Why go the extra mile? How different degrees of post-editing affect perceptions of texts, senders and products among end users

Gys-Walt van Egdom, Utrecht University
Mark Pluymaekers, Zuyd University of Applied Sciences

ABSTRACT

In recent decades, post-editing has received its fair share of attention in the industry as well as in academic circles. What has attracted by far the most attention is the question of quality: together, machine translation and post-editing defy long-standing and commonplace notions of quality. In this paper, we try to observe quality from the vantage point of end users, who are believed to have the final say on a text's fitness for purpose. We will report on an experiment in which end users were asked to pass judgment on manipulated machine translations with different degrees of post-editing. Our findings demonstrate that the additional effort associated with higher degrees of post-editing does not necessarily lead to more positive judgments about text quality. The evidence suggests that text quality is context-dependent and is, therefore, subject to a somewhat opaque process of constant (re)negotiation.

KEYWORDS

Post-editing, machine translation, perceived quality, fitness for purpose, sender image.

1. Introduction

Before the turn of the millennium, the translation industry had little reason to look kindly upon machine translation (MT). Despite the confidence that the earliest ventures in MT had inspired, the ideal of a fully automated high-quality translation (FAHQT) appeared short-lived (ALPAC 1966; see Hutchins 1999, Van der Meer 2016). In recent years, the situation has changed for the better for MT, for at least three reasons. First of all, in today's post-literate society, where written language proficiency is granted a less prominent role, younger generations are increasingly accustomed to reading texts of questionable quality, and it has been claimed that they have adjusted their quality expectations accordingly (Massardo and Van der Meer 2017: 22; see also Hedges 2009). Secondly, statistical phrase-based and neural MT have provided a much-needed boost to the quality of MT output (see Koehn 2009; Bentivogli *et al.* 2016; Koehn 2017). And finally, the industry has managed to find new ways to leverage MT output of suboptimal quality, for example through post-editing (PE) (see Krings 2001; Allen 2002; ISO 18587 2017).

In the current paper, we focus on end users' perceptions of MT output that has been post-edited to a greater or lesser degree. In doing so, we take a view on quality that differs from traditional approaches, in which quality tends to be equated with "excellence" or "flawlessness". Instead, we aim to judge post-edited MT output on its fitness for purpose, which we conceptualize as the degree to which a text fulfils the end users'

informational needs and achieves its communication goals (see Segers and Van Egdom 2018: 41).

The idea of looking at quality from different angles is not new; it was advanced as early as 1984, when David Garvin published his seminal text on Total Quality Management (1984). In his paper, Garvin reviews five definitions of quality he encountered in different domains of knowledge:

1. the transcendent approach;
2. the product-based approach;
3. the user-based approach;
4. the manufacturing-based approach;
5. the value-based approach.

Without stretching the imagination too far, one can see how these approaches are represented in the context of translation[1]. By far the oldest and most prevalent approach to translation quality is the transcendent approach. This approach has been adopted – albeit inadvertently – by a slew of stakeholders (translators, translator trainers, clients, end users). In this view, the quality of a target text eludes the grasp of the receiver. In practice, this means that although no clear arguments can be put forward to justify the judgment passed, it is perceived to stand to reason that a text can be qualitatively poor, mediocre, or good.

Since the 1990s, initiatives have been rolled out to do away with this rather intuitive approach to quality (Drugan 2013: 5-80; see also Saldanha and O'Brien 2014). Since then, both practitioners and academics have been witnessing a proliferation of guidelines, standards, metrics and evaluative models that have been introduced to counter transcendent tendencies. Product-based approaches, which are readily associated with analytical evaluation grids, have often been touted as alternatives to transcendent evaluation (O'Brien 2012; Görög 2014; Lommel *et al.* 2014; Lommel 2014; Mariana *et al.* 2015). Another approach that appears to have a certain appeal in the industry as well as academia is the manufacturing-based approach, which can be linked to standards setting the processual requirements for quality services (e.g. ISO 17100 2015; ISO 18587 2017).

The remaining two approaches (user- and value-based) hold tremendous sway over the translation market (see Massardo *et al.* 2016), yet seem to have attracted only scant attention in Translation Studies. In user-based approaches, the quality of a product or service hinges on the satisfaction of end user needs (Morland 2002; Castilho *et al.* 2014). In other words, a translation is up to standard when the end user is satisfied with it. Value-based approaches, on the other hand, zero in on the price-quality ratio. For example, in some lower-end segments of the translation market,

where cost and timeliness are of the essence, the benchmark for intrinsic linguistic quality can be lower.

In our view, user-based and value-based research on translation quality is warranted, particularly in the subdomains of MT and PE. After all, PE is, more than any other translation-related service, about producing a text that is fit for purpose for the client and the end users. This also becomes evident when looking at the terminology employed in PE (e.g. O'Brien 2010; Hu and Cadwell 2016), which makes explicit (although not well-defined) reference to the underlying needs of clients and end users by using opposing terms such as light/fast/rapid/gist post-editing or full/traditional/heavy post-editing.

What is lacking, however, is empirical research about the way such different degrees or levels of post-editing are perceived by end users. Do they always prefer the most elaborate degree of post-editing, which is relatively costly, or can lesser degrees also be fit for their purpose? Thus far, most research on the reception of PE quality is confined to expert judgements, and thereby fails to take heed of the people that have the final say on text quality: clients and text consumers (see Guerberof 2009; Plitt and Masselot 2010; Tatsumi 2010; García 2011; Daems *et al.* 2013; Ortíz-Boix and Matamala 2015; Daems 2016)[2].

Gaining more insight into what constitutes fitness for purpose from the perspective of end users is important, as it can help translation service providers give evidence-based advice to their clients about the degree of PE required in different contexts. When giving such advice, however, attention should also be paid to the wide array of additional communicational goals a text can have, such as instilling trust in its sender and enhancing the perceived attractiveness of the product that is discussed. Therefore, we include not only perceptions of text quality per se in our study, but also sender image, attitude towards the product and purchase intention.

To summarise, this paper seeks to address the gaps in PE quality research discussed earlier by (1) investigating end users' perceptions of different degrees of post-editing and (2) incorporating not just text perceptions, but also sender and product perceptions. The overarching research question can be formulated as follows:

What is the effect of increasing degrees of post-editing on end users' perceptions of texts, senders and products?

The remainder of the paper will provide an outline of the methods applied in this study, followed by a description of the results and some conclusions that can be drawn from those results. Finally, limitations and suggestions for future research are discussed.

2. Methodology

In order to answer the research question, we conducted an experiment in which we presented end users with machine translated texts that had been post-edited to different degrees. We used a between-subjects design, so that participants could not compare different versions of the same text. Two source texts were included in the experiment: an informative text about phishing intended for the general public and an instructive text about a software package written for IT professionals. The participants who assessed the target versions of these texts were representative of the target group for which the texts were intended. More information about the participants and the materials is provided below.

2.1. Participants and procedure

As stated earlier, two separate participant groups were recruited for this study. The participants who assessed the informative text about phishing (N = 77) were potential customers of the sender of that text (a telecommunications company), and were selected by means of convenience sampling. The sample for the instructive text about a software package (N = 81) consisted of IT students from different universities in the Netherlands and graduates who worked in the industry. No information can be provided about age and gender, as it was not requested in the survey. The main reason for not including real customers in our sample was to prevent interference from their prior experiences in our measurement of sender image and purchase intention. All participants were native speakers of Dutch, which was the target language of the post-edited texts.

Approximately 28% of the participants completed a pen-and-paper version of the experiment. The remaining 72% were directed to an online survey environment created in NetQ (NetQ Internet Surveys 2011).

Participants first read a short introduction in which they were informed about the topic and the target group of the text. Subsequently, they were confronted with one randomly selected post-edited version of the target text, which they were instructed to read carefully, and filled out a questionnaire containing measures for all variables of interest (see Instrumentation below).

2.2. Materials

The experimental materials were based on two source texts (one informative and one instructive) that were originally written in English. In both texts, the company name of the sender was replaced by a fictitious name. Both texts were machine translated into Dutch using SDL Trados Studio and the statistical MT engine Language Cloud. Subsequently, the raw MT output was post-edited by fourth-year translation students

working for Zuyd Vertalingen, the in-house translation bureau of Zuyd University of Applied Sciences in Maastricht (cf. INSTB 2017).

The management of Zuyd Vertalingen was provided with detailed instructions for creating the different PE versions (see Table 1 below). Four degrees of PE were distinguished (minimal PE, light PE, moderate PE and full PE), which were loosely based on the levels or degrees of revision set out by Mossop (2014)[3]. As can be seen in Table 1, the instructions corresponding to the different degrees of PE were formulated in such a way that the number of text manipulations was expected to increase gradually between degree 1 and degree 4.

Degree 1: **minimal PE**	Instructions: correct names, maintain anaphoric relation, parse long sentences, avoid ambiguity
Degree 2: **light PE**	Instructions: correct names, maintain anaphoric relation, parse long sentences, avoid ambiguity, *and* correct grammatical and lexical errors (inversions, congruence and juxtapositions)
Degree 3: **moderate PE**	Instructions: correct names, maintain anaphoric relation, parse long sentences, avoid ambiguity, correct grammatical and lexical errors (inversions, congruence and juxtapositions), *and* improve logic, create cohesion and correct terminology
Degree 4: **full PE**	Instructions: correct names, maintain anaphoric relation, parse long sentences, avoid ambiguity, correct grammatical and lexical errors (inversions, congruence and juxtapositions), improve logic, create cohesion and correct terminology, *and* improve style and add idiomatic constructions

Table 1. Operationalization of Degrees of PE.

The management passed these instructions on to the translators, revisers and QA-managers who were assigned to produce the four versions of each text.

To ensure the validity of the resulting experimental materials, two additional checks were performed. First of all, two experienced lecturers in translation and revision – who had no further involvement in the research

project – were asked to verify whether the instructions had been strictly observed. Having been provided with the list of instructions (Table 1), the lecturers were asked to classify the post-edited versions of each MT into one of the four PE categories. All texts were classified with 100% accuracy.

Secondly, a calculation was made of edit distances to glean an (albeit rough) idea of the differences between the raw MT and the four PE versions, and of the effort required to produce the respective versions. By getting a firmer handle on the effort required to produce a version and combining the calculations with the insights yielded by the end user data, the results of this experiment could be turned to the advantage of language service providers, as it should enable them to strike a happy medium between effort and customer satisfaction. An excellent means to calculate edit distance *ex post facto* is the Damerau-Levenshtein metric (Levenshtein 1966), which is widely used to compute string-to-string similarity[4]. The results of the calculations are presented in Tables 2 and 3. The percentages provide an indication of the degree of similarity between the versions that have been set in opposition. For example, the score of 96.06% indicates that the light PE version and the moderate PE version of the informative text are highly similar: this seems to suggest that, in this case, few manipulations were needed to get from light PE to moderate PE. Conversely, the score of 74.79% suggests that manipulations were manifold in the next phase of PE process: it seems that quite a bit of effort was needed on the part of the post-editor, to get from moderate PE to full PE.

	Raw MT	**Minimal PE**	**Light PE**	**Moderate PE**
Minimal PE	94.68%			
Light PE	90.43%	88.02%		
Moderate PE	87.25%	85.01%	96.06%	
Full PE	70.82%	70.44%	73.57%	74.79%

Table 2. Edit distance between the versions of the informative text.

	Raw MT	**Minimal PE**	**Light PE**	**Moderate PE**
Minimal PE	89.31%			
Light PE	80.36%	86.85%		
Moderate PE	74.03%	80.16%	90.94%	
Full PE	68.20%	73.38%	81.92%	88.99%

Table 3. Edit distance between the versions of the instructive text.

2.3. Instrumentation

In the questionnaire that followed the presentation of the text, participants were asked to voice their opinions about the following

subdimensions of text quality: 1) content, 2) language use, 3) text logic and terminology, 4) style, and 5) usability. Inspiration for the measurement of the first four dimensions was drawn from Mossop's (2014) editing framework and from Rothe-Neves' (2002) translation quality assessment questionnaire. The three statements on usability were based on three mainstays in Skopos theory. Vermeer's formula IA/Trl = f(Sk)[5], first formulated in 1983 (54), is undergirded by the claim that the relative quality of a target text is determined by its fitness to its communicative purpose (1). This purpose is served when the target text fulfils its prospective function (2) and when the expectations and needs of the reader are met (3) (Reiß and Vermeer 1984). A translation is thus considered usable when it serves its intended purpose in a functional context with target addressees who have specific expectations and needs with regard to the offer of information. All items were scored on a 7-point Likert scale.

Sender image, attitude towards the product and purchase intention were gauged using semantic differential scaling with 18 adjective pairs. The list of adjective pairs was derived from Janssen and Gerards (2016), Homer (1990) and MacKenzie *et al.* (1986).

The full version of the questionnaire can be found in the Appendix.

2.4. Analysis

The data were analysed using data processing package SPSS. Separate MANOVAs were conducted for the two texts, with PE version as independent variable and the five dimensions of text quality, sender image, attitude towards the product and purchase intention as dependent variables. MANOVA was used because our design included multiple dependent variables which were likely to be correlated. In such a design, MANOVA provides more power and reduces the likelihood of Type I error for individual dependent variables (Field 2013: 625). To assess the level-by-level differences between different degrees of PE, repeated contrasts were used (Field 2013). This allowed us to assess for each degree of PE (except the lowest) whether it had statistically significant added value over the preceding one.

3. Results

3.1. Informative text

Using Pillai's trace, we found a significant effect of PE degree on the dependent variables ($V = 0.95$, $F(24, 204) = 3.91$, $p < 0.001$). Separate univariate ANOVAs revealed that all perceptions were significantly affected by PE degree (*content*: $F(3, 73) = 29.43$, $p < 0.001$; *language use*: $F(3, 73) = 67.30$, $p < 0.001$; *style*: $F(3, 73) = 43.77$, $p < 0.001$; *logic*: $F(3, 73) = 16.58$, $p < 0.001$; *usability*: $F(3, 73) = 29.85$, $p < 0.001$; *sender*

image: $F(3, 73) = 26.78$, $p < 0.001$; *attitude towards the product*: $F(3, 73) = 35.30$, $p < 0.001$; *purchase intention*: $F(3, 73) = 35.20$, $p < 0.001$).

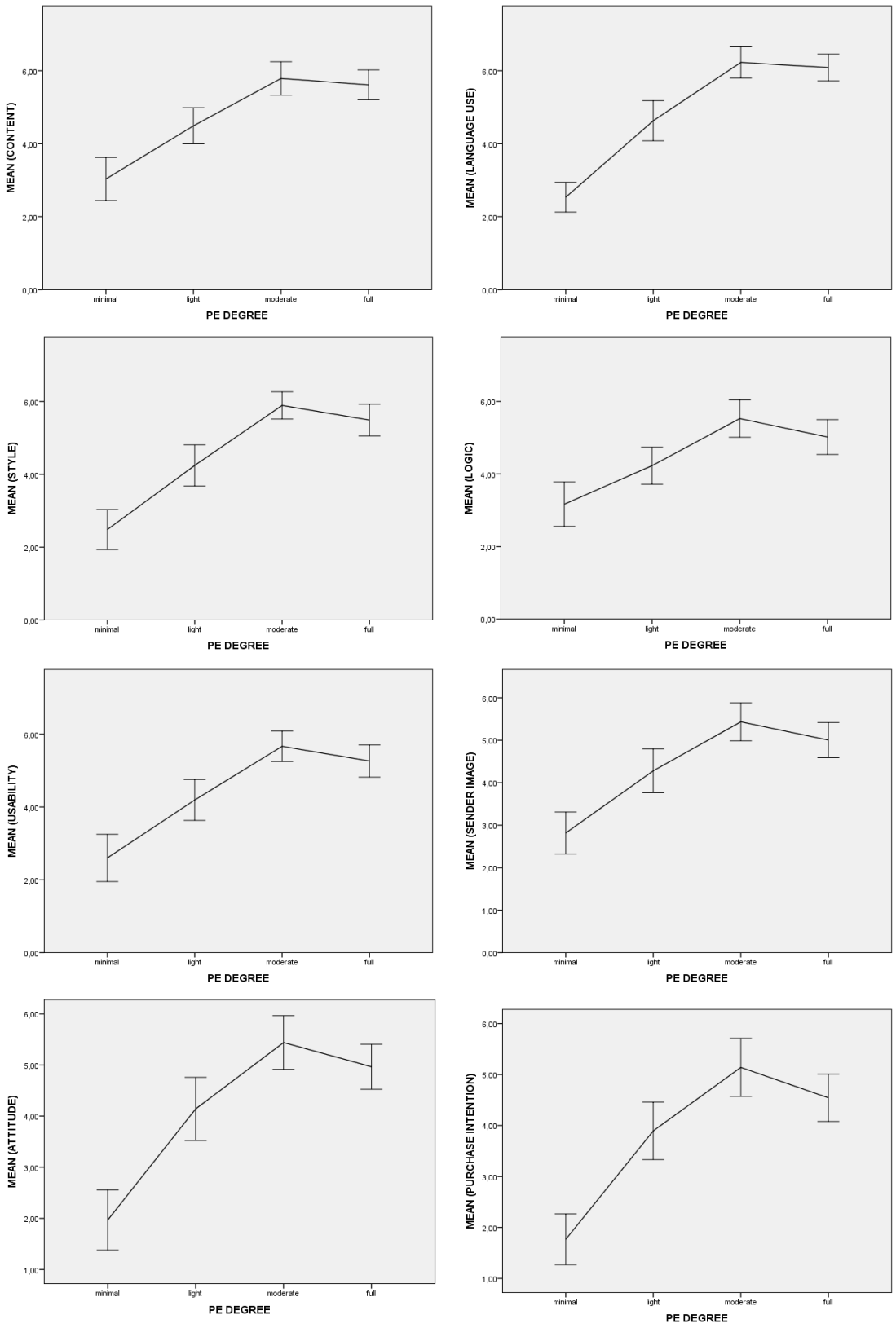

Figure 1. Line plots including error bars for the informative text.

Figure 1 visualizes how the mean for each variable differs as a function of PE degree. A consistent pattern can be observed in these plots: text, sender and product perceptions improve as we move from minimal to light to moderate PE, but then they appear to level off.

The contrasts confirmed that the difference between minimal and light PE was significant for all variables, as was the difference between light and moderate PE. The difference between moderate PE and full PE, however, was never significant. These results are summarized in Table 4 below.

	minimal vs. light	**light vs. moderate**	**moderate vs. full**
Content	Estimate: -1.46 $p < 0.001$	Estimate: -1.30 $p < 0.001$	Estimate: 0.18 $p = 0.60$
Language use	Estimate: -2.10 $p < 0.001$	Estimate: -1.60 $p < 0.001$	Estimate: 0.14 $p = 0.64$
Style	Estimate: -1.76 $p < 0.001$	Estimate: -1.65 $p < 0.001$	Estimate: 0.40 $p = 0.23$
Logic	Estimate: -1.06 $p < 0.005$	Estimate: -1.30 $p < 0.005$	Estimate: 0.51 $p = 0.16$
Usability	Estimate: -1.59 $p < 0.001$	Estimate: -1.47 $p < 0.001$	Estimate: 0.40 $p = 0.27$
Sender image	Estimate: -1.46 $p < 0.001$	Estimate: -1.15 $p < 0.005$	Estimate: 0.43 $p = 0.18$
$A_{product}$	Estimate: -2.17 $p < 0.001$	Estimate: -1.30 $p < 0.005$	Estimate: 0.47 $p = 0.21$
Purchase intention	Estimate: -2.13 $p < 0.001$	Estimate: -1.25 $p < 0.005$	Estimate: 0.60 $p = 0.10$

Table 4: Contrast estimates and *p*-values for the informative text.

3.2. Instructive text

Using Pillai's trace, we again found a significant effect of PE degree on the dependent variables ($V = 0.87$, $F(24, 216) = 3.65$, $p < 0.001$). Separate univariate ANOVAs revealed that all dependent variables were significantly affected by PE degree (*content*: $F(3, 77) = 25.24$, $p < 0.001$; *language use*: $F(3, 77) = 32.99$, $p < 0.001$; *style*: $F(3, 77) = 30.98$, $p < 0.001$; *logic*: $F(3, 77) = 21.04$, $p < 0.001$; *usability*: $F(3, 77) = 15.42$, $p < 0.001$; *sender image*: $F(3, 77) = 22.74$, $p < 0.001$; *attitude towards the product*: $F(3, 77) = 13.77$, $p < 0.001$; *purchase intention*: $F(3, 77) = 5.91$, $p < 0.005$).

Figure 2 below visualizes how the mean for each variable differs as a function of PE degree. Again, the pattern looks similar across different variables, but it differs from the pattern observed for the informative text. First of all, the means keep increasing, also between moderate and full PE. Secondly, the differences between light and moderate PE were not significant for all variables. As can be seen in Table 5, there was no significant difference between light and moderate PE for four of the six dependent variables: content, logic, usability, and purchase intention. On

the other hand, there was a significant difference between moderate and full PE for two variables: language use and style.

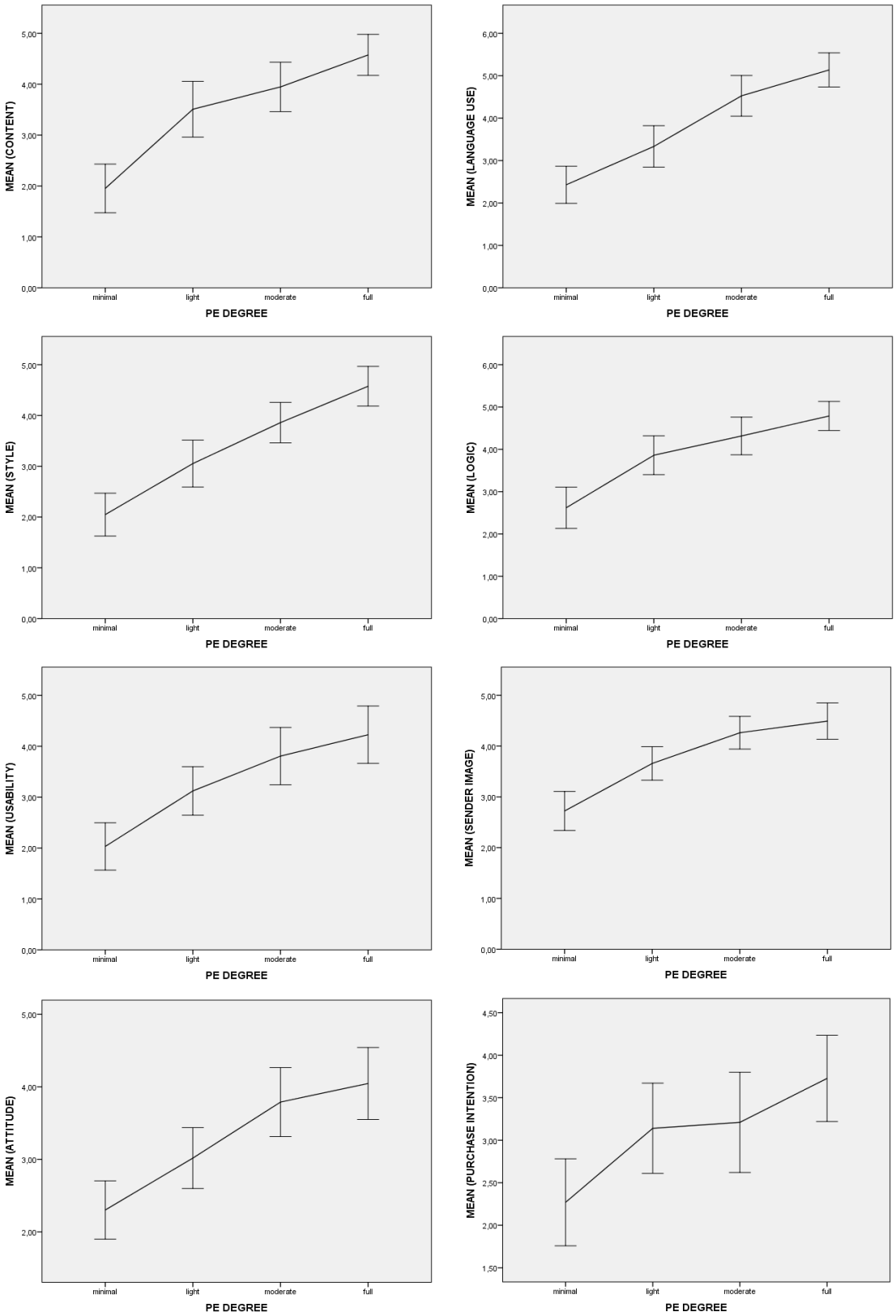

Figure 2. Line plots including error bars for the instructive text.

	minimal vs. light	**light vs. moderate**	**moderate vs. full**
Content	Estimate: -1.56 $p < 0.001$	Estimate: -0.44 $p = 0.19$	Estimate: -0.63 $p = 0.05$
Language use	Estimate: -0.91 $p < 0.005$	Estimate: -1.19 $p < 0.001$	Estimate: -0.61 $p < 0.05$
Style	Estimate: -1.01 $p < 0.005$	Estimate: -0.81 $p < 0.01$	Estimate: -0.72 $p < 0.05$
Logic	Estimate: -1.24 $p < 0.001$	Estimate: -0.46 $p = 0.14$	Estimate: -0.47 $p = 0.11$
Usability	Estimate: -1.09 $p < 0.005$	Estimate: -0.68 $p = 0.06$	Estimate: -0.42 $p = 0.24$
Sender image	Estimate: -0.94 $p < 0.001$	Estimate: -0.61 $p < 0.05$	Estimate: -0.23 $p = 0.34$
$A_{product}$	Estimate: -0.72 $p < 0.05$	Estimate: -0.77 $p < 0.05$	Estimate: -0.26 $p = 0.40$
Purchase intention	Estimate: -0.87 $p < 0.05$	Estimate: -0.10 $p = 0.85$	Estimate: -0.52 $p = 0.15$

Table 5. Contrast estimates and *p*-values for the instructive text.

4. Conclusion and discussion

4.1. Conclusion

In this study, we have sought to investigate the fitness for purpose of post-edited MTs as perceived by end users. Our main goal was to find out whether end users view MT with a high degree of PE more favourably.

The results show convincingly that the degree of PE makes a difference, not just for variables related to perceived text quality, but also for sender and product perceptions. Therefore, it is important that clients of post-editing services consider the consequences of choosing one or another degree of post-editing carefully.

Perhaps surprisingly, full post-editing does not always elicit the most positive judgements from end users. This was very clear in the case of the informative text, where the difference between moderate and full post-editing was never significant, even though the edit distance between the two degrees was relatively large (see Table 2).

For the instructive text, full post-editing elicited the most positive judgements from end users for just two of the eight variables under consideration: language use and style. Again, one might have expected these differences to be more pronounced given the relatively large edit distance between the two degrees (see Table 3).

Overall, the results of our experiment support the idea that there is no single touchstone against which a post-edited MT's fitness for purpose can be measured. This conclusion might have far-reaching consequences for

quality assurance in the translation industry. Instead of meekly submitting to fairly abstract quality standards and client demands, translators and translation service providers are urged to educate their clients, sensitise them to text quality perception and to the way text quality can rub off on sender and product perceptions. By drawing the clients' attention to the complexity of textual fitness for purpose, translators and translation service providers can play a more active part in expectation management.

4.2. Limitations and future research

Our study has obvious limitations and is, thus, likely to spur further user-based research on end users' perceptions of post-edited MT. First of all, it should be noted that the number of texts used in this study is limited. Not only is duplication of this study desirable; similar research with more texts and more text types seems indispensable to corroborate our findings. What could also be interesting, is a follow-up study that explicitly incorporates possible moderating variables such as age and gender, as there are suggestions in literature that younger age groups have lower expectations with regard to text quality (e.g., Hedges 2009). We could not test these moderating effects in the current study because we did not record demographic data and our sampling strategy was not optimised for investigating the effects of continuous moderators such as age.

Another important point to note is that the MT output was produced at a time when computer engineers had not yet capitalised on the potential of neural networks. Given the improvement in the quality of MT output, the manipulation of neural MT output will probably yield different results. Furthermore, the instructions for PE degrees are derived from a basic framework for revision and are not necessarily in keeping with existing PE guidelines. It should also be borne into mind that most of our respondents were selected by means of convenience sampling. Perceived quality is probably best gauged in an authentic communicative situation were the end users' need to acquire information (from a textual source) is more acute.

A final limitation of this study is that it takes no account of text quality as perceived by translation service providers (i.e. translation agencies and freelance translators), clients and other stakeholders. By mapping out their wants and needs, one can more clearly define the notion of fitness for purpose and, by dint of comparison, one can possibly even detect potential sources of service failure. Mapping out text quality perception among translation service providers should perhaps be the first step that will be taken in future research on PE quality.

References

- **Allen, Jeffrey** (2003). "Post-editing. Computers and Translation." Harold Somers (ed.) (2003). *Computers and Translation: A Translator's Guide*. Amsterdam: John Benjamins, 297-317.

- **ALPAC** (1966). *Languages and machines: Computers in translation and linguistics.* Report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Washington [D.C.]: National Academy of Sciences, National Research Council.

- **Arnold, Doug, Balkan, Lorna, Meijer, Siety, Humphreys, R. Lee and Louisa Sadler** (1994). *Machine translation: An introductory guide*. Oxford: NCC Blackwell.

- **Bentivogli, Luisa, Bisazz, Arianna, Cettolo, Mauro and Marcello Federico** (2016). "Neural versus phrase-based machine translation quality: A case study." *arXiv preprint arXiv:1608.04631.*

- **Bowker, Lynne and Melissa Ehgoetz** (2007). "Exploring user acceptance of machine translation output: A recipient evaluation." Dorothy Kenny and Kyongjoo Ryou (eds) (2007). *Across boundaries: international perspectives on translation*. Newcastle-upon-Tyne: Cambridge Scholars Publishing, 209-224.

- **Castilho, Sheila, O'Brien, Sharon, Alves, Fabio and Morgan O'Brien** (2014). "Does post-editing increase usability? A study with Brazilian Portuguese as target language." Marko Tadić, Philipp Koehn, Johann Roturier, Andy Way (eds) (2014). *Proceedings of the 17th Annual Conference of the European Association for Machine Translation, Dubrovnik, Croatia, 16-18 June 2014*, 183-190. http://doras.dcu.ie/19997/1/PE_Usability_EAMT2014_Camera_ready.pdf (consulted 06.06.2018).

- **Daems, Joke, Macken, Lieve and Sonia Vandepitte** (2013). "Quality as the sum of its parts: A two-step approach for the identification of translation problems and translation quality assessment for HT and MT+PE." Sharon O'Brien, Michel Simard, and Lucia Specia (eds) (2013). *Proceedings of the 2nd Workshop on Post-editing Technology and Practice (WPTP-2), Nice, France, September 2.* European Association for Machine Translation, 63-71. http://www.mt-archive.info/10/MTS-2013-W2-TOC.htm (consulted 01.12.2018).

- **Daems, Joke** (2016). *A translation robot for each translator? A comparative study of manual translation and post-editing of machine translations: Process, quality and translator attitude.* PhD Thesis. Ghent University.

- **Drugan, Joanna** (2013). *Quality in Professional Translation: Assessment and Improvement*. London/New York: Bloomsbury.

- **Field, Andy** (2013). *Discovering statistics using IBM SPSS statistics.* London: Sage.

- **García, Ignacio** (2011). "Translating by post-editing: Is it the way forward?" *Machine Translation* 25(3), 217-237.

- **Garvin, David. G.** (1984). "What does product quality really mean?" *Sloan Management Review* 26(1), 25–43.

- **Görög, Atilla** (2014). "Quantification and comparative evaluation of quality. The TAUS Dynamic Quality Framework." *Tradumàtica* 12, 443–454.

- **Guerberof, Ana** (2009). "Productivity and quality in MT post-editing." Marie- Laurie Gerber, Pierre Isabelle, Roland Kuhn, Nick Bemish, Mike Dillinger and Marie-Josée Goulet (eds) (2009). *Beyond Translation Memories Workshop. MT Summit XII. The twelfth Machine Translation Summit. International Association for Machine Translation, Ottawa, August 26-30*. Association for Machine Translation in the Americas. http://www.mt-archive.info/MTS-2009-Guerberof.pdf (consulted 11.12.2017).

- **Hedges, Chris** (2009). *Empire of illusion: The end of literacy and the triumph of spectacle.* New York: Nation Books.

- **Homer, Pamela M.** (1990). "The mediating role of attitude toward the ad: Some additional evidence." *Journal of Marketing Research* 27(1), 78-86.

- **Hu, Ke and Cadwelll, Patrick** (2016). "A Comparative study of post-editing guidelines." *Baltic Journal of Modern Computing* 4(2), 346-353.

- **Hutchins, W. John** (2003). "ALPAC: The (in)famous report." Sergei Nirenburg, Harold L. Somers and Yorick A. Wilks (eds) (2003). *Readings in machine translation.* Cambridge: MIT Press, 131-135.

- **INSTB (Buysschaert, Joost, Fernández Parra, Maria and Gys-Walt van Egdom)** (2017). "Professionalising the curriculum and increasing employability through authentic experiential learning: The cases of INSTB." *Current Trends in Translation Teaching and Learning E (CTTL-E)* 4, 78-111. http://www.cttl.org/uploads/5/2/4/3/5243866/cttl_e_2017_3.pdf (consulted 14.12.2017).

- **ISO 17100** (2015). *Translation services – Requirements for translation services.* Geneva: International Standardization Organization.

- **ISO 18587** (2017). *Translation services – Post-editing of machine translation output – Requirements.* Geneva: International Organization for Standardization.

- **Krings, Hans P.** (2001). *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes* (Geoffrey Koby, ed.). Kent: Kent State University Press.

- **Janssen, Daniel and Valenard Gerards** (2016). "Onze excuses - Over de rol van verontschuldigingen in crisiscommunicatie." *Tijdschrift voor Communicatiewetenschap* 44 (2), 112-133.

- **Koby, Geoffrey, Fields, Paul J., Hague, Daryl, Lommel, Arle, and Alan K. Melby** (2014). "Defining translation quality." *Tradumàtica* 12, 413–420.

- **Koehn, Philipp** (2009). *Statistical machine translation.* Cambridge: Cambridge University.

- **—** (2017). "Introduction to neural machine translation." [Webinar]. *Omniscien Technologies Series*, January 24. https://vimeo.com/201401054 (consulted 22.12.2017).

- **Levenshtein, Vladimir I.** (1966). "Binary codes capable of correcting deletions, insertions, and reversals." *Soviet Physics - Doklady*, 10(8), 707-710. https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf (consulted 05.12.2018).

- **Lommel, Arle, Burchardt, Aljoscha. and Uszkoreit, Hans** (2014). "Multidimensional quality metrics MQM: A framework for declaring and describing translation quality." *Tradumàtica* 12, 455–463.

- **Lommel, Arle** (ed.) (2014). "Multidimensional Quality Metrics (MQM) Definition." http://www.qt21.eu/mqm-definition/definition-2014-08-19.html (consulted 20.12.2017).

- **MacKenzie, Scott B., Lutz, Richard J. and George E. Belch** (1986). "The role of attitude toward the ad as a mediator of advertising effectiveness: A test of competing explanations." *Journal of Marketing Research* 23(2), 130-143.

- **Mariana, Valerie, Cox, Troy and Alan Melby** (2015). "The multidimensional quality metrics (MQM) framework: A new framework for translation quality assessment." *The Journal of Specialised Translation*, 23, 137–161.

- **Massardo, Isabella, Van der Meer, Jaap and Khalilovm Maxim** (2016). *TAUS Translation Technology Landscape Report.* TAUS. https://www.taus.net/think-tank/reports/translate-reports/taus-translation-technology-landscape-report-2016#content (consulted 01.12.2018).

- **Massardo, Isabella and Jaap van der Meer** (2017). *The translation industry in 2022. A report from the TAUS Industry Summit.* TAUS. https://www.taus.net/think-tank/reports/event-reports/the-translation-industry-in-2022 (consulted 01.12.2018).

- **Melby, Alan K., Paul J. Fields and Jason Housley** (2014). "Assessment of post-editing via structured translation specifications." Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard, and Lucia Specia (eds) (2014) *Post-editing of machine translation: Processes and applications.* Newcastle: Cambridge Scholars Publishing, 274-298.

- **Morland, D. Verne** (2002). "Nutzlos, bien pratique, or muy util? Business Users Speak out on the Value of Pure Machine Translation." *Proceedings of Translation and the Computer 24, London, 21-22 November 2002*, 1-17. http://www.mt-archive.info/Aslib-2002-Morland.pdf (consulted 05.12.2018).

- **Mossop, Brian** (2014). *Revising and editing for translators.* Oxon-New York: Routledge.

- **NETQ Internet Surveys 6.7.** (2011). *Software for creating and assessment of internet surveys*. Utrecht: NetQuestionnaires Nederland BV. http://www.netq.nl (consulted 30.06.2017).

- **O'Brien, Sharon** (2006a). *Machine-translatability and post-editing effort: An empirical study using Translog and choice network analysis.* PhD thesis. Dublin City University.

- **—** (2006b). "Pauses as indicators of cognitive effort in post-editing machine translation output." *Across Languages and Cultures* 7 (1), 1–21.

- **—** (2010). "Introduction to post-editing: Who, what, how and where to next." Paper presented at *The Ninth Conference of the Association for Machine Translation in the Americas* (Denver, Colorado 31 October – 4 November 2010). http://www.mt-archive.info/10/AMTA-2010-OBrien.pdf (consulted 21.11.2018).

- **—** (2012). "Towards a dynamic quality evaluation model for translation." *The Journal of Specialised Translation* 17(1), 55-77. http://www.jostrans.org/issue17/art_obrien.pdf (consulted 20.12.2017).

- **Ortiz-Boix, Carla and Anna Matamala** (2015). "Assessing the quality of post-edited wildlife documentaries." *Perspectives. Studies in Translation Theory and Practice* 25(4), 571-593.

- **Plitt, Mirko and François Masselot** (2010). "A productivity test of statistical machine translation post-editing in a typical localisation context." *The Prague Bulletin of Mathematical Linguistics* 93, 7-16.

- **Reiß, Katharina and Hans Vermeer** (1984). *Grundlegung einer allgemeinen Translationstheorie.* Tübingen: Niemeyer.

- **Rothe-Neves, Rui** (2002). "Translation quality assessment for research purposes: An empirical approach." *Cadernos de Tradução* 2(10), 113-131.

- **Roturier, Johann** (2006). *An investigation into the impact of controlled English rules on the comprehensibility, usefulness and acceptability of machine-translated technical documentation for French and German users.* PhD thesis. Dublin City University.

- **Saldanha, Gabriela and Sharon O'Brien** (2014). *Research Methodologies in Translation Studies*. London: Routledge.

- **Segers, Winibert and Gys-Walt van Egdom** (2018). *De kwaliteit van vertalingen. Een terminologie van de vertaalevaluatie.* Kapellen: Pelckmans.

- **Tatsumi, Midori** (2010). *Post-editing machine translated text in a commercial setting: Observation and statistical analysis.* PhD thesis. Dublin City University.

- **Van der Meer, Jaap** (2016). *The Future Does Not Need Translators*. TAUS. http://blog.taus.net/the-future-does-not-need-translators (consulted 11.12.2017).

- **Van Egdom, Gys-Walt, Verplaetse, Heidi, Schrijver, Iris, Kockaert, Hendrik, Segers, Winibert, Pauwels, Jasper, Bloemen, Henri and Bert Wylin** (forthcoming) "How to put the translation test to the test? On preselected item evaluation and perturbation." Elsa Huertas Barros, Sonia Vandepitte and Emilia Iglesias Fernández (eds) (forthcoming). *Quality assurance and assessment practices in translation and interpreting.* Hershey: IGI Global.

- **Vermeer, Hans** (1983). *Aufsätze zur Translationstheorie.* Heidelberg: Groos.

- **Vieira, Lucas N.** (2016). *Cognitive effort in post-editing of machine translation: Evidence from eye movements, subjective ratings, and think-aloud protocols.* PhD thesis. Newcastle University.

- **—** (2017). "From process to product: Links between post-editing effort and post-edited quality." Arnt L. Jakobsen and Bartholomé Mesa-Lao (eds) (2017). *Translation in transition: Between cognition, computing and technology.* Amsterdam: John Benjamins, 162-186.

Biographies

Gys-Walt van Egdom currently teaches Translation Studies and Translation at Utrecht University. In his previous capacity of lecturer-researcher at the Research Centre for International Relationship Management (Zuyd University of Applied Sciences in Maastricht, the Netherlands), he has carried out research in the domains of translation technology, translation didactics, translation evaluation and process-oriented translation studies.

E-mail: g.m.w.vanegdom@uu.nl

Mark Pluymaekers is professor of International Relationship Management at Zuyd University of Applied Sciences in Maastricht, the Netherlands. Mark's research interests include professional communication skills and the application of artificial intelligence in business communication. His scientific work has been published in journals such as *Phonetica*, *Journal of the Acoustical Society of America*, and *International Business Review*. In 2011, he also published a textbook on presentation skills for Dutch students in higher education.

E-mail: mark.pluymaekers@zuyd.nl

Appendix

1. Having read the text once, I am able to summarise the content of the text. (*content*)
2. The message that the sender tries to convey is clear to me. (*content)*
3. Based on the context, I am able to fill in blanks in the text. (*content)*
4. The text contains few to no disturbing spelling errors. (*correctness of language*)

5. The text does not contain unusual sentence structures. (*correctness of language*)
6. To me, the sentences are always of an acceptable length. (*correctness of language*)
7. The text is fluent. (*style*)
8. It suffices to read the sentences one time to grasp their content. (*style*)
9. The style of the text resembles the style employed in other informative texts. (*style*)
10. The text is logically structured. (*logic and terminology*)
11. The text is coherent. (*logic and terminology*)
12. It suffices to read the terms once to comprehend them. (*logic and terminology*)
13. The text achieves its goal. (*usability*)
14. Having read the text, I am able to recognise "phishing". (*usability*)
15. The text meets the expectations I would have as a reader, if I were to seek information on "phishing". (*usability*)
16. Which impression did telecom company D* make:
Sensible - Insensible
Honest – Dishonest
Sympathetic – Unsympathetic
Capable – Incapable
Reliable – Unreliable
Knowledgeable – Ignorant
Sincere – Insincere
Sympathetic – Unsympathetic
Credible – Incredible
Friendly – Unfriendly
Competent – Incompetent
Appealing – Unappealing (*sender image*)
17. Based on my reading of the text, I find the D*'s products:
    Nice – Not nice
    Appealing – Unappealing
    Qualitatively good – Qualitatively bad (*attitude toward the product)*
18. Based on my reading of the text, the products of D* are something
    I certainly would like to try – I certainly would not like to try
    I certainly would purchase – I certainly would not purchase
    I certainly would recommend to a friend – I certainly would not recommend to a friend. (*purchase intention*)

Notes

[1] A more detailed description of the various approaches to translation quality is found in Koby *et al.* (2014), Melby *et al.* (2014) and Van Egdom *et al.* (forthcoming).

[2] There have been laudable attempts to contrast expert and layman perceptions of MT quality (Arnold *et al.* 1994; Roturier 2006; Bowker and Ehgoetz 2007).

[3] The suggestion had been posited that existing guidelines for PE be used for this experiment. However, a comparison of guidelines revealed that there is a striking disparity between guidelines for the same level of text manipulation (Hu and Cadwell 2016). Furthermore, ending up with only two post-edited versions of a source text seemed undesirable.

[4] This metric system was common currency in early research on PE effort. Present-day research on post-editing effort has abandoned this method, for obvious reasons. The metrics do not register deletions, substitutions and insertions – let alone hesitations. In recent years, more effective means to tap into effort (e.g. keystroke logging, eye tracking) have been found (see O'Brien 2006a, 2006b; Vieira 2016, 2017).

[5] The abbreviations stand for *Informationsangebot* (IA, tr. 'offer of information'), *Translatum* (Trl, i.e. the translation product), *Funktion* (f, 'function') and *Skopos* (Sk, i.e. communicative purpose). The formula simply states that the translation product (which is defined as an offer of information) is always determined by a skopos, which allows the product to function in context.