most widespread, being the only challenge that was, directly or indirectly, reported by more than one participant and indeed by the majority of participants (three out of four).

That being said, it deserves to be mentioned that students and novice translators also seem to not have had some problems that Colina's raters were confronted with, namely technical problems, which none of the participants in the replication study complained about. However, a closer discussion of these technical difficulties in the original study is not possible, since it was not detailed how they manifested or how much of an influence they had on the evaluation (see footnote 84).

Following this discussion, a summary of all that has been said above and an answer to RQ-1 may now be attempted. The results of the replication study show that all participants were able to complete the evaluation and that all participants felt entirely prepared to use the model after they had received training on it. This clearly shows that Colina's (2008) TQA model can indeed be used by students and novice translators. At least three out of four participants also felt very or somewhat confident of their scores, and only one participant did not feel confident of the score he/she gave to one of the texts. All this, in combination with

the fact that no problems were attributed (by the participants) or attributable to a lack of theoretical knowledge of translation, clearly proves H-1.1, since students and novice translators were able to use the model and no evidence could be found pointing to the necessity of specialized knowledge of translation theory.

However, while all participants successfully completed the evaluation, only half of the participants did so without problems. One of the two participants to encounter a problem (participant 1) had trouble finding errors in the texts that might have enabled him/her to choose descriptors other than the best ones, a difficulty that seems rather to stem from the texts themselves (being, apparently, of high quality) than originate with the participant, who classified the problem as minor and attributed it to his/her lack of professional translation experience and lack of experience with TQA. However, it is interesting to note that this participant described him/herself as generally insecure regarding quality ratings of texts, and this was also the participant with the lowest confidence rating in scores. Therefore, this proves H-1.2 insofar as more experience in an evaluator would quite probably have led to greater confidence and would have prevented this problem.

The other participant who reported troubles and who took longest to complete the evaluation (participant 2) classified his/her problems as major, although they obviously did not prevent the participant from completing the evaluation. Participant 2 apparently had trouble separating the first two categories, being of the opinion that they should be one category. As discussed above, Colina acknowledges the influence that categories may have on each other, and it may only be presumed that the necessity of having separate categories would have been made clearer to the participant if other texts (containing errors in other categories, highlighting their interrelation or lack thereof) had been chosen. Contrary to the second problem encountered by the same participant, it is unclear in what way this constitutes a major problem, although participant 2 seemed to experience a general difficulty in attributing phenomena found in the text to their correct categories, a problem the participant felt was caused by the tool itself.

Participant 2 also complained that, for multiple reasons, it makes no sense to evaluate the categories Target Language and Functional and Textual Adequacy without reading the ST. That is certainly why the participant attributed this problem entirely to these two categories (TL and FTA). This problem was further compounded by a mistake on the participant's part, who did not hear or failed to remember that switching between categories was allowed and therefore wrongly believed that he/she was stuck with a descriptor once chosen and could not choose another if the evaluation of a later category (involving the ST)

revealed aspects relevant to a previous category. This explains why the participant classified this second problem as major. Nonetheless, participant 2 was the only participant to encounter a problem of this sort (no other participant in the replication study reported any such issues, and neither did Colina's (2008) raters).

Apart from these problems detailed by the participants, some minor troubles could be discerned from the comments, notably that three out of four participants had at some point trouble deciding between one descriptor and another. This is a challenge encountered not only by students and novice translators, although Colina's (2008) participants seem to have had less trouble with it (less than 20% reported this).

Apart from the challenge mentioned last, the problems that participants in the replication study encountered prove H-1.2, since there were problems among them that would, presumably, not have happened to more experienced translators. However, it is unclear how much of an influence the factor experience (both professional and non-professional) has, since some problems were not immediately due to lack of experience and no correlation could be discerned between non-professional experience and problems (or lack thereof). Furthermore, a part of the hypothesis was proven only partially true, since difficulty in deciding between descriptors was the most widespread issue (encountered by three out of four participants), but neither the most serious nor did it prevent a successful evaluation.

The answer to RQ-1 therefore presents itself as follows: students and novice translators were indeed able to use Colina's (2008) TQA model to evaluate the quality of translated texts, but they were unable to do so without encountering problems that more experienced translators in the original study did not have. It bears repeating nonetheless that some (technical) problems Colina's (2008) raters reportedly encountered, students and novice translators did not have.

It is also worth a note at the end of the discussion for this research question that none of the problems mentioned by the participants seemed to be related to either the German or the English language specifically, suggesting that Colina's (2008) model can be effective for the language combination German-English, which had not been tested before in either the first or the second original study. However, larger studies would be needed to confirm this.

5.3 RQ-2: How does the productivity (measured as rating time in minutes) of MA students and novice translators compare to Colina's groups of evaluators?

H-2.1: Students and novice translators will rate more slowly than Colina's professional translators did, due to their lack of professional experience.

H-2.2: Students and novice translators will rate as fast or even faster than Colina's language teachers did, since language teachers rated slower than professional translators and, presumably, spent more time reflecting on their evaluation.

As indicated in the research question, the evaluators' productivity was measured as the time needed by each participant to evaluate one text from start to finish (see Appendix V). The results of the time recordings made by the participants are given below in Table 2, along with the corresponding productivity of each participant expressed as words evaluated per hour. Regarding, at first, only these results from the replication study, they can best be interpreted when looking at the time all participants together needed to evaluate each text (total time per text) and the time each participant took to evaluate all three texts (total time per participant). This will be done in the following paragraphs, before a comparison with the results obtained from the original study will follow.

Table 2: Evaluation time per text and participant (in minutes) as well as productivity in words per hour (indicated between brackets) in the replication study

	Text 1 (428 words)	Text 2 (608 words)	Text 3 (280 words)	Total time per participant
Participant 1	33 (778)	21 (1,737)	20 (840)	74 (1,067)
Participant 2	69 (372)	70 (521)	44 (382)	183 (431)
Participant 3	20 (1,284)	18 (2,027)	11 (1,527)	49 (1,611)
Participant 4	25 (1,027)	40 (912)	25 (672)	90 (877)
Total time per text	147	149	100	

Total time per text: The time needed by all participants together to evaluate Text 1 and Text 2, respectively, is almost identical (difference of 2 minutes), whereas the total time for Text 3 is much shorter (see Table 2 above). This latter result is not surprising, given that Text 3 was the shortest of the three texts with 280 words⁸⁵. It might seem surprising at first that participants took almost as much time to evaluate Text 1, which was 428 words long, as they needed to evaluate Text 2, which had 608 words and was therefore almost 40% longer

⁸⁵ Note that the word count given is the length of the source text, not the target text. The ST length was used in order for the replication study's results to be comparable with those of the original study, since Colina's (2008) article does not feature TT word counts, and also because ST word counts are more easily comparable across multiple texts in each TL.

than Text 1. However, this discrepancy might be explained when taking into account that participants had to use the evaluation model for the first time ever when working on Text 1, whereas for Text 2 they already had a little more routine concerning the work steps, the categories, the scoring procedure etc.

However, this is only true when considering the total time spent on each text, but not so for all evaluators individually. While two out of four evaluators spent more time on Text 1 than on Text 2, one participant needed almost the same amount of time for both texts (only 1 minute longer for Text 2) and the last participant spent more time on Text 2 (15 minutes more than on Text 1). This indicates that some participants perhaps needed more time to get used to the model than others and that, for some participants, their evaluation time correlates more closely with the word count of the evaluated text. That is why it is also interesting to have a look at the total amount of time that each participant spent working on the evaluation of all three texts together and the resulting productivity expressed as words per hour.

Total time per participant: It becomes clear, when looking at Table 2 above and Figure 1 below, that participant 3 was the fastest evaluator both in terms of total evaluation time (only 49 minutes for all three texts, equaling 1,611 words per hour) and for each of the three texts individually (e.g. only 11 minutes for Text 3, equaling 1,527 words per hour). In contrast, participant 2 was by far the slowest evaluator; he/she needed 3 hours and 3 minutes for all texts together (equaling only 431 words per hour) and also had the longest evaluation times for each of the texts (e.g. 70 minutes for Text 2, equaling 521 words per hour). Compared to that, participant 1 and participant 4, with respective total evaluation times of 1 hour 14 minutes (1,067 words per hour) and 1 hour 30 minutes (877 words per hour), range somewhere in between.

This seems to indicate a correlation between the time needed by each participant to evaluate all three texts and that participant's answer to the question of whether or not they encountered any problems (see RQ-1 above). Participant 3, who stated that he/she had no problems at all, was faster than participants 1 and 2. Correspondingly, participant 2 was slowest of all, stating that he/she had encountered serious problems. The only exception to this pattern is participant 4, who also answered that he/she did not encounter any problems, but was still slower than participant 3, leading to a correspondingly lower words per hour ratio. This might simply be due to the personal work ethic of the participant or might be attributable to the fact that participant 3 had already been studying in the Master's program for 12 semesters, whereas participant 4 had only been studying for less than half that time (5 semesters).

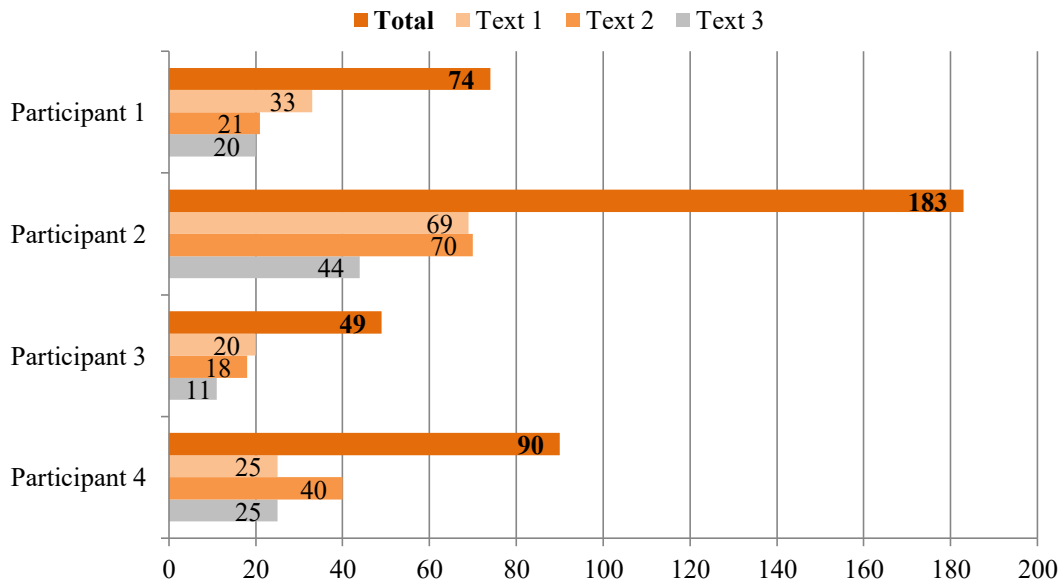


Figure 1: Participants' evaluation time of each text as well as total evaluation time (in minutes) in the replication study

Nonetheless, there could be other possible explanations for some participants being faster than others; one might suspect, for example, that those participants who evaluated faster did so because they worked with less diligence and overlooked issues in the texts that would have warranted further consideration and, possibly, a lower score. For this reason it is interesting to have a look at the evaluation results as submitted by participants, given in Table 3 below.

Table 3: Scores given by participants in each category for each text as well as total scores per text and per participant in the replication study

		TLA	FTA	MEAN	TERM	Total per text	Total per participant
Participant 1	Text 1	20	20	25	15	80	250
	Text 2	25	20	25	15	85	
	Text 3	25	20	25	15	85	
Participant 2	Text 1	20	15	20	15	70	205
	Text 2	20	15	20	10	65	
	Text 3	25	15	20	10	70	
Participant 3	Text 1	20	20	25	15	80	235
	Text 2	20	20	20	10	70	
	Text 3	25	20	25	15	85	
Participant 4	Text 1	15	15	20	15	65	230
	Text 2	20	20	25	15	80	
	Text 3	25	20	25	15	85	

Note: For the full meaning of the category abbreviations, see chapter 3.3. For information on which score corresponds to which descriptor in each category, see Colina (2008: 128ff.).

It becomes immediately apparent, when looking at Table 3, that participant 1 gave the highest scores in total, confirming the participant's self-assessment of having rated very highly throughout the evaluation. It is also readily apparent that participant 2 gave the lowest scores out of all participants, while the total scores by participants 3 and 4 are fairly similar and range somewhere in between. There thus seems to be a certain trend towards fast evaluators (participant 3 and 1 in particular) giving higher scores compared to evaluators who were considerably less fast (especially participant 2) and gave lower scores⁸⁶. However, a closer examination of the comments made by participants reveals that this is probably not due to a lack of diligence on the part of the faster evaluators.

According to the comments, participant 2 only noticed one issue that seems to have escaped the notice of the other participants (namely some hyperlinks in the texts pointing to websites that were in a language other than the TL, without this being indicated in the translation) as well as some minor points, mainly related to spelling, prepositions and frequency of use of expressions, which some other participants also mention noticing. No major issues are mentioned that might account for the considerably lower scores chosen by participant 2. It would rather seem, in fact, as though participants simply judged the same instance differently, with participant 2's decisions diverging from the judgement of the other participants. For example, participant 4 noticed that a specific term had not been explained in Text 2, yet chose descriptor 4c (the best possible descriptor in the fourth category (TERM)) nonetheless, whereas participant 2 chose descriptor 2c (the second-best descriptor in the second category (FTA)) because of it. This example may also serve to point out that participant 2 frequently mixed up categories (see also RQ-1 above), since terminological issues should belong in the fourth category. This might explain, to some degree, why participant 2's ratings diverge from the ratings of the other participants.

Furthermore, participant 2 seems to have been somewhat more critical than the other participants, who commented on noticing some of the same issues, but apparently judged them to be of less importance in the given context. Participant 2 even mentioned in one instance that a descriptor associated with a higher score could also have been chosen, but ultimately decided on the lower score. In addition, no pattern becomes apparent regarding scores for any one text being particularly low across participants, contributing to the assumption that faster participants were no less diligent than slower participants (especially participant 2). This, in turn, suggests that diligence did not have a great influence on

⁸⁶ Note, however, that the correlation is not perfect, since participant 3, who was by far the fastest evaluator, did not give the highest (total) score(s).

productivity, contrary to the problems encountered by some evaluators during the evaluation (as was assumed above in relation to Figure 1).

A comparison of the above-given evaluation time results from the replication study with the results from Colina's raters is somewhat complicated, since the rating time values found in Colina (2008: 118ff.) are average values given per language (Chinese, Russian, Spanish), but across rater groups (consultant, translators, language teachers and, for Spanish, bilinguals), whereas the replication study only had one rater group and one language pair. Furthermore, the texts used by Colina (2008: 111) differ somewhat in length from those used in the replication study, therefore a direct comparison of rating times would not yield very meaningful results. For this reason, the average evaluation time (in seconds) per ten words was calculated for each text in the replication study (see Table 4 below), in order for the results to be comparable to those of Colina's (2008) study⁸⁷.

Table 4: Average evaluation time per text (in minutes) and per ten words (in seconds) as well as average productivity expressed as words per hour, relative to source text length in the replication study

	Text 1	Text 2	Text 3
length of ST in words	428	608	280
average time per text (min.) ⁸⁸	36.75	37.25	25.00
average time per ten words (sec.)	52	37	54
average words per hour	865	1,299	855

For the original study, average rating times were recalculated from the 'Average' column in Colina's (2008: 118ff.) tables to exclude consultants, because the replication study does not feature consultants, and also because the consultants in the original study were only one person (n=1) per language, whereas the other rater groups were all composed of several raters each (n= >1 (min. 3, max. 6)). In Colina's 'Average' calculation, however, the consultant is treated equally to any of the other rater groups (e.g. for Russian, the average times of all translators were added to those of all teachers and the consultant and then divided by three). This is particularly misleading, because the consultants in Colina (2008: 118ff.) tended to produce rating times dissimilar to those of the raters (e.g. average rating time of Russian translators (n=5)/teachers (n=3) for Text 2 was 32.6/60.0 minutes, whereas the consultant

⁸⁷ Note that the word count is again the word count of the source text. Note also that Colina's (2008: 111) article did not contain an exact ST word count; it was only mentioned that one text was 1,200 words long and the other two between 270 and 370 words each, but not which word count corresponds to which of the three texts. This information was obtained through e-mail correspondence with the author of the original study (Colina 2022) (see Appendix IV). According to this, Text 1 had 1,200 words, Text 2 had approximately 270 words (rounded up to the nearest tenth), and Text 3 had 367 words.

⁸⁸ The values for average rating time (Table 4) were rounded commercially to highlight the difference between Text 1 and Text 2. For the calculation of average time per ten words, the decimal values were also used, but the results were rounded to whole numbers, since a differentiation into milliseconds seemed unnecessary.

(n=1) took 15.0 minutes only). Therefore, average rating times including the consultants tend to differ considerably from average rating times excluding the consultants (e.g. average time including consultant for Russian Text 2 is 35.86 minutes vs. average time excluding consultant is 46.30 minutes). Tables 5, 6 and 7 below show the results of the recalculation (average time of each rater group excluding consultant) as well as the calculation of the rating time per ten words for each text in each TL of the original study.

Table 5: Average evaluation time for the Chinese language per text (in minutes) and per ten words (in seconds), relative to source text length in the original study (adapted from Colina (2008: 118))

Chinese	Text 1	Text 2	Text 3
length of ST in words	1,200	270	367
average across all rater groups (excluding consultants) (min.)	57.4	47.6	46.1
average per ten words (sec.) – translators	32	146	88
average per ten words (sec.) – language teachers	26	65	63
average per ten words (sec.) across all rater groups (excluding consultants)	29	106	75

Table 6: Average evaluation time for the Russian language per text (in minutes) and per ten words (in seconds), relative to source text length in the original study (adapted from Colina (2008: 118))

Russian	Text 1	Text 2	Text 3
length of ST in words	1,200	270	367
average across all rater groups (excluding consultants) (min.)	56.0	46.3	43.7
average per ten words (sec.) – translators	24	72	46
average per ten words (sec.) – language teachers	32	133	97
average per ten words (sec.) across all rater groups (excluding consultants)	28	103	71

Table 7: Average evaluation time for the Spanish language per text (in minutes) and per ten words (in seconds), relative to source text length in the original study (adapted from Colina (2008: 120))

Spanish	Text 1	Text 2	Text 3
length of ST in words	1,200	270	367
average across all rater groups (excluding consultants) (min.)	44.2	39.7	37.2
average per ten words (sec.) – translators	21	96	63
average per ten words (sec.) – language teachers	26	92	82
average per ten words (sec.) – bilinguals	20	76	37
average per ten words (sec.) across all rater groups (excluding consultants)	22	88	61

In the following paragraphs, the results of the comparison between the rating time of Colina's rater groups (Chinese, Russian and Spanish translators as well as Chinese, Russian and Spanish language teachers⁸⁹) and the replication study participants will be set out and discussed in order to prove or disprove the two hypotheses made for RQ-2.

Students and novice translators vs. professional translators: A comparison between the data given in Tables 3 and 4 above shows that Chinese translators in the original study needed more time to rate two out of three texts than the participants in the replication study did, since Chinese translators needed 146 seconds per ten words (Text 2) and 88 seconds (Text 3), while novice translators only needed 37 seconds and 54 seconds, respectively. However, for Text 1, Chinese translators were faster than novice translators, who needed 52 seconds per ten words, whereas Chinese translators only needed 32 seconds.

Note that, similar to the replication study, rating time results in the original study do not seem to correlate closely with the word count of the texts, since Text 1, with 1,200 words, was the longest text by far, but Chinese translators, for example, took more time on average for Text 2 (which only had 270 words) than for Text 1, resulting in a great discrepancy between average times per ten words. It is also noteworthy that, for some unknown reason⁹⁰, all translators in the original study, regardless of the TL, seem to have taken a disproportionate amount of time rating Text 2, often twice or even three times as long as for Text 1.

Not dissimilar to the Chinese translator rater group, Russian translators also rated faster than novice translators for two out of three texts, but were slower for Text 2. While novice translators only needed 37 seconds per ten words for Text 2, Russian translators needed 72 seconds, although they were considerably faster for Text 1 (professional translators 24 seconds vs. novice translators 52 seconds) and somewhat faster for Text 3 (professional translators 46 seconds vs. novice translators 54 seconds).

A comparison with Colina's Spanish translators shows results similar to the Chinese translators: their rating took less time than was needed by novice translators for only one out of three texts, namely Text 1, for which professional translators needed only 21 seconds, whereas novice translators needed 52. For Text 2 and Text 3, on the other hand, novice

⁸⁹ Average rating times per ten words were calculated for the rater group 'Spanish bilinguals' in the original study for completeness' sake (see Table 7), but no comparison was made between Spanish bilinguals and students and novice translators, since a bilingual rater group was only available for Spanish in the original study (but not for either Russian or Chinese). Neither was any comparison attempted between students and novice translators and consultants, since consultants were not a proper rater group (see Colina 2008: 107ff.), although they are treated as one in the 'Average' calculations.

⁹⁰ No possible reason for this is given in Colina (2008) either, since Colina's analysis was largely concentrated on comparisons between rater groups and between languages, not between texts.

translators needed less time than professional Spanish translators did, completing their rating in 37 seconds per ten words for Text 2 (vs. 96 seconds needed by professional translators) and 54 seconds for Text 3 (vs. 63 seconds needed by professional translators).

Table 8: Average rating time per ten words (in seconds) of professional translators in the original study compared to that of students and novice translators in the replication study.

	Text 1		Text 2		Text 3		average across texts	
students and novice translators	52		37		54		48	
Chinese professional translators	32	↓	146	↑	88	↑	89	↑
Russian professional translators	24	↓	72	↑	46	↓	47	↓
Spanish professional translators	21	↓	96	↑	63	↑	60	↑

Note: Arrows next to time values indicate whether value is higher (↑) or lower (↓) compared to value of novice translators.

As Table 8 above clearly shows, in two out of three languages in the original study, professional translators were slower than novice translators more often than they were faster, only Russian translators were faster for more texts than they were slower. The same is reflected in the average values calculated across texts (Chinese and Spanish translators having rated slower than students and novice translators), while Russian translators alone were faster, if barely. This disproves H-2.1, since students and novice translators clearly did not rate much slower than Colina's more experienced professional translators. Whether this is due to the texts used in the respective studies (e.g. level of quality, amount of research needed, etc.) or due to some other factor remains unknown. It would seem, however, as though professional experience did not have a significant influence on rating time in this instance, again disproving H-2.1.

Students and novice translators vs. language teachers: When comparing the productivity of Chinese language teachers with the productivity of students and novice translators, it quickly becomes apparent that Chinese language teachers rated two out of three texts more slowly than novice translators did, while being much faster for Text 1 (26 seconds per ten words for teachers vs. 52 seconds for novice translators). For Text 2, Chinese language teachers needed considerably longer than novice translators (65 seconds per ten words vs. 37 seconds) and also somewhat longer for Text 3 (63 seconds per ten words vs. 54 seconds). Once again, the results from the original study do not show any close correlation between rating time and word count, since all language teachers needed longer for Text 2 than for Text 1, and in some cases also longer for Text 2 (which was the shortest text) than for Text 3.

Interestingly, the results of both the Russian and the Spanish language teachers match the pattern described above for the Chinese teachers: teachers always needed less time for Text 1 than novice translators did, but took considerably longer than novice translators in rating Text 2 and Text 3 (see Table 9 below). Specifically, Russian teachers needed only 32 seconds per ten words for Text 1 (compared to 52 seconds for novices translators), but more than thrice the time that novice translators needed for Text 2 (133 seconds per ten words vs. 37 seconds) and almost twice as long as novice translators for Text 3 (97 seconds per ten words vs. 54 seconds). Conversely, Spanish language teachers needed only 26 seconds per ten words for Text 1 (vs. 52 seconds that novice translators needed), but 92 seconds per ten words for Text 2 and 82 seconds for Text 3, whereas novice translators needed only about half that time (37 seconds per ten words for Text 2, and 54 seconds for Text 3).

Table 9: Average rating time per ten words (in seconds) of language teachers in the original study compared to that of students and novice translators in the replication study.

	Text 1		Text 2		Text 3		average across texts	
students and novice translators	52		37		54		48	
Chinese language teachers	26	↓	65	↑	63	↑	51	↑
Russian language teachers	32	↓	133	↑	97	↑	87	↑
Spanish language teachers	26	↓	92	↑	82	↑	67	↑

Note: Arrows next to time values indicate whether value is higher (↑) or lower (↓) compared to value of novice translators.

Table 9 above highlights that, across all languages, novice translators rated two out of three texts faster than language teachers, thus on a whole proving H-2.2. This is reflected in the average calculated across texts, since all language teachers, regardless of TL, rated slower on average than novice translators and students did. Whether this is indeed due to language teachers' tendency to reflect more thoroughly on their evaluation (being often required to give feedback), as Colina (2009: 255) hypothesized in her later study (see Introduction), remains unclear. In future studies, a questionnaire-item relating to task distribution within the time needed for the evaluation (e.g. time spent on reading the texts and deciding between descriptors, time spent on research or on scoring or comments, etc.) might give insight into this matter.

Summarizing the results described and discussed above, an answer to RQ-2 ("How does the productivity (measured as rating time in minutes) of MA students and novice translators compare to Colina's groups of evaluators?") can now be given. Generally, great discrepancies exist between the productivity (rating times) of individual evaluators, as some

needed considerably longer to rate certain texts than others. Some raters were also much slower overall, which is quite possibly due to problems they encountered during their evaluation, since nothing could be found in the comments to suggest that faster raters had worked with any less diligence. While the results of the original study make it difficult to ascertain whether the same can be said for Colina's (2008) study (the results being given only as averages across each rater group, but not for individual raters), a closer look at consultants' rating times (Colina 2008: 118ff.), which in some instances differ considerably from average times of rater groups, allows the assumption that these discrepancies may also not be unique to the replication study.

Contrary to what was put forward in H-2.1, students and novice translators did not rate slower than professional translators in the original study. In fact, novice translators rated faster on average than Chinese and Spanish translators, while Russian translators were only a little faster than novice translators, suggesting that professional experience did not have significant influence on productivity (rating time). H-2.2, on the other hand, was proven correct: students and novice translators rated faster on average than language teachers in the original study did, with the average rating time of Chinese language teachers approaching most closely the time needed by students and novice translators. It would therefore seem as though students and novice translators, in this instance, exhibited similar or even higher productivity (on average) than their professional counterparts in the original study did.

Although not directly related to RQ-2 or to any of the corresponding hypotheses, the start and finish times supplied by raters (Table 10) as part of their time recording coincidentally also yield interesting information regarding another aspect of rating behavior: whether participants evaluated all three texts in a row or preferred to take breaks in between⁹¹.

Table 10: Start and finish time of each participant for each text (given in 24-hours format)

	Text 1		Text 2		Text 3	
	Start time	Finish time	Start time	Finish time	Start time	Finish time
Participant 1	14:12	14:45	14:56	15:17	15:51	16:11
Participant 2	11:30	12:39	16:57	18:07	18:09	18:53
Participant 3	13:57	14:17	14:18	14:36	14:50	15:01
Participant 4	12:10	12:35	12:35	13:15	13:20	13:45

⁹¹ As stated in chapter 4.5, participants were allowed to conduct the evaluation in their own time within a given deadline (between the end of the training session and 10 a.m. the next day).

As becomes apparent from the results, summarized in Table 11 below, all evaluators chose to take at least one break. Participants 1, 2 and 3 took two breaks each, between Texts 1 and 2 and between Texts 2 and 3, respectively, while participant 4 took one break only, between the evaluation of Texts 2 and Text 3. The total length of breaks per participant ranged from over 4 hours to only 5 minutes, with participant 4, of course, having the shortest break time of all.

It may be noted from Table 11 below that participants who reported having had no problems (participants 3 and 4) also took the shortest breaks and that the participant who reported the most problems (participant 2) took the longest break by far, though this might be coincidental. Further studies would be necessary to ascertain whether ease of usage of the model correlates with the length of breaks needed by any one evaluator over the course of e.g. a work day.

For participants who took two breaks, one break was always longer than the other (e.g. participant 2 took two breaks of 260 minutes in total, of which one was 258 minutes and the other only 2 minutes long). However, no pattern becomes apparent in the placement of the breaks (between Texts 1 and 2, or between Texts 2 and 3), since some participants chose to take their longer break after the evaluation of Text 2 (the longest text in terms of word count), but one participant took his/her longer break after the evaluation of Text 1. Once again, further studies would be needed to determine whether text length, break length, and the resulting placement of breaks between texts relate to each other so that a pattern might be discerned.

Table 11: Number, length and placement of breaks between texts for each participant (in minutes)

	Number of breaks	Total length of breaks	Length of longer break	Placement of longer break
Participant 1	2	45	34	between Text 2 and 3
Participant 2	2	260	258	between Text 1 and 2
Participant 3	2	15	14	between Text 2 and 3
Participant 4	1	5	N/A	N/A

5.4 RQ-3: What is the attitude of MA students and novice translators towards the use of Colina's model in professional practice?

H-3.1: Replication study participants may consider the use of Colina's TQA model in practice, but will remain somewhat skeptical, due to their lack of experience with which to judge both the positive and negative potential aspects of practical use.

H-3.2: Students and novice translators will recognize the need for TQA in professional practice as well as the need to use a structured model to do so.

H-3.3: Replication study participants will see the functionalist and descriptive nature of Colina's TQA model in a generally positive way, since at least some translators in the original study did the same.

As RQ-3 asks for results obtained solely from the questionnaire as well as the comments made by participants, it is hardly possible to summarize most of the results in a table or figure before discussing them. Therefore, results for RQ-3 have been divided into all the answers and/or comments relevant to each of the three hypotheses, which will now be discussed one after the other. Examining the hypotheses in order from general to specific, results pertaining to H-3.2 will be given and discussed first.

All participants of the replication study stated, in answer to the questionnaire, that they think TQA to be generally necessary in professional translation to assess the quality of one's own translations or of translations done by fellow translators. Participant 3 also added that he/she thought that TQA can give one "a good idea of where a particular translation needs to be improved", thus partially bearing out H-3.2, since students and novice translators clearly recognized the need for TQA in professional translation practice. However, the other half of H-3.2 was disproven by the results, since only one out of four participants definitely recognized the need for a structured evaluation tool such as Colina's (2008) to carry out the quality assessment. The remaining three participants each stated that they were unsure whether a structured TQA tool (such as Colina's) or subjective judgements (not based on any defined evaluation tool) were best suited for the evaluation of translated texts.

As implied in other hypotheses, this insecurity may be due in part to the participants' lack of professional experience, making it difficult for them to judge questions on practical use. It is interesting to note, however, that none of the participants ticked the purely negative answer (i.e. that a structured tool is not necessary at all, because TQA can be done just as well by other, subjective, means), thus indicating that they recognized the need for structured TQA by means of a defined tool for at least some cases of professional translation practice. One

participant even elaborated in a comment that he/she thought a structured evaluation model would be appropriate in some professional contexts, but not in all. As an example the participant mentioned the TSP he/she was working at (this being participant 2, who had already graduated), and how a structured model like Colina's would be a good idea to use when evaluating the test translations of potential external contractors, since this would allow for scores to be easily compared. For TQA happening on a daily basis, however, participant 2 considered that using Colina's (2008) model would take too long (compared to simply proofreading and then approving translations). In future studies, participants might be educated further on the adaptable nature of Colina's (2008) model and then asked whether they would consider an adapted, simplified and potentially shortened version of the model, perhaps without scores, to be useful in day-to-day evaluation.

Contrary to H-3.2, which was partially disproven by the answers given by participants, H-3.1, relating to Colina's (2008) model in particular and to its potential use in practice, was mostly proven correct. Two participants stated that they would indeed use Colina's (2008) evaluation tool in practice, because of how easy it is to use (participant 4) and because "the different categories make sense" (participant 3). Another participant answered that he/she would consider using Colina's (2008) tool in practice, but was unsure because he/she had no clear idea of how representative Colina's (2008) tool is, compared to the methods generally used for the evaluation of translations⁹². Only one out of four participants refused the use of the tool in practice (this being the same participant who complained of extensive problems and a general dislike of the tool), due to the reasons discussed above under RQ-1. Nevertheless, this participant also underlined that he/she found the descriptors provided with the tool to be "quite useful" (participant 2). Taking into account that two out of four participants were favorably inclined towards the use of the model in practice and one at least considered the use of Colina's (2008) tool, H-3.1 was proven correct on a whole, even if the hypothesis did not include the possibility of some participant(s) simply disliking the model and therefore not considering its practical use to be an option.

As for the participants being skeptical of the model's use in practice due to their lack of professional experience, this part of the hypotheses is given credence by the fact that the one participant who answered that he/she would consider the use of Colina's model, but

⁹² The training session did not include any theoretical information on TQA beyond the basic characteristics needed to understand Colina's (2008) tool, since a lengthy introduction to the theory of translation evaluation would have imposed unduly on the participants who were, after all, not paid for their participation. Therefore, participants had no way to judge the 'normality' of Colina's (2008) TQA model compared to other common means of evaluating translations, such as the ones discussed in chapter 2.2.

would not definitely use it, was participant 1, who had the least experience out of all participants (having studied in the Master's program for only two semesters prior to his/her participation in the replication study). The other two participants who stated that they would use the model, had been studying for five and twelve semesters, respectively, permitting the possible assumption that (slightly) greater, non-professional experience allowed these participants to feel more certain of their attitude towards using the model in practice.

Regarding H-3.3 and the functionalist and descriptive nature of Colina's (2008) tool, participants' opinions were less divided. One participant had no strong opinion either way regarding the functionalist characteristics of the tool, while three evaluators stated that they saw them in a positive way, because they were aware of the importance of a translation fulfilling its function (participant 3 and 4). One participant further added that he/she was positively inclined towards the functionalist nature of the model, because functionalist traits "reflect[...] the nature of a translation as a text that is not closed in itself but part of the 'wider world': it is set in a context, has a target audience etc. and actually works as any other (source) text would" (participant 2⁹³).

Conversely, one participant did not have a strong opinion towards the descriptive nature of the tool, whereas the remaining three participants saw the descriptors in a positive way (see Figure 2 below), agreeing that they made the evaluation easier from the evaluator's point of view and served to guide the evaluators. Two participants further commented that they believe the descriptors to make the evaluation more objective and/or less subjective, which is particularly interesting, considering that it was one of Colina's goals to produce a testable, reproducible way of evaluating translations when she created her descriptive model.

One participant also stated that the descriptors give more meaning to the scores, adding that he/she would therefore consider attributing scores somewhere in-between two descriptors in cases where he/she could not readily decide between the two, to reflect a greater tendency towards one or the other. This point is interesting, because this approach of attributing in-between scores to reflect indecision between descriptors is one of the potential means of adapting the model mentioned by Colina (2008: 123). Another participant further added that the descriptors encourage evaluators to explain themselves in the comments-column, since the descriptors give them a base line to elaborate on when justifying their decisions. If this is true, it would also lead to greater transparency than is common in some

⁹³ That this participant could definitely identify with a positive opinion towards functionalist characteristics, even though he/she specified that he/she did not like how these were implemented in the tool, shows the importance of functionalism in TQA that so many scholars recognize (see chapter 1 and chapter 2.2.6).

analytical models, and enable both commissioners and translators to gain a greater understanding of why a particular score was given. Considering that Colina included this column for evaluators' comments in her model and that she later included a cover sheet with summarizing statements to make it easier for commissioners to understand scores, what this participant recognized is probably something that Colina specifically intended. This was also a point that was indirectly mentioned by the raters in Colina's (2008: 124) original study, who said that the descriptors enabled commissioners to easily see why a particular score was given.

Since three out of four participants saw the respective functionalist and descriptive nature of Colina's (2008) tool in a positive way and the last participant did not see either characteristic in a negative way, the results affirm H-3.3, coinciding with the positive attitude of Colina's (2008: 124f.) raters towards the functionalist and descriptive nature of the tool.

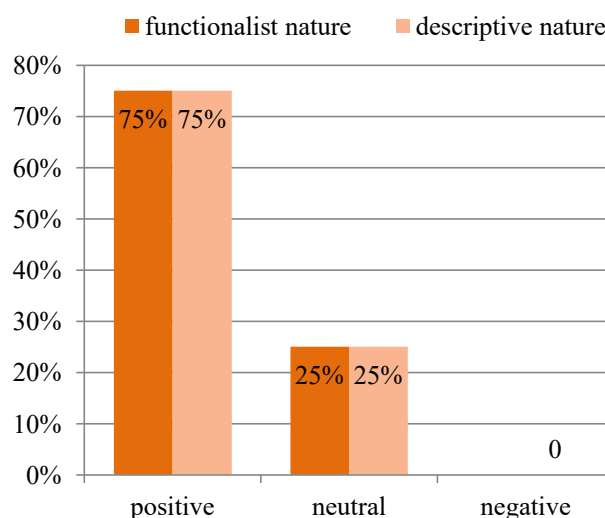


Figure 2: Replication study participants' attitude towards the functionalist and descriptive nature of Colina's (2008) TQA model

The qualitative comments made by participants also revealed some further, interesting points on participants' attitude towards or opinion of Colina's (2008) model, not related to any particular hypothesis. These comments will now be given and discussed in the following paragraphs.

The first interesting expression of attitude in the comments relates to what was stated in chapter 4.4 regarding subjectivity of participants and the different quality expectations that evaluators have. One participant twice made a comment that made it clear that, although two different terms were being used in the ST to describe one and the same concept, he/she would have preferred for the use of terms to be consistent in the TT, because "who says the

translation cannot be better than the ST?” (participant 2). This shows that the participant clearly had the expectation for the TT to be the best possible translation regardless of whether this was true for the ST, an attitude that apparently caused him/her to choose descriptor 4.b (the second-best descriptor out of three descriptors in this category) instead of descriptor 4.c (the best possible descriptor in this category). Since no other participant made any mention expecting the TT to be superior to the ST and no other participant chose any descriptor other than 4.c, it may be assumed that it was indeed this personal quality expectation that had a certain influence on the rating behavior of the participant in question (and also on his/her evaluation time, see RQ-2 above). While not directly related to the research question or any of the hypotheses, this nonetheless is an excellent example of the inherent subjectivity of translation evaluation. It may further serve as an example of what was mentioned above regarding the usefulness of the descriptors and the column for comments that encourage evaluators to justify their rating decisions: if participant 2 had not made his/her comment regarding his/her quality expectation, his/her score for category 4 would have stood out among the other (higher) scores, without any explanation as to whether this participant found a major issue all others overlooked, whether the participant made a mistake him/herself or whether, as was true in this case, the score is due to a personal preference or opinion and might be taken with a grain of salt by the commissioner.

In another comment, participant 1 stated that he/she felt that “the categories are not necessarily complete and that if someone else had created this translation [evaluation] tool, there could have been more or less and completely different categories”. Sadly, the participant failed to mention what these different categories might have been or which parts exactly he/she found to be incomplete. It is important to mention here that the training session did not include a full explanation of the tool and its creation in order not to impose unduly on the participants’ time, and the fact that it was adaptable to different evaluation needs was only mentioned briefly. The participant may therefore have (incorrectly) assumed that the model was fixed and could not be altered by adding or subtracting categories as necessary for an evaluation context or a specific text (see chapter 3).

Interestingly, participant 2 seemed to feel similarly, because he/she suggested that a category ‘Layout’ would be useful to have in the tool⁹⁴. Another participant also noticed this, although indirectly, because he/she made a layout-related comment next to one of the descriptors of category III (Non-specialized content (Meaning)), apparently not knowing where else to put it. Even though the adaptation mechanisms of the tool were not explained to them, these comments show that participants were able to anticipate the need for a versatile, adaptable model that specifically fits the need of the evaluation context and the text under evaluation. Presumably, had the evaluated text been e.g. an audio description guide for visually impaired persons, layout would have seemed less important and another category, e.g. technical integration (with audio description systems) might have been suggested⁹⁵.

Participant 1, in one of his/her comments, also anticipated a change to the model that was in fact implemented after Colina’s first study in 2008: the participant felt that there should also be “descriptive statements for the overall scores so that the end scores would make more sense and could be interpreted easily by looking solely at them.” The first version of the model (the one used in the replication study) was changed for the second study in 2009 to include a cover sheet where raters could tick summarizing statements⁹⁶ and/or make suggestions based on their final score, in order for the evaluation results to be more readily understandable for the commissioners of the evaluation (Colina 2009: 258).

In summary of what has been laid out above, RQ-3 can now be answered based on the results of the replication study. All participants generally recognized the need for quality assessment of translations in a professional context, even though they were mostly unsure about the necessity for a structured evaluation model, thus partially proving and partially disproving H-3.2. Participants’ insecure attitude in this respect is possibly due to their lack of

⁹⁴ Some points of the layout and graphical design of a text are taken into consideration under the category Functional and Textual Adequacy in order to evaluate the TT’s compliance with the required text type (Colina 2015: 253) and its intended function (e.g. cultural appropriateness of colors). However, the participant correctly remarked that the model (in the form presented to the participants) made it impossible to score layout-related matters that do not directly impact the functional or textual adequacy of the TT (such as minor layout issues, inconsistencies etc.).

⁹⁵ At the very least, technical issues might be considered an important part of the category Functional and Textual Adequacy for some types of text, e.g. for games, texts intended for accessibility (e.g. for visually impaired people through screen readers) or sophisticated websites. Even for the texts used in the replication study, which were typical texts found on a website, some participants commented on technical points by adding comments in the FTA category regarding the function or lack of function of hyperlinks, and whether the linked websites offered texts only in the SL (German) or also in the TL (English) and whether an appropriate note had been made to inform readers about the latter.

⁹⁶ Summarizing statements among which evaluators can choose include: “Publish and/or use as is”, “Minor edits needed before publishing*”, “Major edits needed before publishing*”, “Redo translation” and “Translation will not be an effective communication strategy for this text. Explore other options (e.g. create new target language materials)” (Colina 2009: 258). The asterisk behind the second and third statement indicates that the evaluator should describe the necessary edits below under “Notes/Recommended Edits” (Colina 2009: 258).

experience in professional translation, but perhaps also due in part to a reluctance to generalize, thinking that either a structured model or a less structured approach may be appropriate in some contexts. Participants were also mostly in favor of using Colina's (2008) evaluation tool in practice or at least considered its practical use, thus (with the exception of participant 2) proving H-3.1. The fact that the participant with the least experience (only two semesters) was the only one to tick the 'unsure' option regarding whether he/she would use Colina's (2008) tool in practice also gives some support to the hypothesis that lack of experience makes it more difficult to judge matters of practical application, since the participants with more experience seemed more certain of their attitude, whether positive or negative.

The majority of participants further saw the functionalist and descriptive nature of the model in a positive way, with one participant claiming a neutral disposition towards either characteristic. Participants thus clearly recognized the importance of incorporating functionalism into TQA, and even showed a certain initiative, anticipating some points that were later incorporated into Colina's second study (e.g. the summarizing statements to describe scores). This largely positive attitude not only confirmed H-3.3, but is also in accordance with views voiced by professional translators in the original study. Furthermore, participants made some suggestions towards improving Colina's (2008) tool, with some participants stating that they felt it to be incomplete and that a Layout category, in particular, would have been useful. Participants thereby showed a tendency towards a versatile, adaptable model that precisely fits the needs of the text under evaluation. Furthermore, in combination with the anticipation of features that the author of the original study herself later deemed worth including in her model, this shows that students and novice translators, in spite of their lack of professional experience, may offer valuable insights and opinions and are capable of contributing relevant suggestions to the field of translation quality evaluation.

6 Conclusion

The subject of this MA project was the evaluation of translation quality and the evaluation of translated texts, using the TQA model created by Colina (2008). Specifically, a replication study was conducted on the use of the model by MA students from the University of Vienna and recently graduated novice translators in comparison with professionally experienced raters (translators and language teachers) that participated in the original study by Colina (2008).

The aim was to answer three research questions regarding the ability of students and novice translators to use Colina's (2008) model and problems they may experience (RQ-1), the productivity of students and novice translators as compared to the productivity (measured as rating time) of participants from the original study (RQ-2), and also regarding students' and novice translators' attitude towards the use of Colina's (2008) model in practice (RQ-3).

To this end, relevant terms such as *quality*, *translation*, *evaluation/assessment* as well as *translation quality evaluation/assessment* were defined first in chapter 1 and differentiated from various related terms. The concept of translation quality evaluation was then explored in detail in chapter 2, beginning with the cornerstones, the different types and the most common methods of TQA. Following that, different TQA models and approaches were explained and pertinent examples described in order to highlight the main characteristics and drawbacks of each approach. These were later summarized to contrast Colina's (2008) model, which was described in detail in chapter 3. First, the model's rationale of creation was briefly retraced, before the cornerstones of the model were given and its four main characteristics (functionalist, componential, descriptive, user-defined/adaptable) explained. After that followed an overview of the model's structure, the four evaluation criteria (Target Language, Functional and Textual Adequacy, Non-Specialized Content (Meaning), Specialized Content and Terminology). Following that, the original study was described in detail, with a particular focus on participants, procedure, and results and limitations.

In contrast to the original study, the cornerstones and various other important aspects of the replication study were then detailed in chapter 4. It was explained how texts were selected from the internet and according to which criteria; then each ST was briefly described. The translation brief and the weighting of scores used in the replication study were given as well as the process and manner in which they were elaborated, before the participants and their characteristics (language proficiency, degree(s), enrollment in MA program, experience) were described, based on a pre-study questionnaire. In relation with recruitment and

participants, the matter of subjectivity was briefly touched upon. Afterwards, the replication study's procedure was explained: the 40-minute training session with a short try-out example; the rating session in which participants were asked to evaluate three translated texts using Colina's (2008) model; and the post-study questionnaire comprising three pages that participants were asked to fill in. Following that, limitations applying to the replication study (such as limited scope and imperfect replication of the original study) were set out in chapter 5, before the results of the replication study were given. These results were then discussed and compared to Colina's results (where pertinent) to answer the research questions.

The results of the replication study showed that students and novice translators with little or no professional experience were able to use the TQA model by Colina (2008) to evaluate the quality of three texts translated from German into English. All participants returned complete evaluation and score sheets and the great majority felt confident in their scores. Nevertheless, as had been assumed when constructing RQ-1, half of the students and novice translators encountered problems during the evaluation that their more experienced counterparts in the original study did not have. Some problems seem to have been unrelated to professional experience, while some problems, likely rooted in insecurity, would probably not have occurred with more professional experience. Furthermore, the assumption that problems would mainly be related to the matching of descriptors to the texts under evaluation was proven incorrect, since this challenge, while being the most widespread, was far from the only one, and neither was it the most serious. That being said, none of the problems that participants encountered prevented them from successfully completing the evaluation.

Results pertaining to RQ-2 revealed that evaluation times differed considerably between participants, these differences probably being due to whether participants encountered any major problems (causing their evaluation to take more time) or not (meaning they rated faster than other evaluators). The scores given by participants revealed that faster evaluators tended to rate higher than the participant who took considerably more time for his/her evaluation, but no indication could be found in the comments to suggest that this was due to a lack of diligence on the part of the faster participants, suggesting that it was indeed the problems encountered by the slowest participant as well as his/her personal quality expectations that lead to the divergence in scores. Moreover, no pattern of particularly high or low scores could be discerned for any one text.

Results further showed that, within the limitations of the replication study, students' and novice translators' average productivity matched or even exceeded the average productivity of their professional counterparts (professional translators and language teachers) in the original study, productivity being measured as rating time. A comparison of time needed for evaluating ten words of each text showed that replication study participants rated faster than Chinese and Spanish professional translators and only slightly less fast than Russian professional translators. Moreover, in spite of significant discrepancies between individual rater performances, replication study participants were consistently faster than language teachers, as had been hypothesized regarding RQ-2.

The results revealed a largely positive answer to RQ-3, regarding the attitude of students and novice translators towards the use of Colina's (2008) model in practice. As had been assumed, the majority of participants were inclined to consider the professional use of the TQA model, while some remained unsure, the latter probably due to lack of professional experience. Replication study participants also showed a generally positive disposition towards both the functionalist and the descriptive nature of Colina's (2008) model, with three out of four participants claiming a positive attitude and the rest being of neutral disposition. Students and novice translators also summarily recognized the general need for TQA in a professional translation context, even though the majority of participants were unsure about the need for a structured TQA approach.

Further results, unrelated to any specific research question, were also obtained through the replication study. Notably, results suggest that Colina's (2008) model can be effective for the language combination German-English, the evaluation having been successfully completed by all participants without any problems related to either the SL or the TL. Some observations regarding breaks taken by evaluators between the evaluation of texts were also made, but no patterns could be discerned and further studies will be necessary to explore this aspect. It has also become clear from the suggestions that participants made, some of these even anticipating changes made to Colina's (2008) model at a later time by the author herself, that students and novice translators, in spite of their lack of professional experience, may offer interesting evidence, opinions and suggestions regarding translation quality evaluation.

Of course, this replication study is only a very small contribution to the knowledge pool and empirical evidence regarding Colina's (2008) model as well as the field of TQA, and leaves many options for further research. Among these are certainly the options Colina (2008: 125; 2009: 255) suggests in her works: to conduct similar studies on a larger scale (more participants and thus larger raters groups, more texts as well as multiple text types) and

to conduct empirical studies to find out which weighting of criteria would be ideal for which types of text.

Apart from that, consistency over time might also be an interesting aspect to investigate, since most, if not all, studies conducted to this day appear to only examine evaluations from one single point in time. In a study concerned with consistency of evaluation over time, a pool of texts with similar properties might be created to be given to evaluators at certain intervals of several months' time, to see if any tendencies arise regarding the use of the model (e.g. tendency to use the ST more or less, to make more or fewer comments, to rate faster or slower etc.), which might help to optimize its use.

Future studies might also explore specific rater characteristics and their influence on the usability and use of the model. One might, for example, explore whether domain competence would make much of a difference for any kind of text, or whether it would only make a difference for texts that rely heavily on specialized content and terminology. One might also compare the performance of raters with experience in using the model to the performance of raters without experience⁹⁷, in order to be able to make recommendations on which raters would be ideally suited for using the model on certain types of text.

Similarly, one might choose participants who have experience with translation as well as TQA and then let them explore the adaptable nature of Colina's (2008; 2009) model to see whether they can agree on aspects for particular evaluation needs (e.g. day-to-day evaluation needing to be streamlined for application). In this way, a portfolio of adapted versions of the model might be created, with each version being recommended for a particular type of text or evaluation context in order to have specific model versions readily available and save time.

Finally, future studies might also be concerned with efficiency and/or productivity in the sense of break-work ratio. One might examine the length of texts relative to the length of breaks evaluators need, in order to perhaps enable the use of a more efficient order of texts in relation to an evaluator's work day (e.g. longer texts before lunch or alternating longer with shorter texts). Similar studies might also ask questions regarding how much time raters spend on the different activities during an evaluation (reading texts, deciding between descriptors, research, scoring, commenting etc., which could be ascertained through screen recording), since this might give interesting insights into reasons for rating-time discrepancy between rater groups.

⁹⁷ This was not done in Colina (2009: 256), although many of the participants in the study had also been participants in the first study in 2008. Specifically, there were four raters in the second study who had not participated in the first study (Colina 2008) and thus had no prior experience in using the model, yet their performance was not examined separately from that of the raters who were using the model for the second time.

Bibliography

Primary sources

TQA model: Colina, Sonia (2008). Translation Quality Evaluation. Empirical Evidence for a Functionalist Approach. *The Translator* 14 (1), 97-134.

<https://doi.org/10.1080/13556509.2008.10799251> (last accessed: 2021-11-23).

Text 1 German original: ÖIF – Österreichischer Integrationsfonds (2022a). Corona-Sicherheitsmaßnahmen in 17 Sprachen. <https://www.integrationsfonds.at/coronainfo/> (last accessed: 2022-03-03).

Text 1 English translation: ÖIF – Österreichischer Integrationsfonds (2022b). Coronavirus safety measures in 17 languages. <https://www.integrationsfonds.at/coronainfo/en/> (last accessed: 2022-03-03).

Text 2 German original: Wien Tourismus (2022a). Testmöglichkeiten und To-dos bei einer Infektion. <https://www.wien.info/de/reiseinfos/covid-19-informationen/testmoeglichkeiten-wien-gaeste-387066> (last accessed: 2022-03-04).

Text 2 English translation: Wien Tourismus (2022b). Testing options and to-dos in the event of an infection. <https://www.wien.info/en/travel-info/corona-information/testing-options-vienna-guests-388220> (last accessed: 2022-03-04).

Text 3 German original: BMSGPK – Bundesministerium für Soziales, Gesundheit, Pflege und Konsumentenschutz (2021a). Coronavirus: So trage und verwende ich meine FFP2-Maske richtig [PDF document]. https://www.sozialministerium.at/dam/jcr:b0c66427-b3ba-4922-9a9a-0f9fda36b754/210216_Coronavirus_FFP2-Maske_pdfUA.pdf (last accessed: 2022-03-07).

Text 3 English translation: BMSGPK – Bundesministerium für Soziales, Gesundheit, Pflege und Konsumentenschutz (2021b). Coronavirus: How to wear and use your FFP2 mask [PDF document]. https://www.sozialministerium.at/dam/jcr:c22a8859-78aa-47cc-85a4-8d7a5ff5834d/210223_Coronavirus_FFP2-Maske_EN_PDFua.pdf (last accessed: 2022-03-07).

Sample text used during the training session German original: ÖIF – Österreichischer Integrationsfonds (s. a.a). So schützen Sie sich und andere. https://www.integrationsfonds.at/fileadmin/user_upload/Infoblatt_Corona-Sicherheitsmassnahmen.pdf (last accessed: 2022-05-10).

Sample text used during the training session English translation: ÖIF – Österreichischer Integrationsfonds (s. a.b). CORONAVIRUS safety measures.
https://www.integrationsfonds.at/fileadmin/user_upload/Infoblatt_Corona-Sicherheitsmassnahmen_EN.pdf (last accessed: 2022-05-10).

Secondary sources

- Abuja, Gunther & Annau, Eva & Ganster, Siegfried & Keiper, Anita & Tauschitz, Isolde & Mittendorfer, Franz & Nezbeda, Margarete & Steinhuber, Belinda & Winkler, Gabriele (2014). Das Europäische Sprachenportfolio für junge Erwachsene 15 + Sekundarstufe II. http://www.oesz.at/download/publikationen/esp15_gesamt_webversion.pdf (last accessed: 2021-07-06).
- al-Qinai, Jamal (2000). Translation Quality Assessment. Strategies, Parametres and Procedures. *Meta* 45 (3), 497–519. <https://doi.org/10.7202/001878ar> (last accessed: 2021-11-23).
- Ammann, Margret (1990). Anmerkungen zu einer Theorie der Übersetzungskritik und ihrer praktischen Anwendung. *TEXTconTEXT* 5, 209–250.
- Angelelli, Claudia (2009). Using a rubric to assess translation ability | Defining the construct. In: Angelelli & Jacobson (eds.), 13–47.
- Angelelli, Claudia & Jacobson, Holly (eds.) (2009). *Testing and assessment in translation and interpreting studies | A call for dialogue between research and practice*. Amsterdam: John Benjamins.
- Barghout, Mohamed Abdel-Maguid (1990). *Translation quality assessment: an application of a rhetorical model*. Doctoral thesis, University of Salford.
- BDÜ – Bundesverband der Dolmetscher und Übersetzer (2021). Vergünstigungen: Versicherungen. <https://bdue.de/der-bdue/leistungen/verguenstigungen/> (last accessed: 2021-11-19).
- Behr, Dorothée (2017). Assessing the use of back translation: the shortcomings of back translation as a quality testing method. *International Journal of Social Research Methodology* 20 (6), 573–584. <https://doi.org/10.1080/13645579.2016.1252188> (last accessed: 2021-11-23).
- Bensoussan, Marsha & Rosenhouse, Judith (1990). Evaluating student translations by discourse analysis. *Babel* 36 (2), 65–84. <https://doi-org.uaccess.univie.ac.at/10.1075/babel.36.2.02ben> (last accessed: 2022-09-02).

- Bittner, Hansjörg (2020). *Evaluating the Evaluator – A Novel Perspective on Translation Quality Assessment*. New York/London: Routledge.
- BMDW – Bundesministerium für Digitalisierung und Wirtschaftsstandort (2021). Unregulated businesses. <https://www.usp.gv.at/en/gruendung/gewerbe-in-oesterreich/freie-gewerbe.html> (last accessed: 2021-11-05).
- BMSGPK – Bundesministerium für Soziales, Gesundheit, Pflege und Konsumentenschutz (2022). AW: Informationen Masterarbeit [E-mail correspondence with a representative of the BMSGPK from 2022-03-18]. See Appendix IV.
- Bolaños Cuellar, Sergio (2007). *Towards an Integrated Translation Approach. Proposal of a Dynamic Translation Model (DTM)*. Dissertation, University of Hamburg.
- Bond, Esther (2020). Translation Services and the Race for a Coronavirus Vaccine. <https://slator.com/translation-services-and-the-race-for-a-coronavirus-vaccine/> (last accessed: 2021-11-29).
- Bowker, Lynne (2001). Towards a Methodology for a Corpus-Based Approach to Translation Evaluation. *Meta* 46 (2), 345–364. <https://doi.org/10.7202/002135ar> (last accessed: 2021-11-23).
- Brislin, Richard Walter (1969). *Back-Translation for Cross-Cultural Research*. Dissertation, Pennsylvania State University.
- Brunette, Louise (2000). Towards a Terminology for Translation Quality Assessment – A Comparison of TQA Practices. *The Translator* 6 (2), 169–182. <https://doi.org/10.1080/13556509.2000.10799064> (last accessed: 2021-11-24).
- Budin, Gerhard (2007). Entwicklung internationaler Normen im Bereich der Translationsqualität bei ISO/TC 37. In: Schmitt & Jüngst (eds.), 54–65.
- Byrne, Jody (2007). Caveat Translator: Understanding the Legal Consequences of Errors in Professional Translation. *Journal of Specialised Translation* 07, 2–24. https://jostrans.org/issue07/art_byrne.php (last accessed: 2021-10-13).
- Carroll, John (1966). An Experiment in Evaluating the Quality of Translations. *Mechanical Translation and Computational Linguistics* 9 (3/4), 55–66.
- CEN – European Committee for Standardization (s. a.). EN 15038:2006 Translation Services – Service Requirements | CEN Case Study Traducciones Políglota. <ftp://ftp.cencenelec.eu/CEN/WhatWeDo/Fields/Services/CaseStudies/TranslationServices.pdf> (last accessed: 2021-07-06).
- Chesterman, Andrew (2016). *Memes of Translation: The Spread of Ideas in Translation Theory*. Revised Edition. Amsterdam/Philadelphia: John Benjamins.

- Colina, Sonia (2003). *Translation Teaching: From Research to the Classroom: A Handbook for Teachers*. Boston/New York: McGraw Hill.
- Colina, Sonia (2008). Translation Quality Evaluation. Empirical Evidence for a Functionalist Approach. *The Translator* 14 (1), 97–134. <https://doi.org/10.1080/13556509.2008.10799251> (last accessed: 2021-11-24).
- Colina, Sonia (2009). Further evidence for a functionalist approach to translation quality evaluation. *Target* 21 (2), 235–264. <https://doi.org/10.1075/target.21.2.02col> (last accessed: 2021-11-23).
- Colina, Sonia (2011). Evaluation/Assessment. In: Gambier, Yves & van Doorslaer, Luc (eds.) *Handbook of Translation Studies Volume 2*. Amsterdam/Philadelphia: John Benjamins, 43–48.
- Colina, Sonia (2012). Assessment of Translation. In: Chapelle, Carol (ed.) *The Encyclopedia of Applied Linguistics*. <https://doi.org/10.1002/9781405198431.wbeal0053>: Blackwell Publishing.⁹⁸
- Colina, Sonia (2015). *Fundamentals of Translation*. Cambridge: Cambridge University Press.
- Colina, Sonia (2020). Quality, translation. In: Baker, Mona & Saldanha, Gabriela (eds.) *Routledge Encyclopedia of Translation Studies*. 3rd edition. London/New York: Routledge, 458–463.
- Colina, Sonia (2022). Re: [EXT]Question regarding 2008 article for my MA-thesis [E-mail correspondence with author of original study from 2022-08-11]. See Appendix IV.
- Colina, Sonia & Marrone, Nicole & Ingram, Maia & Sánchez, Daisey (2017). Translation Quality Assessment in Health Research: A Functionalist Alternative to Back-Translation. *Evaluation & the Health Professions* 40 (3), 267–293. <https://doi.org/10.1177/0163278716648191> (last accessed: 2021-11-23).
- Conde, Tomás (2011). Translation evaluation on the surface of texts: a preliminary analysis. *Journal of Specialised Translation* 15, 69–86. https://jostrans.org/issue15/art_conde.pdf (last accessed: 2021-11-14).
- CTTIC – Canadian Translators, Terminologists and Interpreters Council (2005). CTTIC Standard Certification Translation Examination | Marker’s Guide. <http://www.cttic.org/examDocs/guide.markersE.pdf> (last accessed: 2021-07-16).
- Dalla-Zuanna, Jean-Marc (2018). Das Ende von “one size fits all” – Dynamische Qualitätsbewertung von Übersetzungen mit MQM. *MDÜ* 3, 10–15.

⁹⁸ Since this volume has been published exclusively online, it features neither page numbers nor can a place of publication be supplied. The DOI (Digital Object Identifier) is given instead.

- Dalla-Zuanna, Jean-Marc & Kurz, Christopher(eds.) (2020). *Translation Quality in the Age of Digital Transformation*. Berlin: BDÜ Weiterbildungs- und Fachverlagsgesellschaft.
- Díaz Cintas, Jorge (2013). Subtitling. Theory, practice and research. In: Millán, Carmen & Bartrina, Francesca (eds.) *The Routledge Handbook of Translation Studies*. Oxon/New York: Routledge, 273–287.
- Do Campo Bayón, María (2020). Insights and Recommendations on MT Training Practices. <https://blog.taus.net/insights-and-recommendations-on-mt-training-practices> (last accessed: 2022-02-15).
- Dörnyei, Zoltán. 2010. *Questionnaires in second language research: Construction, administration, and processing*. 2nd edition. New York, Oxon: Routledge, Taylor & Francis Group.
- Douglas, Susan & Nijssen, Edwin (2003). On the use of “borrowed” scales in cross-national research – A cautionary note. *International Marketing Review* 20 (6), 621–642.
- Drugan, Joanna (2013). *Quality in Professional Translation – Assessment and Improvement*. London/New York: Bloomsbury Academic.
- Eyckmans, June & Anckaert, Philippe & Segers, Winibert (2009). The perks of norm-referenced translation evaluation. In: Angelelli & Jacobson (eds.), 73–93.
- Fields, Paul & Hague, Daryl & Koby, Geoffrey & Lommel, Arle & Melby, Alan (2014). What Is Quality? A Management Discipline and the Translation Industry Get Acquainted. *Revista tradumàtica* 12, 404–412.
- Fourie, Jean & Feinauer, Ilse (2005). The quality of translated medical research questionnaires. *Southern African Linguistics and Applied Language Studies* 23 (4), 349–367. <https://doi.org/10.2989/16073610509486395> (last accessed: 2022-01-27).
- Garant, Mikel (2009). A case for holistic translation assessment. *AFinLA-e Soveltavan kielitieteen tutkimuksia* 1, 5–17. <https://journal.fi/afinla/issue/view/469> (last accessed: 2021-11-26).
- Garvin, David (1984). What Does “Product Quality” Really Mean? *Sloan Management Review* 26 (1), 25–43.
- Gerzymisch-Arbogast, Heidrun (2001). Equivalence Parameters and Evaluation. *Meta* 46 (2), 227–242. <https://doi.org/10.7202/002886ar> (last accessed: 2021-11-24).
- Gouadec, Daniel (1981). Paramètres de l’évaluation des traductions. *Meta* 26 (2), 99–116. <https://doi.org/10.7202/002949ar> (last accessed: 2021-11-23).

- Gouadec, Daniel (1989). Comprendre, évaluer, prévenir – Pratique, enseignement et recherche face à l’erreur et à la faute en traduction. *TTR Traduction, terminologie, rédaction* 2 (2), 35–54. <https://doi.org/10.7202/037045ar> (last accessed: 2021-11-25).
- Gouadec, Daniel (2010). Quality in translation. In: Gambier, Yves & van Doorslaer, Luc (eds.) *Handbook of Translation Studies Volume 1*. Amsterdam/Philadelphia: John Benjamins, 270–275.
- Gutt, Ernst-August (2000/2014). *Translation and Relevance – Cognition and Context*. Oxon/New York: Routledge.
- Hablamos Juntos (2009). Tool 6 – Assessing Translation – A Manual for Requesters. https://riuss.org/wp-content/uploads/2017/10/Hablamos-Juntos_ToolKit6_AssessingTranslation.pdf (last accessed: 2021-07-13).
- Hagemann, Susanne (2007). Zur Evaluierung kreativer Übersetzungsleistungen. *Lebende Sprachen* 52 (3), 102–109. <https://doi-org.uaccess.univie.ac.at/10.1515/LES.2007.102> (last accessed: 2021-11-12).
- Hague, Daryl & Melby, Alan & Zheng, Wang (2011). Surveying Translation Quality Assessment | A Specification Approach. *The Interpreter and Translator Trainer* 5 (2), 243–267. <https://doi.org/10.1080/13556509.2011.10798820> (last accessed: 2021-11-23).
- Harvey, Lee & Green, Diana (1993). Defining quality. *Assessment & Evaluation in Higher Education* 18 (1), 9–34. <https://doi.org/10.1080/0260293930180102> (last accessed: 2021-11-16).
- Holmes, James (1994). *Translated! : Papers on Literary Translation and Translation Studies*. 2nd edition. Amsterdam: Rodopi.
- Holz-Mänttari, Justa (1984). *Translatorisches Handeln – Theorie und Methode*. Helsinki: Suomalainen Tiedeakatemia.
- Horguelin, Paul & Pharand, Michelle (2009). *Pratique de la revision*. 4th edition. Montréal: linguattech éditeur inc.
- House, Juliane (1977). A Model for Assessing Translation Quality. *Meta* 22 (2), 103–109. <https://doi.org/10.7202/003140ar> (last accessed: 2021-11-17).
- House, Juliane (1997). *Translation Quality Assessment – A Model Revisited*. Tübingen: Gunter Narr.
- House, Juliane (2009). Quality. In: Baker, Mona & Saldanha, Gabriela (eds.) *Routledge Encyclopedia of Translation Studies*. 2nd edition. London/New York: Taylor & Francis, 222–225.

- House, Juliane (2014). *Translation: a multidisciplinary approach*. Basingstoke: Palgrave Macmillan.
- House, Juliane (2015). *Translation Quality Assessment | Past and Present*. London/New York: Routledge.
- House, Juliane (2016). *Translation as Communication across Languages and Cultures*. London/New York: Routledge.
- ISO – International Organization for Standardization (2015a). *DIN EN ISO 17100:2015-05 Deutsche Fassung – Übersetzungsdienstleistungen – Anforderungen an Übersetzungsdienstleistungen (ISO 17100:2015)*. Brüssel: CEN.
- ISO – International Organization for Standardization (2015b). *ISO 9000:2015(en) Quality management systems — Fundamentals and vocabulary*.
<https://www.iso.org/obp/ui/#iso:std:iso:9000:ed-4:v1:en> (last accessed: 2021-10-22).
- Karoubi, Behrouz (2016). Translation quality assessment demystified. *Babel* 62 (2), 253–277.
<https://doi.org/10.1075/babel.62.2.05kar> (last accessed: 2021-11-23).
- Kingscott, Geoffrey (2007). Translation quality assessment. In: Schmitt & Jüngst (eds.), 317–325.
- Kiraly, Don (2000). *A Social Constructivist Approach to Translator Education – Empowerment from Theory to Practice*. Manchester: St. Jerome Publishing.
- Kockaert, Hendrik & Segers, Winibert (2017). Evaluation of legal translations: PIE method (Preselected Items Evaluation). *Journal of Specialised Translation* 27, 148–163.
https://jostrans.org/issue27/art_kockaert.pdf (last accessed: 2021-10-15).
- Koehn, Philipp (2020). *Neural Machine Translation*. New York: Cambridge University Press.
- Kurz, Christopher (2018). Eine Frage der Definition – Qualitätsbewertung professioneller Übersetzungen. *MDÜ* 3, 16–21.
- Kurz, Christopher (2020). Translation quality, translation evaluation, error categories and score sheets. In: Dalla-Zuanna & Kurz (eds.), 38–68.
- Larose, Robert (1994). Qualité et efficacité en traduction : réponse à F. W. Sixel. *Meta* 9 (2), 362–373. <https://doi.org/10.7202/003405ar> (last accessed: 2021-11-19).
- Larose, Robert (1998). Méthodologie de l'évaluation des traductions. *Meta* 43 (2), 163–186.
<https://doi.org/10.7202/003410ar> (last accessed: 2021-11-28).
- Lauscher, Susanne (2000). Translation Quality Assessment – Where Can Theory and Practice Meet? *The Translator* 6 (2), 149–168. <https://doi.org/10.1080/13556509.2000.10799063> (last accessed: 2021-11-29).

- Lommel, Arle & Uszkoreit, Hans & Burchardt, Aljoscha (2014). Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Revista tradumàtica* 12, 455–463. <https://doi.org/10.5565/rev/tradumatica.77> (last accessed: 2022-01-28).
- Lommel, Arle & Burchardt, Aljoscha & Uszkoreit, Hans (eds) (2015). *Multidimensional Quality Metrics (MQM) Definition*. <http://www.qt21.eu/mqm-definition/definition-2015-12-30.html> (last accessed: 2022-01-28).
- Maier, Carol (2000). Introduction. *The Translator* 6 (2), 137–148. <https://doi.org/10.1080/13556509.2000.10799062> (last accessed: 2021-11-21).
- Maneesriwongul, Wantana & Dixon, Jane (2004). Instrument translation process: a methods review. *Journal of Advanced Nursing* 48 (2), 175–186. <https://doi-org.uaccess.univie.ac.at/10.1111/j.1365-2648.2004.03185.x> (last accessed: 2022-01-27).
- Martínez Melis, Nicole & Hurtado Albir, Amparo (2001). Assessment In Translation Studies: Research Needs. *Meta* 46 (2), 272–287. <https://doi.org/10.7202/003624ar> (last accessed: 2021-11-25).
- Muzii, Luigi (2016). The Importance & Difficulty of Measuring Translation Quality. <https://kv-emptypages.blogspot.com/2016/10/the-importance-difficulty-of-measuring.html?m=1> (last accessed: 2022-05-06).
- Nida, Eugene (1964). *Toward a Science of Translating – With special reference to principles and procedures involved in bible translating*. Leiden: Brill.
- Nida, Eugene & Taber, Charles (1982). *The Theory and Practice of Translation*. Leiden: Brill.
- Nishino, Ryutaro (2020). Translation quality and JTF’s Evaluation Guidelines. In: Dalla-Zuanna & Kurz (eds.), 154–164.
- Nord, Christiane (2005). *Text Analysis in Translation – Theory, Methodology and Didactic Application of a Model for Translation-oriented Text Analysis*. 2nd edition. Translated from German to English by Nord, Christiane & Sparrow, Penelope. Amsterdam: Rodopi.
- O’Brien, Sharon (2012). Towards a dynamic quality evaluation model for translation. *Journal of Specialised Translation* 17, 55–77. https://jostrans.org/issue17/art_obrien.php (last accessed: 2021-07-08).
- ÖIF – Österreichischer Integrationsfonds (2021). AW: AW: Informationen für Masterarbeit [E-mail correspondence with a representative of the ÖIF from 2021-07-06]. See Appendix IV.

- Pöhhacker, Franz (1994). *Simultandolmetschen als komplexes Handeln*. Tübingen: Gunter Narr.
- Quah, Chiew Kin (2006). *Translation and Technology*. Basingstoke/New York: Palgrave Macmillan.
- Reinart, Sylvie (2014). *Lost in Translation (Criticism)? – Auf dem Weg zu einer konstruktiven Übersetzungskritik*. Berlin: Frank & Timme.
- Reiß, Katharina (1971). *Möglichkeiten und Grenzen der Übersetzungskritik*. München: Max Hueber
- Reiß, Katharina (1986). *Möglichkeiten und Grenzen der Übersetzungskritik – Kategorien und Kriterien für eine sachgerechte Beurteilung von Übersetzungen*. 3rd edition. München: Max Hueber.
- Reiß, Katharina & Vermeer, Hans (1984). *Grundlegung einer allgemeinen Translationstheorie*. Tübingen: Max Niemeyer Verlag.
- Risku, Hanna & Pichler, Theresa & Wieser, Vanessa (2017). Transcreation as a Translation Service: Process Requirements and Client Expectations. *Across Languages and Cultures* 18 (1), 53–77. <https://doi.org/10.1556/084.2017.18.1.3> (last accessed: 2021-11-23).
- RIUSS – Research Institute of United States Spanish (2021). Content of Hablamos Juntos. <https://riuss.org/content-of-hablamos-juntos> (last accessed: 2021-07-13).
- Rodríguez Rodríguez, Beatriz (2007). *Literary Translation Quality Assessment*. München: Lincom.
- SAE International (2021). Translation Quality Metric. https://saemobilus.sae.org/content/j2450_201608 (last accessed: 2022-01-14).
- Sager, Juan (1983). Quality and standards – the evaluation of translations. In: Picken, Catriona (ed.) *The Translator's Handbook*. London: Aslib, 121–128.
- Schmitt, Peter & Jüngst, Heike (eds.) (2007). *Translationsqualität*. Frankfurt am Main: Peter Lang.
- Schnierer, Madeleine (2019). *Qualitätssicherung | Die Praxis der Übersetzungsrevision im Zusammenhang mit EN 15038 und ISO 17100* (TransÜD. Arbeiten zur Theorie und Praxis des Übersetzens und Dolmetschens Band 105). Berlin: Frank & Timme.
- Secară, Alina (2005). Translation Evaluation – a State of the Art Survey. *Proceedings of the eCoLoRe/MeLLANGE Workshop*, Leeds, 39–44.

- TAUS – Translation Automation User Society (2021). Data Marketplace For Language Data Monetization and Acquisition. <https://datamarketplace.taus.net/> (last accessed: 2022-02-16).
- TAUS – Translation Automation User Society (2022). QT21 Project – Harmonized DQF-MQM Error Typology. <https://o.taus.net/qt21-project#harmonized-error-typology> (last accessed: 2022-10-03).
- televic education (2022). translationQ. <https://www.televic-education.com/en/translationq> (last accessed: 2022-02-16).
- Thelen, Marcel (2009). Quality Management for Translation. In: Forstner, Martin & Lee-Jahnke, Hannelore & Schmitt, Peter (eds.) *CIUTI-Forum 2008 – Enhancing Quality: Ways, Means, Methods*. Bern: Peter Lang, 195–212.
- Toulmin, Stephen & Rieke, Richard & Janik, Allan (1984). *An Introduction to Reasoning*. 2nd edition. New York: Macmillan.
- Txabarriaga, Rocío (2009). IMIA Guide on Medical Translation. <https://www.imiaweb.org/uploads/pages/823..pdf> (last accessed: 2021-11-03).
- Tyupa, Sergiy (2011). A Theoretical Framework for Back-Translation as a Quality Assessment Tool. *New Voices in Translation Studies* 7, 35–46. <https://www.iatis.org/images/stories/publications/new-voices/Issue7-2011/article-tyupa-2011.pdf> (last accessed: 2021-11-24).
- Universitas Austria (2017). Vorsicht ist besser als Nachsicht: Berufshaftpflichtversicherung für TranslatorInnen. <https://www.universitas.org/vorsicht-ist-besser-als-nachsicht-berufshaftpflichtversicherung-fuer-translatorinnen/> (last accessed: 2021-11-19).
- Universitas Austria (2019). Home – The Benchmark of Interpreting and Translation. <https://www.universitas.org/en/home/> (last accessed: 2021-11-05).
- UniVie – University of Vienna (2017). Curriculum für das Masterstudium Translation. https://transvienna.univie.ac.at/fileadmin/user_upload/z_translationswiss/Studium/Curricula/Curriculum_MA_Translation_Juni2018.pdf (last accessed: 2022-04-28).
- van den Broeck, Raymond (1985/2014). Second Thoughts on Translation Criticism – A Model of its Analytic Function. In: Hermans, Theo (ed.) *The Manipulation of Literature – Studies in Literary Translation*. Oxon/New York: Routledge Revivals, 54–62.
- Waddington, Christopher (2001). Different Methods of Evaluating Student Translations: The Question of Validity. *Meta* 46 (2), 311–325. <https://doi.org/10.7202/004583ar> (last accessed: 2021-11-23).

- Waddington, Christopher (2004). Should student translations be assessed holistically or through error analysis? *Lebende Sprachen* 49 (1), 28–35. [https://doi-org.uaccess.univie.ac.at/10.1515/LES.2004.28](https://doi.org.uaccess.univie.ac.at/10.1515/LES.2004.28) (last accessed: 2021-11-28).
- Wien Tourismus (2022). AW: Informationen für Masterarbeit [E-Mail correspondence with a representative of Wien Tourismus from 2022-03-07]. See Appendix IV.
- Wild, Diane & Grove, Alyson & Martin, Mona & Eremenco, Sonya & McElroy, Sandra & Verjee-Lorenz, Aneesa & Erikson, Pennifer (2005). Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) Measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. *Value in Health* 8 (2), 94–104. <https://doi.org/10.1111/j.1524-4733.2005.04054.x> (last accessed: 2021-11-24).
- Williams, Malcolm (1989). The Assessment of Professional Translation Quality: Creating Credibility out of Chaos. *TTR Traduction, terminologie, rédaction* 2 (2), 13–33. <https://doi.org/10.7202/037044ar> (last accessed: 2021-11-25).
- Williams, Malcolm (2001). The Application of Argumentation Theory to Translation Quality Assessment. *Meta* 46 (2), 326–344. <https://doi.org/10.7202/004605ar> (last accessed: 2021-11-23).
- Williams, Malcolm (2004). *Translation Quality Assessment: An Argumentation-Centred Approach*. Ottawa: University of Ottawa Press.
- Williams, Malcolm (2009). Translation Quality Assessment. *Mutatis Mutandis* 2 (1), 3–23. <https://revistas.udea.edu.co/index.php/mutatismutandis/article/view/1825> (last accessed: 2021-11-24).

Appendix I: Comparison between TQA approaches mentioned in Kingscott (2007) and Garvin (1984)

TQA approach mentioned in Kingscott (2007)	match with categories mentioned in Garvin (1984)		shortcomings of the approach mentioned in Kingscott (2007)
	<u>complete match</u>	<u>partial match(es)</u>	
<i>the authority figure approach:</i> A translated text is evaluated by a translator and/or revisor who is more experienced than the translator who translated the text and who gives his or her opinion on the quality of the translated text. (Kingscott 2007: 318)	<i>transcendent approach</i> (Garvin 1984: 25f.) ((taking into account what Nishino (2020: 156) says about this being applicable to experts who have developed an intuitive understanding of quality through experience)		This approach is subjective, hinging on the opinion of the evaluator alone. It is not easy to find such authority figures (i.e. more experienced translators/revisers) in sufficient numbers. (Kingscott 2007: 318f.)
<i>the qualifications of performer approach:</i> The competences and qualifications of translators (e.g. language proficiency, professional training, specialized knowledge in a certain field etc.) are stipulated in order to prevent unqualified translations. (Kingscott 2007: 319)		<i>manufacturing-based approach</i> (Garvin 1984: 26ff.) (being based on the same assumption that quality of process and quality of 'material' (in this case performer) are sufficient to result in quality of the product)	This approach does not allow judging the quality of the end product, since the end product is never evaluated. It is difficult to define what qualifications are necessary and/or sufficient for the translation of a certain text (Does a postgraduate-level education suffice for the translation of chemistry texts or is a degree in chemistry required?). (Kingscott 2007: 319f.)

TQA approach mentioned in Kingscott (2007)	match with categories mentioned in Garvin (1984)		shortcomings of the approach mentioned in Kingscott (2007)
	complete match	partial match(es)	
<i>the constrained procedures approach:</i> Constrained procedures are applied to select translators for the translation of a certain text, to the processes of translation and quality assurance etc. An example of this would be procedures as specified in ISO 9000. (Kingscott 2007: 320)	<i>manufacturing-based approach</i> (Garvin 1984: 26ff.)		Assuring the quality of the process can only guarantee consistency but not the quality of the end product itself. (Kingscott 2007: 320)
<i>the theoretical translation model approach:</i> The center of this approach is what Kingscott calls the classical theoretical model of translation, which is based on (1) comprehension of what intent the author of the ST pursued in the writing of the ST (on the side of the translator) and (2) equivalence of response to the TT (on the side of the target audience). (Kingscott 2007: 321)		<i>product-based approach</i> (Garvin 1984: 25ff.) (since there are criteria against which to measure the translation) <i>user-based approach</i> (Garvin 1984: 27)	No shortcomings are mentioned in Kingscott (2007), however, this approach seems subject to all the shortcomings of the equivalence-of-response models, i.a. that equivalence of response is hard to define and measure, that the target audience needs to be precisely known in order to compare responses, that reader response is not equally meaningful for all text types (e.g. advertisement vs. legal text) etc. (Colina 2008: 100; Colina 2009: 237).
<i>compliance with specifications-approach or PEXIS approach:</i> The translation product must comply with Purpose-oriented Explicit or Implicit Specifications. (Kingscott 2007: 321f.)	no complete match in Garvin (1984), but matches Harvey & Green's (1993) <i>fitness for purpose approach</i>	<i>user-based approach</i> (Garvin 1984: 27) (insofar as the user and/or customer defines the needs and specifications)	Not every customer provides specifications, but expectations and purpose can usually be deduced from the context and type of the document. (Kingscott 2007: 322)

Appendix II: Disadvantages of previous TQA models overcome (or not) by Colina's (2008) model

Type and name* of previously existing TQA model		Shortcoming(s) of previously existing TQA model**	Over-come by Colina's model?	If yes, how are shortcomings overcome? If no, why are they not overcome?
<i>experience-based</i>	SICAL (by the Translation Bureau of the Government of Canada)	1) fixed error categories (one-size-fits-all) 2) evaluation at word- and sentence-level only 3) lack of definitional clarity regarding the yardstick of 'central message'	Yes (all)	1) Colina's (2008: 107; 2015: [269f.]) model is adaptable. 2) Texts are evaluated as a whole (Colina 2008: 104; 2015: [266]). 3) Essential terms are defined and assumptions clearly stated (Colina 2008: 104; 2015: [243]).
<i>research-based</i>	Analytical model by Gouadec (1981)	1) lack of definitional simplicity regarding different error types 2) over-complexity of error coding (675 possible combinations) and thus limited applicability 3) almost exclusive focus on negative evaluation (error marking)	Yes (all)	1) Evaluation criteria are unambiguously and clearly described (descriptive statements). Their number is limited and each statement is easily attributable to its criterion (Colina 2008: 128ff.). 2) The evaluation criteria are limited in number (four in Colina (2008: 106f.), though this number is adjustable). 3) The model takes both positive and negative aspects into account and is not based on error marking (Colina 2008: 106).
<i>research-based</i>	MQM-DQF error typology (by TAUS)	1) based on error counting and error categorization 2) only partially implemented positive evaluation (marking but undefined score impact)	Yes (all)	1) The model is not based on error counting, and evaluation parameters are not merely error categories but allow for mixed positive and negative evaluation (Colina 2008: 106f.). 2) There is no need for separate implementation of positive evaluation, since positive aspects are taken into account as much as negative aspects are, due to the nature of the model (Colina 2008: 106; Hablamos Juntos 2009).

<i>research-based</i>	PIE method (calibration method) by Kockaert & Segers (2017)	1) dichotomy of judgement (correct or incorrect, no error grading) 2) partial assessment only (disregard of all non-preselected parts of the text) 3) reliance on statistical calculations with little room for human judgement 4) limited applicability due to complexity and the method being time-consuming	Yes (all)	1) Descriptive statements allow for a continuum of grading (positive and negative aspects) (Colina 2008: 107, 123). 2) Texts are evaluated as a whole (evaluators read both the ST and TT in their entirety) (Colina 2015: [266]). 3) The model allows for evaluators to bring their judgement to bear while still quantifying results and limiting subjectivity by providing criteria and descriptors (Colina 2015: [248]). 4) The model is easy to learn and apply (Colina 2008: 121) and allows for reasonably timely evaluation (average rating time in Colina (2008: 118) for a text of 1,200 words was between 44 and 50 minutes).
<i>research-based</i>	Mathematical model by Fan (1990)	• no or highly limited practical applicability	Yes	• Colina's (2008; 2009) model is intended and has been tested for practical applicability. Its applicability is only limited insofar as that large amounts of text presumably entail longer evaluation times, which is a common problem in TQA (Drugan 2013: 52; Kurz 2018: 20).
<i>research-based</i>	Argumentation approach by Bittner (2020)	• limited applicability (due to necessary knowledge and the approach being time-intensive)	Yes	• Colina's model is easy to learn and apply (Colina 2008: 121) and allows for reasonably timely evaluation (average rating time in Colina (2008: 118) for a text of 1,200 words was between 44 and 50 minutes).
<i>research-based</i>	TQA model by Gutt (2000/2014)	• no or highly limited practical applicability (due to lack of quantification of results, the model being time-consuming etc.)	Yes	• Colina's (2008; 2009) model is intended and has been tested for practical applicability. Its applicability is only limited insofar as that large amounts of text presumably entail longer evaluation times, which is a common problem in TQA (Drugan 2013: 52; Kurz 2018: 20). The model allows for results to be quantified in relation to the translation brief (Colina 2008: 107, 130).
<i>reader-response</i>	TQA model by Carroll (1966)	1) partial evaluation (raters never read the entire ST or TT) 2) difficulties in measuring reader response and limited importance of reader response for certain text types 3) exclusive focus on one aspect of quality	Yes (all)	1) Texts are evaluated as a whole, and evaluators are instructed to read both ST and TT (Hablamos Juntos 2009: 5). 2) Reader response (as part of the criterion Functional and Textual Adequacy) is only one of the aspects of quality that Colina's (2008; 2009) model recognizes (see 3)). 3) Colina's (2008: 105) model is componential.

<i>equivalence-based</i>	Back-translation by Brislin (1969)	1) reliance on the problematic notion of equivalence 2) limited validity (the method does not serve to reliably find errors in the TT since many factors may lead to seeming equivalence between ST and BT) 3) does not allow functionalist aspects to be taken into account (since alterations would be penalized as mismatches) 4) one-size-fits-all approach (no possibility for adaptation, no defined evaluation criteria)	Yes (all)	1) Colina's (2008; 2009) model is functionalist, not based on an a-priori notion of equivalence. One could argue, however, that the functionalist approach still includes the notion of equivalence, only that its degree and nature are drawn from the requirements of the brief (Colina 2015: [243]). 2) Testing showed that Colina's (2008: 114; 2009: 248) model produces reasonably reliable and valid results. 3) see 1) 4) Colina's (2008: 107; 2015: [269f.]) model is adaptable. Both evaluation criteria and their weighting can be adjusted and/or changed in accordance with the evaluation context.
<i>equivalence-based</i>	Model by van den Broeck (1985/2014)	1) limited applicability (due to the necessity to guess the translator's thoughts, the lack of quantification of results) 2) no defined evaluation criteria (resulting in high levels of subjectivity) 3) exclusive focus on one aspect of quality	Yes (all)	1) Colina's (2008; 2009) model is intended and has been tested for practical applicability. The model allows for results to be quantified in relation to the translation brief (Colina 2008: 107, 130). It does not examine the translation process, only the translation product (Colina 2008: 105). 2) Colina's model (2008: 128ff.) includes defined (described) evaluation criteria, limiting subjectivity. 3) Colina's model (2008: 105) is componential.
<i>functional-textual</i>	TQA model by al-Qinai (2000)	1) no quantification of results (also unclear how to counterbalance negative and positive results) 2) no weighting of categories 3) limited applicability (due to how time-consuming the analysis is)	Yes (all)	1) Colina's (2008: 106f.) model allows for quantification of results in relation to the translation brief. The nature of the model renders it unnecessary to balance negative and positive aspects on a point-based system. 2) Evaluation criteria can be weighted in accordance with the brief, though this need not be done if the brief does not demand it (Colina 2008: 107, 130). 3) Colina's (2008; 2009) model is intended and has been tested for practical applicability. Its applicability is only limited insofar as that large amounts of text entail longer evaluation times, which is a common problem in TQA (Drugan 2013: 52; Kurz 2018: 20).

<i>functional-pragmatic</i>	Functional-pragmatic model by House (1977; 1997; 2014)	1) limited practical applicability (due to lack of quantification of results) 2) reliance on the notion of equivalence (problematic since TT might be intended for different function than ST)	Yes (all)	1) Colina's (2008; 2009) model is intended and has been tested for practical applicability. Results can be quantified in relation to the translation brief via the score sheet (Colina 2008: 106, 130). 2) Colina's (2008; 2009) model is functionalist, not based on an a-priori notion of equivalence. One could argue, however, that the functionalist approach still includes the notion of equivalence, only that its degree and nature are drawn from the requirements of the brief (Colina 2015: [243]). As long as the brief provides for a different TT function, this divergence can be incorporated into the evaluation without problems (Colina 2008: 101, 104).
<i>functional-textual</i>	TQA model by Ammann (1990)	1) reliance on the vague scenes-and-frames semantic 2) lack of quantification of results	Yes (all)	1) The evaluation method and evaluation criteria are clearly described (Colina 2008: 128ff.; Hablamos Juntos 2009). 2) Results can be quantified in relation to the translation brief via the score sheet (Colina 2008: 130).
<i>text-based</i>	Textual TQA model by Reiß (1986)	1) reliance on the controversial notion of equivalence 2) lack of definitional clarity regarding the concept of 'optimum equivalence' 3) limited applicability (not suitable for all types of TT)	Yes (1, 2) Probably (but unproven) (3)	1) Colina's (2008; 2009) model is functionalist, not based on an a-priori notion of equivalence. One could argue, however, that the functionalist approach still includes the notion of equivalence, only that its degree and nature are drawn from the requirements of the brief (Colina 2015: [243]). 2) Essential terms are defined and assumptions clearly stated (Colina 2008: 104f.). 3) Colina's (2008; 2009) model is intended and has been tested for practical applicability. Though testing involved only one text type (health education material) and the model is originally intended for the use in health care organizations (Hablamos Juntos 2009: 1f.), the functionalist nature of the model allows the assumption that it could be used for other types of TT as well. This is substantiated by examples given by Colina (2008: 124) in which she mentions using the tool on creative texts.

<i>text-based</i>	Translation-relevant text analysis model by Nord (2005)	1) over-complexity of analysis, resulting in limited applicability 2) lack of quantification of results 3) disproportionate focus on the ST	Yes (all)	1) Colina's (2008; 2009) model is intended and has been tested for practical applicability. Its applicability is only limited insofar as that large amounts of text entail longer evaluation times, which is a common problem in TQA (Drugan 2013: 52; Kurz 2018: 20). 2) Results can be quantified in relation to the translation brief via the score sheet (Colina 2008: 130). 3) Evaluators are instructed to read both the ST and the TT (Hablamos Juntos 2009: 5). The object of the evaluation is the TT (Colina 2008: 105).
<i>text-based</i>	ARTRAQ model by Williams (2004; 2009)	1) complexity of model (appears difficult to learn) 2) limited applicability (nature of model precludes the possibility of sampling) 3) questionable importance of argument structure for certain texts and text types 4) exclusive focus on one aspect of quality	Yes (1, 3, 4) No/unclear (2)	1) Testing showed that Colina's (2008: 121) model is easy to learn. 2) Colina's (2015: [266]) model is intended to be used for the evaluation of entire texts. The author does not state whether sampling would be possible; however, the nature of the model and its evaluative process do not seem to preclude sampling. That being said, all the disadvantages and risks of sampling (missing major errors, not getting an impression of the entire text etc., see chapter 2.2.6) would come to bear, as they would with any other model. 3) Colina's (2008: 107) model does not rely on argument structure as a criterion. Furthermore, criteria can be adjusted in accordance with the evaluation context. 4) Colina's (2008: 105) model is componential.
<i>functional-textual</i>	Teleological model by Larose (1994)	• limited applicability (due to lack of criteria)	Yes	• Colina's (2008: 107, 128ff.) model includes clearly described criteria, which can be changed in accordance with the translation brief.

<i>corpus-based</i>	Corpus-based model by Bowker (2001)	1) lack of quantification of results 2) limited applicability (due to the need to construct an evaluation corpus, which is time-intensive)	Yes (all)	1) The model allows for quantification of results via the score sheet (Colina 2008: 130). Additionally, the original model from 2008 was later adapted (Colina 2009: 244) to include a cover sheet, where evaluators can choose between different recommendations for requesters based on their evaluation (e.g. “Publish and/or use as is” or “Redo translation” (Colina 2009: 258) (see chapter 3.1). 2) Colina’s (2008; 2009) model is intended and has been tested for practical applicability. Testing revealed that the model allows for reasonably timely evaluation (average rating time in Colina (2008: 118) for a text of 1,200 words was between 44 and 50 minutes), since it does not require any preparation beyond filling in the translation brief (Colina 2008: 128ff.).
---------------------	--	---	-----------	--

*Please note that the names of the TQA models given in this table do not necessarily feature verbatim in the work of the author(s) cited next to the model. Some authors have neglected to name their model, these models are therefore labeled “TQA model by [author]”. In addition, some central characteristics of a model were sometimes added to the name if one was not given by the author, as is the case, for example, with “Translation-relevant text analysis model by Nord (2005)”.

**This table merely constitutes an overview of shortcomings already discussed in chapter 2.2. Refer to this chapter for the precise citations on each shortcoming, since they cannot be given in the table due to spatial reasons.

Appendix III: Texts

Text 1 (German original/ST)⁹⁹

Informationen zur Corona-Schutzimpfung



Alle Personen, die in Österreich wohnen, können sich in Österreich kostenlos gegen das Coronavirus impfen lassen. Die Anmeldung zur Corona-Schutzimpfung erfolgt über die Impfplattformen der Bundesländer unter www.oesterreich-impft.at/ietzt-impfen



Im Rahmen der Onlineberatung „Sprechstunde Corona: Lockdown und Impfung“ informiert der ÖIF zu aktuell geltenden Maßnahmen sowie zur Corona-Schutzimpfung. Die Beratungen werden für Menschen mit geringen Deutschkenntnissen

gedolmetscht. Die Teilnahme ist kostenlos und ohne Anmeldung möglich.

Alle Termine sind unter www.integrationsfonds.at/onlinekurse zu finden.



Telefonisch stehen Ihnen die Mitarbeiterinnen und Mitarbeiter des ÖIF von Montag bis Donnerstag von 10:00 – 14:00 Uhr sowie Freitag von 09:00 – 13:00 unter +43 1/715 10

51 - 263 für Fragen zum Coronavirus zur Verfügung!



Nähere Informationen zu Corona oder zur Corona-Schutzimpfung finden Sie auf der Seite des Sozialministeriums, unter [Informationen zur Corona-Schutzimpfung in Österreich \(sozialministerium.at\)](http://Informationen%20zur%20Corona-Schutzimpfung%20in%20%C3%96sterreich%20(sozialministerium.at)) sowie unter [Coronavirus \(sozialministerium.at\)](http://Coronavirus%20(sozialministerium.at)).

Hier geht es zu den Onlineberatungen rund um COVID-19.

COVID-19 Impfpflicht ab Februar 2022

Eine Impfpflicht bedeutet, dass alle Personen in Österreich verpflichtet sind, sich gegen das Coronavirus impfen zu lassen (wenn nicht bereits geimpft), wenn die Behörde das vorschreibt. Sonst kann es zu einer Strafe kommen.

Folgende Bestimmungen gelten seit 5. Februar 2022. Nähere Informationen unter [Informationen zur Impfpflicht gegen das Coronavirus in Österreich \(sozialministerium.at\)](http://Informationen%20zur%20Impfpflicht%20gegen%20das%20Coronavirus%20in%20%C3%96sterreich%20(sozialministerium.at)).

⁹⁹ Note that the layout of the texts and the distribution of elements on the page as depicted here may vary slightly from the layout presented to evaluators, since page margins needed to be adjusted in the thesis to allow for printing.

- Gilt seit 5. Februar 2022 bis 31. Jänner 2024 – Strafen ab 16. März 2022
- Gilt für alle Personen ab 18 Jahren, die in Österreich wohnen.
Ausgenommen sind: Das Vorliegen von einem Ausnahmegrund ist von einem Arzt zu bestätigen.
 - Personen unter 18 Jahren
 - Schwangere
 - Personen, die aus gesundheitlichen Gründen, nicht geimpft werden können
 - Personen, die trotz Impfungen keine oder nicht ausreichend viele Antikörper gebildet haben (z.B. bei Allergien gegen Corona-Impfstoffe)
 - Personen, die in den letzten 180 Tagen eine Corona-Infektion überstanden haben
- Das Vorliegen von einem Ausnahmegrund ist von einem Arzt zu bestätigen.
- Die Impfpflicht umfasst alle empfohlenen Corona-Schutzimpfungen, derzeit sind das:
 - Erstimpfung
 - Zweitimpfung
 - Drittimpfung
- Anerkannte Impfstoffe zur Erfüllung der Impfpflicht sind von:
 - BioNTech-Pfizer - Comirnaty
 - Moderna Biotech - Spikevax
 - AstraZeneca - Vaxzevria
 - Johnson & Johnson - Covid-19 Vaccine Janssen
 - Novavax - Nuvaxovid
- 2 chinesische Impfstoffe von:
 - Sinopharm - Vero Cell
 - Sinovac - CoronaVac
- 3 indische Impfstoffe von:
 - Bharat Biotech - COVAXIN
 - Serum Institute of India - COVOVAX und Covishield
- Personen, die zwei Impfungen mit keinem der oben angeführten Impfstoffe im Ausland erhalten haben, gelten in Österreich als erstgeimpft.
- Bei Corona-Infektion zwischen den Impfungen, hat spätestens 180 Tage ab dem Tag der Probenahme (positiver PCR-Test) eine Impfung zu erfolgen.
- Bei Nichteinhaltung droht Geldstrafe (bis 600 Euro bzw. bis 3.600 Euro bei Verfahren vor Gericht)

Text 1 (English translation/TT)

Information on the coronavirus vaccination



All those who live in Austria may receive a free vaccine in Austria to protect them against coronavirus. Registration for the coronavirus vaccine must be carried out via the vaccination platform of the federal states at [www.oesterreich-impft.at/ietzt-](http://www.oesterreich-impft.at/ietzt-impfen)

[impfen](http://www.oesterreich-impft.at/ietzt-impfen).



As part of the online consultation “Coronavirus chat: lockdown and vaccination”, the ÖIF provides information on the measures currently in place as well as on the vaccination to protect against coronavirus. For people with insufficient knowledge of

German, the consultation sessions will be interpreted. Participation is free of charge and does not



require registration. All dates can be found at www.integrationsfonds.at/onlinekurse.

The ÖIF staff are available for you via telephone from Monday to Thursday from 10:00 – 14:00 as well as on Friday from 09:00 – 13:00 at +43 1/715 10 51 - 263 for all

your questions regarding coronavirus!



More detailed information on coronavirus and the coronavirus vaccination is available on the webpage of the Ministry for Social Affairs under [Informationen zur Corona-Schutzimpfung in Österreich \(sozialministerium.at\)](http://Informationen_zur_Corona-Schutzimpfung_in_Oesterreich_(sozialministerium.at)) (*information on the coronavirus vaccination in Austria, German language*) as well as under [Coronavirus \(sozialministerium.at\)](http://Coronavirus_(sozialministerium.at)).

Click [here](#) for online advice on all aspects of COVID-19.

COVID-19 compulsory vaccination from February 2022

Compulsory vaccination means that all persons in Austria are obligated to get themselves vaccinated against coronavirus if they are not yet vaccinated and the authority so specifies. Fines are possible in the event of non-compliance. The following provisions have been in force since 5 February

2022: Further information is available at [Information on compulsory vaccination against coronavirus in Austria \(sozialministerium.at\)](http://Information_on_compulsory_vaccination_against_coronavirus_in_Austria_(sozialministerium.at)).

- Applies from 5 February 2022 until 31 January 2024; violations will be subject to fines from 16 March 2022 onwards
- Applies for all persons aged 18 and over who live in Austria.
The exceptions to this are:
 - Persons under the age of 18

- Women who are pregnant
- Persons who cannot be vaccinated for health reasons
- Persons who have not formed any or have only formed an insufficient number of antibodies despite vaccinations (e.g. in the event of allergies to the coronavirus vaccines)
- Persons who have recovered from a coronavirus infection in the last 180 days
- Reasons that justify an exception must be confirmed by a doctor.
- Compulsory vaccination incorporates all recommended coronavirus vaccinations; these are currently:
 - Initial vaccination
 - Second vaccination
 - Third vaccination
- Vaccines recognised to meet the requirements of compulsory vaccination are:
 - BioNTech-Pfizer - Comirnaty
 - Moderna Biotech - Spikevax
 - AstraZeneca - Vaxzevria
 - Johnson & Johnson - Covid-19 Vaccine Janssen
 - Novavax - Nuvaxovid
 - 2 Chinese vaccines:
 - Sinopharm - Vero Cell
 - Sinovac - CoronaVac
 - 3 Indian vaccines:
 - Bharat Biotech - COVAXIN
 - Serum Institute of India - COVOVAX and Covishield
- People who received two vaccinations abroad, but none of the vaccines listed above, are considered to have received only the first dose of vaccination.
- In the event of a coronavirus infection between the vaccinations, a vaccination must be given at the latest 180 days from the sample (positive PCR test).
- If this requirement is not complied with, a fine is possible (up to 600 euros or up to 3,600 euros in the case of proceedings in court)

Testmöglichkeiten und To-dos bei einer Infektion

Für Gäste gibt es viele Möglichkeiten, sich testen zu lassen. Sollten Sie in Wien erkranken, wird bestens für Sie gesorgt.

Testmöglichkeiten

Österreich bietet so viele (kostenlose) Testmöglichkeiten wie kaum ein anderes Land weltweit - auch für Gäste, die keine österreichische Sozialversicherungsnummer besitzen. Nutzen Sie also die Gelegenheit, lassen Sie sich testen und schauen Sie auf sich und Ihre Mitmenschen.

WAS SIE AKTUELL IN WIEN BEACHTEN MÜSSEN

Kostenlose Testmöglichkeiten

- ◆ Aktion [Alles gurgelt!](#) der Stadt Wien: Viele Wiener Hotels bieten **kostenlose PCR-Gurgeltests** für ihre Gäste an. Diese können vor Ort gemacht und entweder wieder an der Rezeption abgegeben oder zu einer der Abgabestationen (Billa, Billa Plus, Bipa, Penny, BP-Merkur Inside, Jet-Billa Stop & Shop, Shell-Billa unterwegs, ÖBB Reisezentren Wien Hauptbahnhof und Wien Meidling, ÖBB Reisebüros in Wien Hauptbahnhof und Wien Westbahnhof) gebracht werden. Das Ergebnis liegt innerhalb von 24 Stunden vor und gilt 48 Stunden.
Bitte fragen Sie an der Rezeption, ob Ihr Hotel an der Aktion teilnimmt.
- ◆ [Alles gurgelt! für Tourist:innen](#)
- ◆ Tourist:innen können sich in den [Teststraßen und Gurgelboxen der Stadt Wien](#) kostenlos testen lassen (PCR-Tests und teilweise Antigen-Schnelltests). Eine Anmeldung ist erforderlich!
Bitte beachten Sie, dass manche Teststraßen nur ohne Symptome besucht werden dürfen!
- ◆ In Apotheken gibt es darüber hinaus für Gäste mit österreichischer E-Card/Sozialversicherung die Möglichkeit, sich kostenlos testen zu lassen ([Vor Anmeldung notwendig](#)).

Kostenpflichtige Testmöglichkeiten

- ◆ Zahlreiche private Labors in Wien bieten Testmöglichkeiten mit PCR- und Antigen-Tests an. Eine Auswahl bietet eine [Liste des Gesundheitsministeriums](#).
- ◆ Gäste können einen PCR- oder Antigen-Test bei zahlreichen [Wiener Apotheken](#) in Wien durchführen lassen. Voranmeldung ist nötig.
- ◆ Schnelle Ergebnisse von PCR-Tests innerhalb weniger Stunden liefern Unternehmen wie [FasTest](#), [ViroTrust](#), [DermaCare](#), [PCRTestwien.at](#) mit Standorten in der Innenstadt.
- ◆ Am Flughafen Wien besteht im [Health Center](#) die Möglichkeit sowohl einen PCR-Test wie auch einen Antigen-Schnelltest (Roche SARS-CoV-2 Rapid Antigen Test) zu machen.

- ◆ Hotelgäste können auch direkt im Hotel einen PCR-Test über das Service [DeliveryTest](#) ➞ durchführen lassen. Geschulte Boten nehmen die Proben ab.
- ◆ Das [Hotel Sacher](#) ➞ und das [Palais Hansen Kempinski](#) ➞ bieten ihren Gästen die Möglichkeit einer Testung auf Covid-19 (PCR-Schnelltest oder Antigentest) im Hotel. Näheres zum Angebot erfahren Sie auf den Webseiten der Hotels bzw. bei direkter Kontaktaufnahme.

Covid-19-Tests in Wien

GALERIE



© HANS PUNZ / APA / picturedesk.com

Verdacht auf Infektion

Sollten Sie während Ihres Aufenthaltes erkranken, werden Sie selbstverständlich versorgt und können kostenlose medizinische Behandlung in Anspruch nehmen. Bitte beachten Sie folgende Vorgehensweise:

- ◆ Wenn Sie den Verdacht haben, sich infiziert zu haben, bleiben Sie bitte in der Unterkunft und meiden Sie Menschenansammlungen sowie öffentliche Verkehrsmittel.
- ◆ Rufen Sie die Gesundheits-Hotline 1450 an: Sie werden von medizinischem Fachpersonal telefonisch betreut und erhalten weitere Informationen.
- ◆ Benutzen Sie auch den [Symptomchecker](#) ➞.

Krankheitsfall während Ihres Aufenthalts für Wien

- ♦ Wer trotz aller Vorkehrungen im Urlaub in Wien erkrankt, kann auf die umfangreichen Hygiene- und Sicherheitsmaßnahmen der Stadt vertrauen.
- ♦ Wer sich aufgrund einer behördlichen Maßnahme bzw. eines behördlichen Bescheids (auf Basis des Epidemiegesetzes) in eine der Einrichtungen der Stadt Wien in Quarantäne begibt, muss für die Kosten nicht selbst aufkommen.

Freitesten

- ♦ Einreisende, die sich nach fünf Tagen aus der Quarantäne freitesten lassen wollen (mittels PCR- oder Antigen-Test), können dies in allen hier angeführten Einrichtungen tun. Für die Dauer der Testung darf die Quarantäne unterbrochen werden.
- ♦ Für Covid-Infizierte beträgt die Quarantänefrist in Wien mindestens 10 Tage.
- ♦ Seit **8. Jänner 2022 gilt für Kontaktpersonen**: Es gibt es keine Unterscheidung mehr zwischen K1- und K2-Personen. Man ist keine Kontaktperson mehr, wenn man dreimal geimpft ist oder wenn alle Beteiligten eine FFP2-Maske getragen haben. Das gilt auch für Kinder, die noch keine dritte Impfung erhalten. **Für alle, die als Kontaktpersonen eingestuft werden, gilt: Freitesten ist ab dem fünften Tag mit einem PCR-Test möglich.**

Testing options and to-dos in the event of an infection

There are many options for guests to get themselves tested. Should you fall sick in Vienna, you'll be well looked after.

Testing options

Austria offers more (free) options for getting tested than virtually any other country in the world – including for guests without an Austrian social security number. So take the opportunity to get yourself tested and take care of yourself and those around you.

WHAT TO CONSIDER IN VIENNA AT THE MOMENT

Free of charge testing options

- ◆ The City of Vienna's [Alles gurgelt!](#) campaign: Many Viennese hotels are offering **free PCR gargle tests** for their guests. These can be taken on site and either returned to reception or taken to one of the drop-off stations (Billa, Billa Plus, Bipa, Penny, BP-Merkur Inside, Jet-Billa Stop & Shop, Shell-Billa unterwegs, ÖBB travel centers at Vienna Main Station and Vienna Meidling Station, ÖBB travel agencies in Vienna Main Station and Vienna Westbahnhof Station). The result is available within 24 hours and is valid for 48 hours.
- ◆ Please ask at the reception if your hotel participates in the promotion.
- ◆ [Everything gargles! for tourists](#)
- ◆ Tourists can get themselves tested in the [test centers and gargle boxes operated by the city of Vienna](#) for free (PCR tests and some antigen quick tests). An appointment is required!
- ◆ **Please note that some test centers may only be visited if you do not have any symptoms!**
- ◆ Guests with an Austrian E-Card/social security cover can also have themselves tested free of charge in pharmacies ([registration required](#)).

Fee-based testing options

- ◆ Numerous private laboratories in Vienna offer the possibility to get PCR and antigen tests. You can find a selection in a [list from the Ministry of Health](#).
- ◆ Guests can have a PCR or antigen test performed at any one of dozens of [drugstores](#) in Vienna. An appointment is required.
- ◆ Fast results from PCR tests are delivered within a few hours by companies such as [FasTest](#), [ViroTrust](#), [DermaCare](#), [PCRTTestwien.at](#), with locations in the city center.
- ◆ The [Health Center](#) at Vienna Airport offers the possibility to get a PCR test or a rapid antigen test (Roche SARS-CoV-2 Rapid Antigen Test).
- ◆ Hotel guests can also have a PCR test performed directly in the hotel via the service [DeliveryTest](#). Trained couriers collect the samples.

- ◆ The [Hotel Sacher](#) and the [Palais Hansen Kempinski](#) offer their guests the possibility to get tested for COVID-19 (rapid PCR test or antigen test) in the hotel. For more information, please refer to the websites of the hotels or contact them directly.

COVID-19 tests in Vienna

GALLERY



© HANS PUNZ / APA / picturedesk.com

Suspected infection

Should you fall sick during your stay, you will be taken care of and can avail yourself of free medical treatment. Please pay attention to the following procedures:

- ◆ If you suspect that you are infected, please stay in your room and avoid gatherings of people and public transport.
- ◆ Call the health hotline on 1450: You will be looked after by medical staff over the phone and given further information.
- ◆ You should also use the [symptom checker](#).

Sickness during your stay in Vienna

- ◆ If, despite taking every precaution, you fall sick when on vacation in Vienna, you can always trust in the city's extensive hygiene and safety measures.
- ◆ If you have to quarantine yourself in one of Vienna's facilities on the basis of an official measure or government notice (in accordance with the Epidemics Act), you won't have to pay the costs of doing so.

Testing to leave quarantine early

- ♦ Incoming travelers who want to release themselves from quarantine after five days by getting themselves tested (with a PCR or antigen test) can do so at any of the facilities listed here. Quarantine may be interrupted in order to go for the test.
- ♦ For Covid-infected persons, the quarantine period in Vienna is at least 10 days.
- ♦ **The following applies to contact persons since January 8, 2022:** There is no longer any distinction made between K1 and K2 people. You are no longer considered to be a contact person when you have been vaccinated three times or if everyone involved was wearing an FFP2 mask. That also applies to children who have not yet received a third vaccination. **The following applies to everyone classified as contact persons: Testing to leave quarantine early is possible from the fifth day with a negative PCR test.**



Coronavirus: So trage und verwende ich meine FFP2-Maske richtig



Beim Kauf einer FFP2-Maske schau ich auf Folgendes:

- Bezeichnung als FFP2-Maske*
- CE-Prüfzeichen
- EN-Kennzeichnung
- Vierstellige Nummer zur Identifikation des zertifizierenden Testinstituts
- kein Ausatemventil



Bevor ich die FFP2-Maske aufsetze, wasche ich meine Hände gründlich mit Seife oder desinfiziere sie.



Beim Auspacken der FFP2-Maske achte ich darauf, die Innenseite der Maske nicht zu berühren. Ich forme den Nasenbügel mit den Zeigefingern vor.



Ich fixiere die Schlaufen hinter den Ohren, ziehe zuerst die Maske über das Kinn und dann über die Nase. Den Nasenbügel passe ich mit beiden Zeigefingern fest an.



Ich überprüfe, ob die FFP2-Maske dicht am Gesicht anliegt. Es darf keine Luft über die Ränder austreten.



Während des Tragens berühre oder verschiebe ich die FFP2-Maske nicht.



Nachdem ich meine FFP2-Maske abgelegt habe, wasche ich meine Hände gründlich mit Seife oder desinfiziere sie.



Ich transportiere die FFP2-Maske nicht in meiner Hosentasche oder in einer Tasche, in die ich häufig greife. Am besten benutze ich einen separaten Beutel.



Ich lasse die FFP2-Maske nicht an meinem Hals baumeln, denn dadurch können sich Keime auf der Haut verteilen.



Ich nehme die FFP2-Maske nur von hinten an den Bändern vom Gesicht ab und berühre sie sonst nicht.



FFP2-Masken sind grundsätzlich Einwegmasken. Bei sorgfältigem Gebrauch kann sie mehrmals verwendet werden. Ist die FFP2-Maske durchfeuchtet oder kaputt, entsorge ich sie im Mistkübel.



Trotz FFP2-Maske beachte ich weiterhin allgemeine Hygiene- und Schutzmaßnahmen sowie den Mindestabstand von 2 Metern.



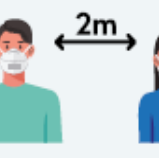
Einen umfassenden Überblick über alle bundesweit gültigen Maßnahmen finde ich auf sozialministerium.at/faqcoronavirus.

* Als mindestens gleichwertig gelten folgende Standards: FFP3 (Europa), N95 (NIOSH-42C FR84, USA), P2 (AS/NZ 1716:2012, Australien/Neuseeland), KF94 (Korea 1st Class KMOEL-2017-64), DS (Japan JMHLW-Notification 214,2018), KN95 (GB2626-2006, China)

Text 3 (English translation/TT)



Coronavirus: How to wear and use your FFP2 mask

 When buying an FFP2 mask, look out for the following: • Labelled as FFP2 mask* • CE marking • EN identification • Four-digit number to identify the certifying testing institution • No exhalation valve	 Before putting on your FFP2 mask, wash your hands thoroughly with soap or sanitise them.	 When unwrapping the FFP2 mask, make sure you don't touch the inside of the mask. Use your index fingers to push the nose wire into shape.	 Position the loops behind your ears and pull the mask over your chin and then over your nose. Use your index fingers to press the nose wire firmly into place.
 Check that the FFP2 mask fits closely on your face. No air should escape round the edges.	 Do not touch or adjust your FFP2 mask while wearing it.	 After taking off your FFP2 mask, wash your hands thoroughly with soap or sanitise them.	 Do not carry your FFP2 mask around in your trouser pocket or a bag that you use frequently. Ideally, use a separate bag.
 Do not leave the FFP2 mask hanging round your neck because this could allow microorganisms to get on to your skin.	 Always remove the FFP2 mask from the back, using the straps, and otherwise do not touch it.	 FFP2 masks are essentially single-use masks. If used carefully, they can be used more than once. If the FFP2 mask becomes moist or damaged, dispose of it with household refuse.	 Even when wearing an FFP2 mask, continue to observe general hygiene and protective measures and the 2-metre social distancing rule.



A comprehensive summary of all measures applicable throughout Austria is available at sozialministerium.at/faqcoronavirus.

* The following standards are regarded as at least equivalent: FFP3 (Europe), N95 (NIOSH-42C FR84, USA), P2 (AS/NZ 1716:2012, Australia/New Zealand), KF94 (Korean 1st Class KMOEL-2017-64), DS (Japan JMHLW-Notification 214,2018), KN95 (GB2626-2006, China)

Appendix IV: Correspondence

E-mail correspondence with the ÖIF, corresponding to the bibliographic entry found under ÖIF (2021) (under secondary sources)

Liebe Frau Hengsberger,
untenstehend finden Sie die um die Antworten ergänzten Fragen. Melden Sie sich jederzeit, sollten Sie noch etwas brauchen.

Beste Grüße

■■■■■■■■■■ [representative of the ÖIF]¹⁰⁰

stv. Leiterin Team Öffentlichkeitsarbeit

Österreichischer Integrationsfonds

A-1030 Wien, Schlachthausgasse 30

Mehr zum Thema Integration in Österreich:

www.integrationsfonds.at

INTEGRATION = ÖSTERREICH

Von: Sonja Hengsberger <a01601237@unet.univie.ac.at>

Gesendet: Freitag, 2. Juli 2021 14:49

Betreff: Re: AW: Information für Masterarbeit

Sehr geehrte Frau ■■■■■■■■!

Vielen Dank für Ihre rasche Antwort. Folgende Fragen wären für meine Arbeit relevant:

Alle Fragen drehen sich um diese Website (<https://www.integrationsfonds.at/coronainfo/>) bzw. deren englische Version, wobei ich davon ausgehe, dass es sich bei der englischen Version um eine Übersetzung der deutschen Version handelt.

Ist meine Annahme, dass es sich bei der deutschsprachigen Version der Website um die Originalversion und bei der englischen um die Übersetzung handelt korrekt?

Genau. Die deutschsprachige Version ist die Originalversion.

Aus welchem Grund wurden die Informationen auf dieser Website vom Integrationsfond in die 17 angebotenen Sprachen übersetzt? (möchte man dadurch z.B. dem öffentlichen Informationsauftrag gerecht werden oder richtet sich das nicht-deutschsprachige Informationsangebot speziell an Menschen, die der deutschen Sprache nicht mächtig sind etc.) Es existiert ja meines Wissens in Österreich kein Gesetz, dass öffentliche Stellen zur Bereitstellung von Informationen in anderen Sprachen als dem Deutschen zwingt.

Der Österreichische Integrationsfonds (ÖIF) ist ein Fonds der Republik Österreich und ein Partner des Bundes in der Integrationsförderung. In seiner Tätigkeit richtet sich der ÖIF an Asylberechtigte, subsidiär Schutzberechtigte und Drittstaatsangehörige sowie Menschen mit Migrationshintergrund. Der ÖIF hat in seiner täglichen Arbeit naturgemäß vielfach mit Personen zu tun, die Deutsch noch nicht oder nur in geringem Ausmaß sprechen. Aufgrund der Relevanz des Thema und der Bedeutung der Einhaltung der Corona-Maßnahmen durch alle Menschen in Österreich, unabhängig vom Grad

¹⁰⁰ Names of correspondents have been blackened out for reasons of data protection.

der Sprachbeherrschung, hat der ÖIF in Absprache mit anderen öffentlichen Stellen die Zurverfügungstellung von mehrsprachigen Angeboten gebündelt.

Von wem (Personen, Abteilung, Organisation o.Ä.) wurde die Übersetzung durchgeführt?

Interessant ist für mich vor allem, ob es sich dabei um eine eigenen Übersetzungsabteilung oder freiberufliche ÜbersetzerInnen handelt, oder ob Personen ohne Übersetzer Ausbildung diese durchgeführt haben. Ich erwarte mir selbstverständlich keine Name oder dergleichen, da dies sicher datenschutzrechtlich problematisch wäre.

Der ÖIF greift auf unterschiedliche professionelle Übersetzungsagenturen zurück.

Wurden dabei ExpertInnen (z.B. ÄrztInnen, Pflegepersonal, Mitarbeiter des Roten Kreuzes o.Ä.) zu Rate gezogen?

Ob im Zuge der Übersetzungen seitens der beauftragten Agenturen Dritte hinzugezogen wurden, ist dem ÖIF nicht bekannt.

Einige Informationen auf der genannten Website stammen ursprünglich von verschiedensten Bundesministerien oder anderen Websites. **Wurde aber die Übersetzung dieser Informationen vom Integrationsfonds durchgeführt oder stammen schon die übersetzten Informationen von Ministerien o.Ä.?**

Die meisten Übersetzungen wurden vom ÖIF in Auftrag gegeben, vereinzelt wurden bereits übersetzte Dokumente von der Bundesregierung zur Verfügung gestellt.

Gehe ich Recht in der Annahme, dass es der hauptsächliche Zweck dieser Website bzw. der darin enthaltenen Informationen ist, die LeserInnen über Corona und die damit verbundenen Themen wie Impfung und dergleichen zu informieren?

Ja, die Website soll Personen mit geringeren Deutschkenntnissen über die aktuell geltenden Sicherheitsmaßnahmen informieren, aufklären und Flüchtlinge sowie Migrantinnen und Migranten ebenso bei der Inanspruchnahme von Testungen und Impfungen unterstützen.

Dies sind meine hauptsächlichen Fragen, ich freue mich allerdings auch über alle Informationen, die Sie mir zusätzlich dazu zu dieser Website und ihrer Übersetzung geben können.

Ich bedanke mich im Voraus für Ihre Mühe und verbleibe mit freundlichen Grüßen,
Sonja Hengsberger

E-mail correspondence with the Vienna Tourism Board, corresponding to the bibliographic entry under Wien Tourismus (2022) (under secondary sources)

Sehr geehrte Frau Hengsberger,
herzlichen Dank für Ihre Anfrage. Ich habe mir erlaubt, Ihre Fragen kurz und bündig zu beantworten (siehe unten).
Beste Grüße, [REDACTED]

[REDACTED] [representative of WienTourismus] | Teamleiter Content Creation | Medienhaus
WienTourismus | Invalidenstraße 6 | 1030 Wien, Österreich | www.wien.info | #ViennaWaitsForYou

-----Ursprüngliche Nachricht-----

Von: Sonja Hengsberger <a01601237@unet.univie.ac.at>

Gesendet: Donnerstag, 3. März 2022 11:14

An: Wien-Hotels & Info <info@wien.info>

Betreff: Informationen für Masterarbeit

Sehr geehrte Damen und Herren!

Ich schreibe derzeit meine Masterarbeit an der Universität Wien (Studium: Übersetzen und Dolmetschen) und wäre sehr dankbar für Ihre Hilfe. Im Zuge meiner Arbeit beschäftige ich mich mit der Evaluierung übersetzter Texte und suche daher Textbeispiele für meine Arbeit. Der Inhalt dieser Webseite (<https://www.wien.info/de/reiseinfos/covid-19-informationen/testmoeglichkeiten-wien-gaeste-387066>) wäre für mich interessant und ich hätte einige Fragen zu der Englischen Übersetzung (<https://www.wien.info/en/travel-info/corona-information/testing-options-vienna-guests-388220>). Es wäre für mich sehr hilfreich, wenn Sie mich an eine Stelle verweisen könnten, wo man die unterstehenden Fragen kurz beantworten kann.

1. Gehe ich recht in der Annahme, dass die deutsche Version der Webseite das Original darstellt (und die englische Version die Übersetzung)?

Ja, die deutsche Version eines Artikels ist immer Ausgangsversion eines Artikels in Fremdsprachen.

2. Von wem wurde die Übersetzung durchgeführt? (z.B. von professionellen ÜbersetzerInnen, von einer internen Übersetzungsabteilung, durch automatische (maschinelle) Übersetzung etc.)

Wir arbeiten seit vielen Jahren mit externen professionellen Übersetzer:innen zusammen. Zum Teil sind das auch Natives. Derzeit wird bei uns nichts automatisch/manuell übersetzt.

3. Ich nehme an, dass der Inhalt der Webseite darauf abzielt, Einreisenden aus anderen Ländern (und daher mit keinen oder geringen Deutschkenntnissen) Informationen über die Einreise während Covid-19 zu bieten. Ist dieser Annahme korrekt?

Unsere Website gibt es in insgesamt 12 Sprachen. In unterschiedlichem Umfang produzieren wir Content für diese Sprachen. Bei den Covid-Seiten haben wir uns dazu entschlossen, diese Informationen in fünf Sprachen anzubieten: Deutsch, Englisch, Italienisch, Französisch, Spanisch. Uns ist wichtig, unsere Gäste so gut wie möglich in ihrer eigenen Sprache zu informieren.

Vielen Dank im Voraus und mit freundlichen Grüßen, Sonja Hengsberger

E-mail correspondence with the Federal Ministry for Social Affairs, Health, Care and Consumer Protection, corresponding to the bibliographic entry found under BMSGPK (2022) (under secondary sources)

AW: Informationen für Masterarbeit

Absender: [REDACTED] [representative of the BMSGPK]

Datum: Heute (18.03.2022) 09:43

Sehr geehrte Frau Hengsberger,

1. Die deutsche Version ist das Original.
2. Die Übersetzung wurde von professionellen Übersetzer:innen durchgeführt.
3. Ihre Annahmen sind korrekt.

Wir wünschen Ihnen viel Erfolg mit Ihrer Arbeit.

Beste Grüße,

Bundesministerium für Soziales, Gesundheit, Pflege und Konsumentenschutz Abteilung
Kommunikation und Service

Stubenring 1, 1010 Wien | kommunikation@sozialministerium.at | sozialministerium.at

-----Ursprüngliche Nachricht-----

Von: Sonja Hengsberger a01601237@unet.univie.ac.at

An: * Briefkasten <briefkasten@sozialministerium.at>; * kommunikation

kommunikation@sozialministerium.at

Betreff: Informationen für Masterarbeit

Sehr geehrte Damen und Herren!

Ich schreibe derzeit meine Masterarbeit an der Universität Wien (Studium: Übersetzen und Dolmetschen) und wäre sehr dankbar für Ihre Hilfe. Im Zuge meiner Arbeit beschäftige ich mich mit der Evaluierung übersetzter Texte und hätte daher einige kurze Fragen zu dem Infoblatt "Coronavirus: So trage und verwende ich meine FFP2-Maske richtig" und der Englischen Übersetzung "Coronavirus: How to wear and use your FFP2 mask" (siehe Anhang). Es wäre für mich sehr hilfreich, wenn Sie mich an eine Stelle verweisen könnten, wo man die unterstehenden Fragen kurz beantworten kann.

1. Gehe ich recht in der Annahme, dass die deutsche Version des Merkblatts das Original darstellt (und die englische Version die Übersetzung)?
2. Von wem wurde die Übersetzung durchgeführt? (z.B. von professionellen ÜbersetzerInnen, von einer internen Übersetzungsabteilung, durch automatische (maschinelle) Übersetzung etc.)
3. Ich nehme an, dass das (englische) Merkblatt sich an alle BürgerInnen und/oder Reisende richtet, die Deutsch nicht oder nicht ausreichend beherrschen, um sich auf Deutsch zu informieren. Weiters nehme ich an, dass damit zu einer Aufklärung über Corona-Maßnahmen (Maskentragen) unter der Bevölkerung beigetragen werden soll. Sind diese Annahmen korrekt?

Vielen Dank im Voraus und mit freundlichen Grüßen, Sonja Hengsberger

E-Mail correspondence with the author of the original study, corresponding to the bibliographic entry under Colina (2022) (under secondary sources)

Betreff: Re: [EXT]Question regarding 2008 article for my MA-thesis

Von: Colina, Sonia – [REDACTED]

An: Sonja Hengsberger a01601237@unet.univie.ac.at

Dear Ms. Hengsberger:

Thank you for your email and for your continued interest in my work. I can confirm that Text 1 is the longest (1200 words) and that the other texts had 270 (Hep. B, text 2) and 367 words (Text 3 Medicare). I was unable to do a word count on Text 2, but the word counts provided in the article were rounded up to the nearest tenth.

I hope this helps.

Best regards, Sonia Colina, PhD

From: Sonja Hengsberger a01601237@unet.univie.ac.at

Date: Wednesday, August 10, 2022 at 1:56 AM

To: Colina, Sonia – [REDACTED]

Subject: Re: [EXT]Question regarding 2008 article for my MA-thesis

Dear Ms Colina,

you might remember that I contacted you last year regarding my Master's thesis. I am sorry to bother you again, but I am now analyzing the results of my little study and trying to compare them to the results you obtained in your study ("Translation Quality Evaluation - Empirical Evidence for a Functionalist Approach" (The Translator 14: 1; 2008)). Specifically, I am trying to compare rating times, but since my texts do not have the same length as the ones you used, I am calculating the average rating time per 10 words in order for the results to be comparable. For that to happen, however, I would need to know the word count of the source texts you used. You mention in the article that one source text had approx. 1200 words and the others between 270 and 370 words each, but it would be extremely helpful to know, which text exactly was the longest. Based on your timing-results, I am assuming that it was source Text 1 that had 1200 words, but it would be great if you could confirm this and also perhaps the length of the other two source texts.

I hope my request is not too much of a bother for you and I thank you in advance for your time.

Cordially,

Sonja Hengsberger

Appendix V: Checklist

Checklist

1. Preparations:

If you want to do the evaluation by hand, print out all three documents labelled “Tool – Text 1/2/3” as well as the “Questionnaire” now. You can also print out the texts. Then take to hand the documents “Tool – Text 1”, “Text 1 – source text” and “Text 1 – target text”.

If you want to do the evaluation on your computer, open the documents “Tool – Text 1” as well as the documents “Text 1 – source text” and “Text 1 – target text”.

2. Start Time recording:

Look at the clock and take note of the time now. Write it down on the top right corner of the “Tool - Text 1” under “Start time”.

3. Evaluation:

First, read the **translation brief** (text type, target audience, purpose) in “Tool – Text 1”. You can also click on the link and have a look at where the text came from, if you want to.

Now read “Text 1 – target text”, preferably without looking at the source text. Once you have read the target text, do the evaluation for the category **Target Language** and for the category **Functional and Textual Adequacy**.

Remember, in each category choose the description that best fits the text given. You can make comments if you want to. If you do not have enough space in the column, you can write your comment anywhere else on the page or in a separate document, but please make sure to say which category and text the comment belongs to.

4. Evaluation:

Once you have completed step 3, read “Text 1 – source text”. You can also re-read the target text or parts of it at any time. Now do the evaluation for the category **Non-Specialized Content (Meaning)** and for the category **Specialized Content and Terminology**.

Once again, in each category choose the description that best fits the text given. Any comments you make will be welcome.

5. Scoring:

You should now have chosen for each category one description that best fits the target text (i.e. chosen one box per category). Take note of the category number associated with the descriptions you chose. Now go to the last page of “Tool – Text 1” where it says “Scoring Worksheet”. Here you can see the scores associated with each category number. Write down the score associated with the category number you chose in each category in the box “Score” in the correct line of each category. Once you have done this, add up all four partial scores into one total score and write it down at the bottom of the page under “Total Score”.

To do this, have a look at what box you chose for the category Target Language (in step 3). Did you choose description 1a, 1b, 1c, 1d or 1e? If, for example, you chose 1d for Target Language, that means Text 1 gets a Target Language-score of 20. Repeat this for all the categories (Functional and Textual Adequacy, Meaning, Specialized Content and Terminology). Now you have four scores (e.g. 20 for 1d, 15 for 2c, 15 for 3b and 10 for 4b). Add up these scores (20+15+15+10=60) and write the result down under “Total Score”.

- 6. Finish Time recording:** Now look at the clock again and note down your “Finish time” on the top right corner of the “Tool - Text 1”.
-

Congratulations, you have finished the evaluation of Text 1!

Now repeat steps 2-6 for the other two texts and then have a look at steps 19 and 20.

If you feel unsure, you can simply go on following the checklist to the end (steps 7-20).

7. Preparations:

If you want to do the evaluation by hand, take out your print of “Tool – Text 2”, and “Text 2 – source text” and “Text 2 – target text”.

If you want to do the evaluation on your computer, open the documents “Tool – Text 2” as well as the documents “Text 2 – source text” and “Text 2 – target text”.

- 8. Start Time recording:** Look at the clock and take note of the time now. Write it down on the top right corner of “Tool - Text 2” under “Start time”.

- 9. Evaluation:** First, read the **translation brief** (text type, target audience, purpose) in “Tool – Text 2”. You can also click on the link. Now read “Text 2 – target text”, preferably without looking at the source text. Once you have read the target text, do the evaluation for the category **Target Language** and for the category **Functional and Textual Adequacy**.

- 10. Evaluation:** Once you have completed step 9, read “Text 2 – source text”. You can also re-read the target text or parts of it at any time. Now do the evaluation for the category **Non-Specialized Content (Meaning)** and for the category **Specialized Content and Terminology**.

- 11. Scoring:** You should now have chosen one description for category that best fits the target text (i.e. chosen one box per category). Now go to the last page of “Tool – Text 2” where it says “Scoring Worksheet”. Here you can see the scores associated with each category number. Write down the score associated with the category number you chose in each category in the box “Score” in the correct line. Once you have done this, add up all four partial scores into one total score and write it down at the bottom of the page under “Total Score”.

- 12. Finish Time recording:** Now look at the clock again and note down your “Finish time” on the top right corner of “Tool - Text 2”.
-

Congratulations, you have finished the evaluation of Text 2!

13. Preparations:

If you want to do the evaluation by hand, take out your print of “Tool – Text 3”, and “Text 3 – source text” and “Text 3 – target text”.

If you want to do the evaluation on your computer, open the documents “Tool – Text 3” as well as the documents “Text 3 – source text” and “Text 3 – target text”.

- 14. Start Time recording:** Look at the clock and take note of the time now. Write it down on the top right corner of “Tool - Text 3” under “Start time”.

- 15. Evaluation:** First, read the **translation brief** (text type, target audience, purpose) in “Tool – Text 3”. You can also click on the link. Now read “Text 3 – target text”, preferably without looking at the source text. Once you have read the target text, do the evaluation for the category **Target Language** and for the category **Functional and Textual Adequacy**.
- 16. Evaluation:**
Once you have completed step 15, read “Text 3 – source text”. You can also re-read the target text or parts of it at any time. Now do the evaluation for the category **Non-Specialized Content (Meaning)** and for the category **Specialized Content and Terminology**.
- 17. Scoring:** You should now have chosen one description out of each category that best fits the target text (i.e. chosen one box per category). Now go to the last page of “Tool – Text 3” where it says “Scoring Worksheet”. Here you can see the scores associated with each category number. Write down the score associated with the category number you chose in each category in the box “Score” in the correct line. Once you have done this, add up all four partial scores into one total score and write it down at the bottom of the page under “Total Score”.
- 18. Finish Time recording:** Now look at the clock again and note down your “Finish Time” on the top right corner of “Tool - Text 3”.

Congratulations, you have finished the evaluation of Text 3!

Thank you for completing the evaluation! Now please take 20 more minutes to fill in the questionnaire and then send me all your filled-in materials.

- 19. Questionnaire:** Open/take to hand the document “Questionnaire” and answer all the questions. This should not take you more than twenty minutes.

20. Last steps:

Please check that you have given four partials scores and calculated once total score for each text, and that you have noted down your Start Time and Finish Time for each text.

If you did the evaluation by hand, please take a photo or scan of the filled-in documents listed below, and then send them to me at a01601237@unet.univie.ac.at.

If you did the evaluation on your computer, please send me the filled-in documents listed below to a01601237@unet.univie.ac.at.

- “Tool – Text 1”
- “Tool – Text 2”
- “Tool – Text 3”
- “Questionnaire”
- any separate document(s) you may have written your comments in

You have made it through the checklist and have officially finished the study!
Thank you so much for helping me with my MA project!

Appendix VI: Questionnaire

After-evaluation questionnaire

Dear participant, thank you for completing the evaluation. Now that you have used the evaluation tool, I am very interested in **finding out about your experience**. This is important for my thesis, because I will compare your experience and any problems you might have had (or not) to the experience of the raters in the original study that I am replicating.

Please note that there are no right or wrong answers to the questions below. Answering the questionnaire should not take you more than twenty minutes in total. Any answers you give will only be used for my MA project and only in anonymized form (see consent form).

Once again, please fill in the questionnaire directly in Word by clicking on the check boxes, using the drop-down lists or typing in text where this is required. If this does not work for some reason, you can also print out this questionnaire and answer the questions by hand (then you will need to send me a scan or photo of your finished questionnaire).

*Questions marked with * allow you to give one answer only, to questions marked with ** you can give more than one answer (e.g. tick multiple check boxes).*

1. After our training session (video conference and Q&A session), did you feel prepared to use the evaluation tool to evaluate translated texts? *

- ☐ Yes, I felt entirely prepared.
- ☐ Yes, I felt somewhat prepared.
- ☐ No, I did not feel prepared, because

2. Did you experience any problems (of any kind, e.g. technical, theoretical, practical) when using the tool? (e.g. did you have trouble matching the descriptive statements to your evaluation of the text, or did you have trouble distinguishing between the different descriptors of one criterion?) *

- ☐ Yes, I found that I had problem(s).....
- ☐ No, I did not have any problems.

3. If you answered the previous question with “Yes”, do you feel that the problems you experienced prevented you from effectively using the tool? (If you answered the previous question with “No”, you can leave this question as well as question 4, 5 and 6 blank.) *

- ☐ Yes, I had major problems that prevented me from effectively using the tool.
- ☐ No, the problems I had were minor in nature and did not prevent my using the tool.
- ☐ I did have some problems that had some influence on my use of the tool, but they were neither major nor minor, but somewhere in between.

4. If you experienced problems in using the tool, were these problems related to any particular criterion (Target Language, Functional and Textual Adequacy, Non-Specialized Content (Meaning), Specialized Content and Terminology)? **

☐ Yes, I found that I had trouble with one criterion in particular, namely, because.....

☐ No, my problems were unrelated to any particular criterion.

☐ Some of my problem(s) were related to one particular criterion, while others were unrelated to any criterion in particular. *(If this is true, please tick the first answer as well and describe which problem was related to one particular criterion.)*

5. If you experienced any problems, what do you think was the reason? **

I feel that my problem(s) was/were due to.....

☐ insufficient preparation.

☐ lack of experience with translation quality assessment.

☐ lack of professional experience as a translator.

☐ insufficient specialized knowledge of translation theory.

☐ the tool itself or some part of the tool.

☐ the texts to be evaluated.

☐ some other reason, namely.....

6. Please explain your answer(s) to the previous question. What do you think would have been necessary or would have helped you to avoid the problem(s) you had? **

.....

7. Do you have any other remarks about your experience in using the tool that have not been covered by the questions above? *

☐ Yes, I would like to add that

☐ No.

8. How confident do you feel of the score you gave for each of the texts? **

	Very confident	Somewhat confident	Not confident
Text 1 (Informationen zur Corona-Schutzimpfung)	☐	☐	☐
Text 2 (Testmöglichkeiten und To-Dos bei einer Infektion)	☐	☐	☐
Text 3 (So trage und verwende ich meine FFP2 Maske richtig)	☐	☐	☐

9. Do you think that, in professional translation practice, it is generally necessary to assess the quality of translated texts (either your own translations or translations done by other translators)? *

- ☐ Yes, I think that translation quality assessment is generally necessary in practice.
- ☐ No, I do not think that translation quality assessment is a necessary part of translation practice.
- ☐ I am not sure.

10. Do you think that, if the quality of translations is being assessed in professional practice, that this should be done through a structured evaluation tool such as the one you used (as opposed to subjective judgements that are not based on structured criteria or a defined evaluation tool)? *

- ☐ Yes, I think that a structured tool is best suited for translation quality evaluation.
- ☐ No, I do not think that a structured tool is necessary to assess translation quality, since this can be done just as well by other means, such as.....
- ☐ I am not sure.

11. Would you use the evaluation tool by Colina (2008), which you used in the study, to assess the quality of translated texts in practice? *

- ☐ Yes, I would use Colina's evaluation tool in practice, because.....
- ☐ I might consider using Colina's evaluation tool, but I am not sure, because.....
- ☐ No, I would not use Colina's evaluation tool in practice, because.....

12. In your opinion, do you see the functionalist and descriptive nature of Colina's (2008) tool as positive or negative? (*functionalist means that the evaluation takes place relative to the function, the audience and the context of the evaluated texts, and descriptive means that you do not rate on a purely numerical scale (10, 20, 30...), but that each quality level is represented by a descriptive statement for each criterion and only later converted into scores*) **

- ☐ I see the functionalist nature of the tool as something positive, because.....
- ☐ I see the functionalist nature of the tool as something negative, because.....
- ☐ I do not have a strong opinion either way regarding the functionalist nature of the tool.
- ☐ I see the descriptive nature of the tool as something positive, because.....
- ☐ I see the descriptive nature of the tool as something negative, because.....
- ☐ I do not have a strong opinion either way regarding the descriptive nature of the tool.

13. Do you wish to tell me anything else about your participation in the study? *

☐ Yes, I would like to say that.....

☐ No.

😊 You have reached the end of the questionnaire. Thank you for completing it and helping me with my thesis! 😊

Appendix VII: Comparison of original study and replication study

		Original study by Colina (2008)	Replication study
corner-stones	evaluation method	TQA tool by Colina (2008)	TQA tool by Colina (2008)
	evaluation object	translated health-education texts (3 STs + translations into the TLs)	translated health-education texts (3 STs + translations into TL)
	evaluation goal	to assess quality of texts / to test capabilities and usability of tool	to assess quality of texts / to test usability of tool for students and novice translators
languages		SL: English TLs: Russian, Chinese, Spanish	SL: German TL: English
brief	used as reference	yes	yes
	extrapolated through	expert opinion of professional translator and features/context of selected texts	information obtained from agencies that published the texts and features/context of texts
criteria	number of criteria	4	4
	evaluation criteria and weighting	• Target Language (25) • Functional and Textual Adequacy (20) • Non-Specialized Content (25) • Specialized Content and Terminology (15)	• Target Language (25) • Functional and Textual Adequacy (20) • Non-Specialized Content (25) • Specialized Content and Terminology (15)
	number of descriptors	varies for each criterion (as opposed to Colina (2009), where all criteria have four descriptors each)	varies for each criterion (as opposed to Colina (2009), where all criteria have four descriptors each)
texts	nature of texts	• health education material • publically available (through government agencies) • reader-oriented; little specialized terminology; aimed at a lay audience	• health education material • publically available (through government/public agencies) • reader-oriented; little specialized terminology; aimed at a lay audience
	number of texts	3 STs + 3 Russian TTs + 3 Chinese TTs + 3 Spanish TTs = 12	3 STs + 3 English TTs = 6
	text length (ST)	total: 1,837 words • Text 1: 1,200 words • Text 2: 270 words • Text 3: 367 words	total: 1,316 words • Text 1: 428 words • Text 2: 608 words • Text 3: 280 words
method /steps	pre-pilot session	yes	no
	training session	yes • internet/phone conference (~90 min) • question-and-answer session • PowerPoint presentation • PDF materials	yes same material as in original study used: no (see footnote 56) • video conference (~40 min) • question-and-answer session • PowerPoint presentation • try-out example
	rating session	yes • materials sent via e-mail • materials in the form of PDF documents • training materials available	yes • materials sent via e-mail • materials in the form of Microsoft Word documents (except one text as a PDF)

		• possibility to contact author: unknown	document and the PowerPoint presentation) • possibility to contact author: yes (not used)
	questionnaire	yes • 19 questions in total ○ 8 open-ended questions ○ 11 MC questions • aspects asked for in the questionnaire: ○ preparedness to use the tool ○ training materials (usefulness, possible improvements) ○ confidence in scores given ○ ability to use the tool and possible problems in doing so ○ feelings about the tool and study	yes • 13 questions in total ○ 1 open-ended question ○ 4 MC questions ○ 8 combined questions • aspects asked for in the questionnaire: ○ preparedness to use the tool ○ experience using the tool ○ problems using the tool ○ confidence in scores given ○ attitude towards TQA (and Colina's tool) in practice
proce- dure	training materials available during evaluation	• PowerPoint presentation • other PDF training materials (not specified)	• PowerPoint presentation • try-out example • checklist
	time available to evaluators	one afternoon and one night (approximately 20 hours)	one afternoon and one night (approximately 22 hours)
	time of rating each text recorded	yes	yes
	advice from consultants	yes	no
rat- ers/ evalua- tors	number of raters	37(+3 consultants)=40	4
	raters paid	yes	no
	rater profile	• all: native or near-native language proficiency (method of ascertaining this: unknown) • most: university/college degree (not necessarily language-related) • raters divided into groups: yes ○ professional translators: rely on translation as income, min. 5 years experience, relevant education, professional membership ○ language teachers: teach language courses, may have translation experience, but do not rely on it as source of income ○ bilinguals: little or no formal proficiency, no translation education	• all: native or near-native language proficiency (method of ascertaining this: self-assessment of participants) • all: language-related university degree (BA and/or MA) • raters divided into groups: no • all: students or novice translators (no or less than one year professional experience) • all: no membership in a professional organization
limita- tions		• recruitment: uneven numbers of raters across groups and languages; no bilingual group for Chinese and Russian • study size: 37 raters, 3 texts + translations into 3 languages	• recruitment: raters unpaid • study size: only 4 raters, 3 texts + translations into 1 language, only 2 languages in total • imperfect replication

Appendix VIII: Abstracts

English abstract

This MA project features the replication of a study conducted originally by Colina (2008). The replication study concerned the quality evaluation of translations from German to English by means of a model created by Colina (2008; 2009). The evaluation in the replication study was carried out by students and novice translators from the University of Vienna's Centre for Translation Studies, who had little or no professional experience, as opposed to professionally experienced translators, language teachers and bilingual persons in the original study. The replication study sought to provide data on students' and novice translators' ability to use the above-mentioned model, on differences in productivity compared to the evaluators in the original study as well as on participants' attitude towards the model.

Results show that students and novice translators were able to use the model, although not without some of them encountering problems that the more experienced participants in the original study did not have. The average productivity of replication study participants, measured as time needed for the evaluation, was shown to be equal to or to even exceed the average productivity of participants in the original study, with few exceptions. The replication study also revealed a largely positive attitude among participants towards Colina's model, notably regarding the use of the model in practice and its functionalist and descriptive nature. Last but not least, the replication study also showed that students and novice translators, despite their lack of professional experience, may offer interesting opinions and suggestions in the field of translation quality evaluation.

German abstract

Die vorliegende Masterarbeit umfasst die Wiederholung einer Studie, welche ursprünglich von Colina (2008) durchgeführt wurde. Die Wiederholungsstudie beschäftigt sich mit der Qualitätsevaluierung von Übersetzungen aus dem Deutschen ins Englische mittels des von Colina (2008; 2009) geschaffenen Evaluierungsmodells. Bei der Wiederholungsstudie wurde die besagte Evaluierung von BerufsanfängerInnen und StudentInnen des Zentrums für Translationswissenschaft der Universität Wien durchgeführt, wohingegen bei der Originalstudie TranslatorInnen mit Berufserfahrung, SprachenlehrerInnen und zweisprachige Personen die EvaluatorInnen waren. Die vorliegende Masterarbeit beschäftigte sich im Speziellen mit der Fähigkeit von StudentInnen und BerufsanfängerInnen, Colinas (2008) Modell zu benutzen, mit deren Produktivität (gemessen als die für die Evaluierung benötigte

Zeit) im Vergleich zur Produktivität der EvaluatorInnen der Originalstudie sowie mit der Einstellung der TeilnehmerInnen zu Colinas Modell.

Die Resultate zeigen, dass StudentInnen und BerufsanfängerInnen in der Lage waren, Colinas Modell zu benutzen, obwohl einige TeilnehmerInnen dabei auf Probleme stießen, welche die erfahreneren TeilnehmerInnen der Originalstudie nicht hatten. Es stellte sich weiters heraus, dass die durchschnittliche Produktivität der Wiederholungsstudien-TeilnehmerInnen bis auf wenige Ausnahmen der durchschnittlichen Produktivität der TeilnehmerInnen der Originalstudie entsprach oder diese sogar übertraf. Darüber hinaus konnte eine überwiegend positive Einstellung der TeilnehmerInnen gegenüber Colinas Modell aufgezeigt werden, unter anderem betreffend die praktische Anwendung des Modells sowie dessen funktionalistische und deskriptive Eigenschaften. Letztlich konnte die Wiederholungsstudie ebenfalls aufzeigen, dass StudentInnen und BerufsanfängerInnen, trotz ihrer mangelnden Berufserfahrung, im Bereich der Qualitätsevaluierung von Übersetzungen interessante Meinungen und Vorschläge beitragen können.