To my eight-year-old self, who said she'd write a book
and to her wonderful parents, who taught her she could.

Promotor	Prof. dr. Lieve Macken
	Vakgroep Vertalen, tolken en communicatie
Copromotor	Prof. dr. Sonia Vandepitte
	Vakgroep Vertalen, tolken en communicatie

Decaan	Prof. dr. Marc Boone
Rector	Prof. dr. Anne De Paepe

UNIVERSITEIT GENT

Faculteit Letteren & Wijsbegeerte

Joke Daems

A translation robot for each translator?

A comparative study of manual translation and post-editing of machine translations: process, quality and translator attitude

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de vertaalwetenschap

2016

Translation is a fine and exacting art, but there is much about it that is mechanical and routine and, if this were given over to a machine, the productivity of the translator would not only be magnified but his work would become more rewarding, more exciting, more human.

- Martin Kay, The Proper Place of Men and Machines in Language Translation, 1980

Abstract

To keep up with the growing need for translation in today's globalised society, post-editing of machine translation is increasingly being used as an alternative to regular human translation. While presumably faster than human translation, it is still unsure whether the quality of a post-edited text is comparable to the quality of a human translation, especially for general text types. In addition, there is a lack of understanding of the post-editing process, the effort involved, and the attitude of translators towards it.

This dissertation contains a comparative analysis of post-editing and human translation by students and professional translators for general text types from English into Dutch. We study process, product, and translators' attitude in detail.

We first conducted two pretests with student translators to try possible experimental setups and to develop a translation quality assessment approach suitable for a fine-grained comparative analysis of machine-translated texts, post-edited texts, and human translations. For the main experiment, we examined students and professional translators, using a combination of keystroke logging tools, eye tracking, and surveys. We used both qualitative analyses and advanced statistical analyses (mixed effects models), allowing for a multifaceted analysis.

For the process analysis, we looked at translation speed, cognitive processing by means of eye fixations, the usage of external resources and its impact on overall time. For the product analysis, we looked at overall quality, frequent error types, and the impact of using external resources on quality. The attitude analysis contained questions about perceived usefulness, perceived speed, perceived quality of machine translation and post-editing, and the translation method that was perceived as least tiring. One survey was conducted before the experiment, the other after, so we could detect changes in attitude after participation. In two more detailed analyses, we studied the impact of machine translation quality on various types of post-editing effort indicators, and on the post-editing of multi-word units.

We found that post-editing is faster than human translation, and that both translation methods lead to products of comparable overall quality. The more detailed error analysis showed that post-editing leads to somewhat better results regarding

adequacy, and human translation leads to better results regarding acceptability. The most common errors for both translation methods are meaning shifts, logical problems, and wrong collocations. Fixation data indicated that post-editing was cognitively less demanding than human translation, and that more attention was devoted to the target text than to the source text. We found that fewer resources are consulted during post-editing than during human translation, although the overall time spent in external resources was comparable. The most frequently used external resources were *Google Search*, concordancers, and dictionaries. Spending more time in external resources, however, did not lead to an increase in quality. Translators indicated that they found machine translation useful, but they preferred human translation and found it more rewarding. Perceptions about speed and quality were mixed. Most participants believed post-editing to be at least as fast and as good as human translation, but barely ever better. We further discovered that different types of post-editing effort indicators were impacted by different types of machine translation errors, with coherence issues, meaning shifts, and grammatical and structural issues having the greatest effect. HTER, though commonly used, does not correlate well with more process-oriented post-editing effort indicators. Regarding the post-editing of multi-word units, we suggest 'contrast with the target language' as a useful new way of classifying multi-word units, as contrastive multi-word units were much harder to post-edit. In addition, we noticed that research strategies for post-editing multi-word units lack efficiency. Consulting external resources did lead to an increased quality of post-edited multi-word units, but a lot of time was spent in external resources when this was not necessary.

Interestingly, the differences between human translation and post-editing usually outweighed the differences between students and professionals. Students did cognitively process texts differently, having longer fixation durations on the source text during human translation, and more fixations on the target text during post-editing, whereas professional translators' fixation behaviour remained constant. For the usage of external resources, only the time spent in dictionaries was higher for students than for professional translators, the usage of other resources was comparable. Overall quality was comparable for students and professionals, but professionals made fewer adequacy errors. Deletions were more noticeable for students than for professional translators in both methods of translation, and word sense issues were more noticeable for professional translators than for students when translating from scratch. Surprisingly, professional translators were often more positive about post-editing than students, believing they could produce products of comparable quality with both methods of translation. Students in particular struggled with the cognitive processing of meaning shifts, and they spent more time in pauses than professional translators.

Some of the key contributions of this dissertation to the field of translation studies are the fact that we compared students and professional translators, developed a fine-grained translation quality assessment approach, and used a combination of state-of-

the-art logging tools and advanced statistical methods. The effects of experience in our study were limited, and we suggest looking at specialisation and translator confidence in future work. Our guidelines for translation quality assessment can be found in the appendix, and contain practical instructions for use with brat, an open-source annotation tool. The experiment described in this dissertation is also the first to integrate *Inputlog* and *CASMACAT*, making it possible to include information on external resources in the *CASMACAT* logging files, which can be added to the CRITT Translation Process Research Database.

Moving beyond the methodological contributions, our findings can be integrated in translation teaching, machine translation system development, and translation tool development. Translators need hands-on post-editing experience to get acquainted with common machine translation errors, and students in particular need to be taught successful strategies to spot and solve adequacy issues. Post-editors would greatly benefit from machine translation systems that made fewer coherence errors, meaning shift errors, and grammatical and structural errors. If visual clues are included in a translation tool (e.g., potentially problematic passages or polysemous words), these should be added to the target text. Tools could further benefit from integration with commonly used external resources, such as dictionaries.

In the future, we wish to study the translation and post-editing process in even more detail, taking pause behaviour and regressions into account, as well as look at the passages participants perceived as the most difficult to translate and post-edit. We further wish to gain an even better understanding of the usage of external resources, by looking at the types of queries and by linking queries back to source and target text words.

While our findings are limited to the post-editing and human translation of general text types from English into Dutch, we believe our methodology can be applied to different settings, with different language pairs. It is only by studying both processes in many different situations and by comparing findings that we will be able to develop tools and create courses that better suit translators' needs. This, in turn, will make for better, and happier, future generations of translators.

Samenvatting

Door toenemende globalisatie wordt er steeds vaker gebruikgemaakt van automatische vertaalsystemen, waarvan de output vervolgens gepost-edit wordt. Hoewel dit post-editen vermoedelijk sneller is dan gewoon vertalen, is het niet zeker of de uiteindelijke kwaliteit zo goed is als die van een manuele vertaling, zeker niet wat algemene teksten betreft. Verder hebben we nog niet voldoende inzicht in het post-editproces, de inspanning die ermee gepaard gaat, en de attitude van vertalers.

In dit proefschrift wordt een vergelijkende studie gemaakt van post-editen en manueel vertalen uit het Engels naar het Nederlands door studenten en professionele vertalers voor algemene teksten. We kijken hierbij in detail naar het proces, het product, en de attitude van de vertalers.

We voerden een vooronderzoek uit met studenten om de experimentele opstelling te testen en om een methode voor kwaliteitsanalyse te ontwikkelen. Deze methode moest geschikt zijn voor een fijnmazige vergelijkende analyse van de output van automatische vertaalsystemen, gepost-edite teksten, en manuele vertalingen. Voor het hoofdonderzoek werkten we met studenten en professionele vertalers en gebruikten we een combinatie van toetsregistratiesoftware, oogregistratiesoftware en enquêtes. We voerden zowel kwalitatieve analyses als geavanceerde statistische analyses uit.

Voor de procesanalyse keken we naar snelheid, cognitieve belasting op basis van oogbewegingen, het gebruik van externe bronnen en de impact ervan op de totale benodigde tijd. Voor de productanalyse keken we naar globale kwaliteit en veelvoorkomende fouten. De analyse voor de attitude bevatte vragen rond het nut van automatische vertaalsystemen en post-editen, de snelheid, kwaliteit, en de vertaalmethode die het minst vermoeiend was. We brachten verschillen in attitude voor en na deelname in kaart. Naast de globale analyse voerden we nog twee meer gedetailleerde analyses uit, waarin we keken naar de impact die de kwaliteit van de output van het automatische vertaalsysteem had op verschillende vormen van inspanning tijdens het post-editen, en op het post-editen van collocaties.

We stelden vast dat post-editen sneller was dan manueel vertalen, en dat het eindproduct van beide vertaalmethodes een vergelijkbare kwaliteit had. Op basis van de

gedetailleerde foutenanalyse werd duidelijk dat post-editen beter scoort op vlak van adequaatheid (adequacy) en dat manueel vertalen beter scoort op vlak van aanvaardbaarheid (acceptability). Veelvoorkomende fouten waren betekenisverschillen, logische problemen, en fouten tegen collocaties. Post-editen was cognitief minder belastend dan manueel vertalen, en er werd meer aandacht geschonken aan de doeltekst dan aan de brontekst. Er werden minder externe bronnen geraadpleegd bij het post-editen dan bij manuele vertaling, hoewel de totale tijd die in externe bronnen gespendeerd werd vergelijkbaar was. De meest gebruikte bronnen waren Google Zoeken, concordantietools, en woordenboeken. Meer tijd doorbrengen in externe bronnen leidde echter niet tot betere kwaliteit. Vertalers vonden automatische vertaalsystemen nuttig, maar gaven de voorkeur aan manueel vertalen en haalden daar ook meer voldoening uit. De meningen over snelheid en kwaliteit waren verdeeld. De meeste deelnemers geloofden dat post-editen minstens even snel en minstens even goed kon zijn als manueel vertalen, maar vrijwel niemand dacht dat het beter was. Verder ontdekten we dat verschillende soorten fouten in de output van het automatisch vertaalsysteem een impact hadden op verschillende soorten post-editinspanning. Hierbij hadden coherentieproblemen, betekenisverschillen, en grammaticale en structurele fouten het grootste effect. HTER, een vaak gebruikte maat voor het aantal benodigde aanpassingen in een automatisch vertaalde tekst op basis van menselijke referentievertalingen, stemt niet goed overeen met procesaspecten die post-editinspanning voorspellen. Voor het post-editen van collocaties stelden we een nieuwe categorisatie voor op basis van 'contrast met de doeltaal', aangezien contrastieve collocaties moeilijker te post-editen zijn. Daarnaast waren de strategieën die gebruikt werden bij het post-editen van collocaties niet erg efficiënt. Het raadplegen van externe bronnen leidde tot een verbeterde kwaliteit, maar er werd ook veel tijd gestoken in het opzoeken van externe bronnen voor collocaties waarbij dit niet nodig was.

Opvallend genoeg waren de verschillen tussen manueel vertalen en post-editen vaak groter dan de verschillen tussen studenten en professionele vertalers. Studenten leken de tekst cognitief wel anders te verwerken: ze hadden langere fixaties op de brontekst bij manueel vertalen, en meer fixaties op de doeltekst bij post-editen. Het fixatiegedrag van de professionele vertalers bleef constant. Studenten maakten langer gebruik van woordenboeken, het gebruik van de overige externe bronnen was vergelijkbaar. De globale kwaliteit was eveneens vergelijkbaar, maar professionele vertalers maakten minder fouten tegen adequaatheid. Voor studenten kwamen omissies in beide vertaalmethodes vaker voor dan bij professionele vertalers, woordbetekenisfouten kwamen dan weer frequenter voor bij professionele vertalers dan bij studenten bij manueel vertalen. Vreemd genoeg waren professionele vertalers vaak positiever over post-editen dan studenten. Ze waren overtuigd dat ze met beide vertaalmethodes kwaliteitsvolle vertalingen konden leveren. Studenten hadden het vooral cognitief

moeilijk bij het verwerken van betekenisverschillen en hun totale pauzetijd was langer dan die van professionele vertalers.

De belangrijkste bijdragen van dit proefschrift aan de vertaalwetenschap zijn het feit dat we studenten en professionele vertalers vergeleken hebben, dat we een fijnmazige methode voor kwaliteitsanalyse ontwikkeld hebben, en dat we een combinatie gebruikten van moderne registratiesoftware en geavanceerde statistische methodes. Het effect van ervaring in onze studies was beperkt, en we raden aan om in toekomstig onderzoek te kijken naar factoren als specialisatie en zelfvertrouwen bij vertalers. De appendix bevat onze richtlijnen voor de kwaliteitsanalyse. Hier staan praktische instructies in voor het gebruik van brat, een open source annotatietool. De studie beschreven in dit proefschrift is ook de eerste waarbij *Inputlog* met *CASMACAT* gecombineerd wordt, waardoor het mogelijk wordt om informatie over externe bronnen toe te voegen aan de outputbestanden van *CASMACAT*. Deze kunnen op hun beurt toegevoegd worden aan de CRITT database voor onderzoek naar vertaalprocessen (Translation Process Research Database).

Naast deze methodologische bijdragen kunnen onze bevindingen ook gebruikt worden voor vertaalopleidingen, het ontwikkelen van automatische vertaalsystemen, en het ontwikkelen van vertaaltools. Vertalers moeten praktische post-editervaring opdoen om vertrouwd te raken met veelvoorkomende fouten in de automatische vertaaloutput. Studenten in het bijzonder moeten strategieën aangeleerd krijgen die hen helpen adequaatheidsproblemen te ontdekken en op te lossen. Post-editors zouden het meest baat hebben bij automatische vertaalsystemen die minder coherentiefouten maken, betekenisverschillen introduceren, en grammaticale en structurele fouten maken. Als er visuele hints aan vertaaltools toegevoegd worden (zoals waarschuwingen bij problematische passages of polysemie), dan kunnen die best aan de doeltekst toegevoegd worden. Daarnaast zouden deze tools ook verbeterd kunnen worden door integratie van vaak gebruikte externe bronnen zoals woordenboeken.

In de toekomst willen we het vertaal- en post-editproces nog meer in detail bekijken. Hierbij zouden we kijken naar pauzegedrag en regressies, alsook naar de passages die onze deelnemers moeilijk te vertalen en te post-editen vonden. Daarnaast willen we ook meer inzicht krijgen in het gebruik van externe bronnen, door te kijken naar de zoekopdrachten en door deze terug te koppelen aan bron- of doeltekstwoorden.

Hoewel onze bevindingen beperkt blijven tot het post-editen en manueel vertalen vanuit het Engels naar het Nederlands voor algemene teksttypes, zijn we overtuigd dat onze methodologie ook nuttig kan zijn voor andere situaties, met andere talencombinaties. Pas wanneer beide vertaalmethodes in verschillende situaties onderzocht worden en deze gegevens met elkaar vergeleken worden, kunnen we vertaaltools en lessen ontwikkelen die beter aansluiten bij de noden van vertalers. En het is zo dat we ervoor kunnen zorgen dat de volgende generatie vertalers nog beter en gelukkiger is.

Acknowledgements

Who would have thought that after four years of writing papers, these next few pages would be the most daunting to write? For fear of forgetting anyone, I would like to use this space to thank the people that have helped and supported me, both before I even thought of attempting such a project (arguably a smarter me) and during the attempt itself.

I would not be where I am today without the constant availability, feedback, and support of my supervisor, Lieve Macken. I've heard it said that a good supervisor is paramount to a successful PhD, and I could not possibly have wished for a better one. Lieve, thank you for believing in me when you had not much more than my determination and enthusiasm to go by. Despite me being your first PhD student, you instantly found the balance between letting me figure things out on my own and helping me out when I couldn't. I am sorry for often working so close to the deadline and I am grateful for the frequent last-minute work you put in. Thank you for keeping my feet on the ground when necessary, and for making me aware of my accomplishments when I couldn't see them myself.

My sincerest thanks also go out to the other members of my doctoral guidance committee (DGC). First of all my co-supervisor, Sonia Vandepitte, for her input and feedback when writing papers, and in particular for her engaging conversations regarding the more fundamental aspects of linguistics and terminology. And of course Robert (Rob) Hartsuiker, our insider from the Department of Experimental Psychology, for his expertise with eye trackers, and his help with the more challenging aspects of the experimental setup and analysis. In addition to the DGC, I owe my gratitude to the members of the jury, Gert De Sutter, Veronique Hoste, Sharon O'Brien, and Arnt Lykke Jakobsen, for taking the time to read my work and for their insightful questions that paved the way for future work. I'm also grateful to Marleen Van Peteghem, for presiding over the jury meetings, and Orphée, for her amazing work as secretary, providing me with a clear synthesis of the jury's feedback. Furthermore, I want to thank Gitte for preparing the cover and manuscript for print.

The me from four years ago could not have written the book you are holding. Part of what makes research so satisfying is being given the opportunity to learn new things, to pick up new skills. Before I thank (some of) the people that gave me the tools I needed, I would first and foremost like to thank Elke Van Steendam, my master's thesis supervisor, for without her input, I would never even have considered pursuing an academic career. Furthermore, I am grateful to the instructors of the course on Paradigms and Instruments of Experimental Psychology at Ghent University for expanding my horizon on a theoretical as well as a technical level. I'm indebted to Luuk Van Waes, Mariëlle Leijten, and Michael Carl, both for helping to develop the main tools used in my final experiments (Inputlog and CASMACAT) and for helping me figure out how best to combine them and analyse the output. For the statistical analyses, the feedback from Koen Plevoets was invaluable. In addition, I have nothing but fond memories of my time spent in Copenhagen, where the brilliant minds of Arnt, Michael, Moritz, and so many others helped deepen my knowledge of translation process research, and where I was later given the opportunity to share my own knowledge with other researchers. From midnight data analysis to creating the perfect smørrebrød, Copenhagen has been an amazing experience and I'm glad to have met such wonderful people there.

Speaking of wonderful people, few are as wonderful as the (ex-)members of the LT³ team. Colleagues are a crucial element in any work environment, and I doubt I could find better ones anywhere else. From the start, everyone has been welcoming and supportive. Veronique, Lieve, Els, Kathelijne, Isabelle, Klaar, Peter, Bart, Orphée, Cynthia, Arda, Nils, Sarah, Mariya, Ayla, Gilles, and Stef, thank you for the many inspiring conversations during lunch and chats over coffee, thank you for letting me rant when I needed it, for cheering me up when I had a bad day, and for helping me even when you were busy yourselves, thank you for the regular birthday treats and barbecues, for indulging (and taking part in!) my quidditch obsession, and for joining me on after-lunch Pokémon hunts. Thanks in particular to Bart for teaching me how to use the brat tool, to Ayla and Cynthia for proofreading my dissertation, and to Orphée for helping me navigate the tricky waters of that final stretch. Perhaps even more so than the research topic, it's been our group's dynamic and atmosphere that made these four years so fulfilling.

Of course, there is one name missing from the above list of colleagues: Marjan, my workplace-proximity-associate-turned-friend, my ranting, binge-watching, and drinking buddy. We haven't shared an office in a while, but you will always be my favourite roomie. Thank you for making the time at the office more fun, and for the entertaining double-dates with your trophy husband.

Thank you to all of my friends who've supported me, especially the ones that remember what it is I do exactly (no, sorry, I still don't work for Google). Zoë, thank you for regularly checking up with me and for wanting to proofread the dissertation. And to

everyone from Ghent Gargoyles Quidditch: thank you for helping me maintain my play-work balance. A special thank you to Annemarie for the almost daily Skype talks with mutual symbolic rock-lifting, for your sincere interest and continuous support, for proofreading my work, and for helping me take a break from writing by, well, writing some more.

I'm grateful to my family for encouraging me to perform better while allowing me to choose my own path, for believing in me even when I didn't believe in myself. I'm grateful in particular to my parents, Moema and Chievo, for providing a loving home, quality education, and room to grow. I'd also like to thank my 'little' brothers for keeping me sharp, and Thomas in particular for his help with my cover (even though I didn't end up using his design).

Last, but, as cliché would have it, most certainly not least, Kirsten, my *keppie*. I don't have the words to express how grateful I am to have you in my life. Thank you for tolerating my absence and lack of input in the household whenever another deadline was around the corner. Thank you for believing in me, for being there for me through the ups and downs, and for looking after me when I needed it the most. Without your unwavering patience and support, these four years would have been close to impossible. Thank you for being on my team.

List of Tables

List of Figures

Table of Contents

Introduction

...the computer, far from being a threat to your livelihood, can become an essential tool which will make your job easier and more satisfying.

- Harold Somers, Computers and Translation

The main goal of technology is to improve human life. Whether it is used to increase productivity, comfort, or safety, technological advances have worked their way into every aspect of modern society. In that sense, describing the 'ideal translation tool' would not be so different from, for example, describing the 'ideal car'. Simply put, the goal of a translation is to go from point A - the source text - to point B - the target text, just like a car literally gets people from point A to point B. In theory, any car might be able to take you where you want to go, just as any tool might allow you to produce a translation. In practice, however, the journey will be quite different when you are driving a fully manually operated car without any form of technological support compared to when you are driving a fully automated vehicle. The first car requires the driver to do all the work, whereas the second has technological advances that have proven their worth: GPS has been shown to improve drivers' travel time and prevent traffic congestions (Taylor, Woolley, & Zito, 2000), drivers prefer individually tailored route planning systems (Letchner, Krumm, & Horvitz, 2006), intelligent driver-assist technology increases road safety (Pilipovic, Spasojevic, Velikic, & Teslic, 2014), and fatigue detection systems can reduce traffic accidents (Hailin, Hanhui, & Zhurnei, 2010). While, objectively speaking, the advances of the car with technological support outweigh those of the manually operated car, the human factor needs to be taken into account as well. Novice drivers, for example, have been shown to benefit more from automation than expert drivers (Young & Stanton, 2007). In addition, a lot of people take pride in knowing how to 'drive a stick', and while most people have accepted the usefulness of having a GPS installed, they might feel that other monitoring systems are too invasive, causing them to reject the technology.

It is hard to determine what exactly makes humans - or users - accept technology. According to Dillon (2001), "technology must satisfy basic usability requirements and be

perceived as useful by its intended user community" (para. 7) in order to be accepted. He further adds that experience, training, and implementation will additionally impact the acceptance levels. These assumptions were confirmed by Demiris et al. (2004), showing that even older adults were willing to accept technology, provided the advantages of using the technology were clear, the devices were user-friendly, and the training was adapted to the users. In a study looking at whether or not users would accept persuasive interfaces in their car, Meschtscherjakov, Wilfinger, Scherndl, and Tscheligi (2009) found that the intention to use the interface goes hand in hand with its perceived usefulness. But there is also a danger in trusting technology too much. Drivers' response to critical situations, for example, was much slower in an automatic driving condition compared to a manual driving condition (Merat & Jamson, 2009).

Just as cars have become increasingly automated to the point where they can practically drive themselves, the translation process is becoming increasingly automated:

> Above all, the degree of computerization permeating all aspects of the translation work environment has risen. Software is used for creating translation memories, aligning texts, managing terminology, checking spelling and grammar, accessing and searching electronic corpuses, and carrying out machine translation. (Gambier, 2016)

The computer has become the most important tool for the modern day translator. Word processing tools such as spell checkers and electronic dictionaries have been around for quite a while now (Sager, 1994), and most translators will no longer question their use. Even translation memory systems, which allow translators to reuse previously made translations, have been marketed commercially since the mid-1990s and "in a short period of time they were quickly accepted by users" (Somers, 2003a, p. 33). The use of fully-automated translation, or 'machine translation', is also on the rise (Koponen, 2016a). Technological advances in translation tools have been shown to reduce cognitive effort and increase speed (Screen, 2016), efficiency, and consistency (Austermühl, 2001), yet the human factor should not be forgotten either. In fact, translators often tend to dislike machine translation, as they are used to producing high-quality products and take pride in their work, they do not believe the technology is all that useful (Koehn, 2009), and they see machine translation as something that might eventually take their job away (Krings, 2001). As with cars, automation (in this case, machine translation) has been shown to be beneficial especially to novice users (Garcia, 2010; Yamada, 2015), but there is also the risk of them trusting the MT output too much (Depraetere, 2010).

In order to truly determine the acceptance and usefulness of translation technology, we need to take various factors into account. Different aspects of the automation need to be compared to the manual task, to determine areas where automation improves or

facilitates the manual tasks. In addition, it is important to take the users into account, their attitude as well as their experience. For translation in particular, the manual process is regular human translation; the automated process consists of correcting the output of a machine translation system to create a text of publishable quality, a process known as post-editing. In this dissertation, we compare the differences between human translation and post-editing for student translators as well as professional translators in terms of process, product, and attitude.

Motivation

A 2009 report on the European translation industry showed it to have an annual growth rate of 10% (Rinsche & Portera-Zanotti), and it does not look like that trend is about to change anytime soon. Shiyab (2010) predicted there would be a big demand for translation in the years to come, with globalization moving faster, people wanting to learn about different cultures, international markets becoming accessible, and technology reducing communication costs. This, indeed, seems to be the case now. The strength of the translation industry today can be attributed to the interplay between technological advances and information accessibility (Drugan, 2013). Technologies - such as ever more complex phones or cars - are produced increasingly fast, while at the same time being distributed to more regions, leading to an increase in volume as well as reach. Translation is needed to access these foreign - often stronger - markets, but it is also time-sensitive because information becomes rapidly outdated. The technical nature of the documents to be translated and the fact that international organisations and traders need to be aware of global changes also add to translation's importance. In sum, the translation industry is currently one of the healthiest industries despite global economic downturn, to the extent that "[t]he volume of potentially available translation work goes beyond the capacity of all professionals put together" (Gambier, 2014, p. 4). As traditional human translation cannot keep up with current translation needs (Doherty, 2016), we need to, on the one hand, find ways to help professional translators manage the greater workload, and, on the other hand, find ways of engaging novice translators. This is where translation technology and, more specifically, machine translation and post-editing come in.

When Kay (1980, reprint 1997) first described his so-called *translator's amanuensis*, he described a tool consisting of a text editor with an integrated dictionary, morphological rules, the option to recall and use previous translations, placeholders, and interactive machine translation, but he also described it as a tool that "does not exist and probably never will" (p. 12). Today's Translation Environment Tools, however, look exactly like

that. Translation tools have gone through at least as many technological advances as the car industry (Doherty, 2016; Gambier, 2016). They contain all basic features that are common in advanced text editors, such as spelling and grammar checking, with additional access to (online) dictionaries and term banks for terminology (Somers, 2003b). The option to recall and use previous translations is the exact function of a Translation Memory (TM), which is one of the most popular aspects of translation tools (Somers, 2003a). Modern tools such as *SDL Trados* and *Wordfast* integrate both translation memory functionalities and machine translation: when there is no match in the translation memory, machine translation output is given instead (Reinke, 2013). Even interactive machine translation is no longer a thing of the future, but something that is already being integrated into today's translation tools such as *TransType* (Macklovitch, 2006), *MateCat* (Bentivogli et al., 2016; Federico et al., 2014), and *Lilt*, the commercial successor of *Predictive Translation Memory* (Green, Chuang, Heer, & Manning, 2014).

Machine translation is here, but, does it work? Industry results seem to indicate it does, with companies reporting significant time gains when post-editing machine translation output compared to when translating from scratch (Aranberri, Labaka, Diaz de Ilarraza, & Sarasola, 2014; Groves & Schmidtke, 2009; Plitt & Masselot, 2010; Zhechev, 2014), whereas productivity findings for general text types are more mixed (Carl, Dragsted, Elming, Hardt, & Jakobsen, 2011; Garcia, 2010, 2011; Krings, 2001). In fact, some researchers question whether machine translation should be used for anything but technical texts: "MT is only suited for a very limited range of text types, and source texts have to be carefully tailored to the capabilities and restrictions of an MT system to minimize the amount of time and effort needed for post-editing" (Reinke, 2013, p. 35).

Surprisingly, most researchers note that post-editing leads to products of comparable quality to human translation, and sometimes even to products of better quality than human translations (Fiederer & O'Brien, 2009; Garcia, 2010; O'Curran, 2014; Plitt & Masselot, 2010). The focus is starting to shift from 'can post-edited texts be as good as human-translated texts' to understanding the quality differences in more detail by introducing more fine-grained error analyses (Koponen, Aziz, Ramos, & Specia, 2012; Stymne et al., 2012) and by finding ways to automatically predict machine translation quality (de Souza, Turchi, & Negri, 2014; Denkowski & Lavie, 2012).

What is presumably different, however, is the post-editing process compared to the human translation process. More and more post-editing research is being conducted, and this research often gives rise to new questions specific to the post-editing process (Dillinger, 2014). According to Dillinger (2014), the advancement of technological research tools such as keystroke loggers and eye trackers revolutionized translation research, enabling us to look at the translation and post-editing process in more detail, giving rise to more specific questions every day. The post-editing process is expected to be different from the translation process (O'Brien, 2002), because post-editors do not only need text production and interlingual communication skills, but they also need to

understand the MT jargon (Schütz, 2008), as well as be able to spot typical machine translation errors (Schäfer, 2003). The various elements at play during human translation and post-editing are visualised in Figure 1. The main process is the same for human translation and post-editing: anyone trying to convert a text from language A to language B needs to understand the goal (skopos) of the text to guide the process. They further need to understand the source text, be able to transfer this meaning into the target language, and produce the actual target text. This process is usually followed by a revision and review phase. Where, during the regular translation process, external resources are the main additional source of information for a translator (these can be consulted at any time during the process, and can be monolingual or bilingual), a post-editor has the machine translation output to work from. The machine translation system more or less takes over the 'transfer of meaning' step of the translation, although the post-editor needs to understand the output in order to turn it into a publishable target text. Post-editors consult external resources as well, although they presumably consult fewer resources than regular translators (Krings, 2001).

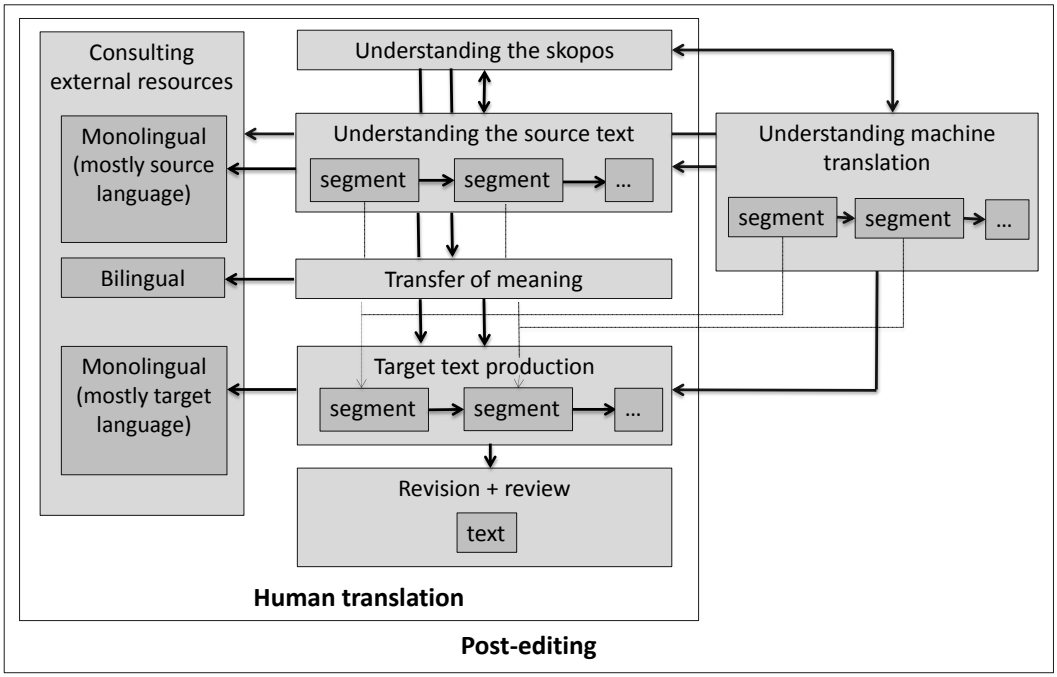


Figure 1 Schematic representation of the regular human translation process and the post-editing process.

Understanding the differences between both processes is important to learn if and when post-editing is a viable alternative to human translation, and which aspects can still be improved in order to make it a viable alternative more often. Post-editing has frequently been shown to be faster than human translation, but speed is not the only factor at play. Krings (2001), for example, found that post-editing leads to an increase in cognitive effort compared to human translation. In addition, we should not forget about

the human factor. As users only accept technologies they find useful (Dillon, 2001), we need to take translators' attitude into account when studying post-editing processes and products. In opinion pieces or translators' blogs and online groups, the language used to describe machine translation is often pretty strong, with post-editing being described as "linguistic janitorial work" (Kelly, 2014). More academic surveys investigating the attitudes of a larger number of respondents are harder to find. Fulford (2002) surveyed freelance translators and found that, although few of them used machine translation, most of them were interested to learn more about it. A decade later, a survey conducted by Optimale (2012) showed that only 28% of respondents considered the ability to post-edit machine translation 'essential' or 'important', compared to 76% of respondents when asked about the ability to use translation memory systems. This suggests that, while translation memories are now commonplace, machine translation is not quite as established. Guerberof (2013) found that translators are aware of the fact that post-editing will become a necessary skill in the translation industry, but that their experiences and feelings are mixed.

One way to make sure future translators are aware of machine translation and its limitations is by integrating post-editing in the translation curriculum. One of the best-known proposals for post-editing course content was developed by O'Brien (2002). In addition to the fact that post-editing will become a necessary skill for most future translators, Depraetere (2010) noted that students may be less averse to machine translation than experienced translators who are already used to producing high-quality content. A more recent study indeed indicated that certain types of college students would make decent post-editors (Yamada, 2015).

In sum, the usage of machine translation is expected to increase with the increased demand for translation, making post-editing a necessary skill for modern-day translators. While translation memory systems have become well established and form an integral part of modern translation courses, machine translation is used reluctantly and post-editing courses are still being developed. We need more detailed research to understand how and when post-editing works, by taking into account as many aspects as possible, namely the process (speed as well as cognitive aspects of the process), the quality of the final product, and the translators' attitude. This knowledge can then be used to improve translators' understanding (and, indirectly, acceptance) of post-editing. In addition, it can be integrated into translator training, and it can be used to improve machine translation systems to better suit post-editors' needs.

Research objectives

The main goal of the research presented in this dissertation, is to gain a better understanding of some of the differences between human translation and post-editing, for translation of general text types from English into Dutch, by student translators and professional translators. We chose general text types because, while post-editing for technical and controlled texts is somewhat established already, findings for post-editing with general text types are generally more mixed. We do this by studying the process, the product, and the translators' attitude, while trying to answer the following research questions:

What are the differences in process between human translation and post-editing?

1) Is post-editing faster than human translation?
2) Is post-editing cognitively more demanding than human translation?
3) Is the eye fixation behaviour different for post-editing and human translation?
4) Are more (or other) external resources consulted in human translation compared with post-editing?

In order to gain insight into the various aspects of the translation and post-editing process, we need to use keystroke logging tools and eye-tracking tools. Combined, these tools can register translation speed, keystrokes and pauses, fixations, and external resources.

What are the differences in product between human translation and post-editing?

1) Is there a difference in overall quality between the product of human translation and the product of post-edited machine translation output?
2) Is there a difference in the most common error types in human translations and post-edited texts?

What is the impact of machine translation quality on post-editing?

1) What is the impact of overall machine translation quality on post-editing effort?
2) What is the impact of specific machine translation errors on post-editing effort?
3) How does the machine translation output for multi-word units affect post-editing quality?

4) How does the machine translation output for multi-word units affect the consultation of external resources during post-editing?

In order to study the quality aspects, we need a translation quality assessment approach suitable for a fine-grained comparative analysis of human translations, post-edited texts, and machine translation output.

What are the differences in attitude towards human translation and post-editing?

1) How rewarding is post-editing compared to human translation?
2) How useful is MT output according to translators?
3) Which translation method is perceived as being faster?
4) How is the quality of both methods of translation perceived?
5) Which translation method is the most preferred translation method?
6) Is there a difference in perception before and after the experiment?

Attitude research up to now is limited and has led to very mixed results. To gain better insights into translators' attitude towards machine translation and post-editing, we need to conduct surveys before translators participate in a post-editing experiment, and after they participated.

What are the differences between student translators and professional translators?

If we want to determine how we can make post-editing work for its users, we need to understand the differences between different types of potential users, in this case professional translators and students. The factor 'experience' does not lead to isolated research questions, but should be seen as the following addition to the abovementioned questions: *...and is there a difference between students and professional translators?*

Thesis outline

Chapter 1 contains an overview of the related research in the fields of translation process research (more specifically, speed, cognitive load, and the usage of external resources), translation quality assessment, and the limited research available that deals with translators' attitudes.

Chapter 2 adds onto the work presented in the first chapter, and contains the hypotheses for the main analysis of this dissertation. The other chapters are dedicated to the experiments and their analysis.

In Chapter 3, we describe two pretests we conducted with students of translation. We used the pretests to explore possible experimental setups and tools, and to develop our translation quality assessment approach. The remaining chapters deal with the main experiments.

The method of our main experiment can be found in Chapter 4. It contains a discussion of our text selection process, the creation of the surveys, and the tools we used, as well as a description of the participants (students and professional translators), and the procedure during each session of the experiment. We also discuss how the data was prepared for statistical analysis and addition to the *Translation Process Research Database (TPR-DB)*.

Chapter 5 contains the global analysis of the main experiment: a comparative analysis of process, product, and attitude between human translation and post-editing as well as between students and professional translators. The process is studied through speed, fixations, and the usage of external resources. The quality is studied at various levels of granularity. Attitude is studied via two surveys: the first taken before participating in the experiment, the second after participating. We further look at the impact of the usage of external resources on overall time, and on final quality.

The final two chapters contain more specific analyses of the impact of machine translation output quality on subsequent post-editing. The relevant literature as well as the hypotheses for those chapters will be discussed within the chapters, and not in Chapters 1 and 2. In Chapter 6 we look at the impact of machine translation quality on post-editing effort. We discuss the error types found in the MT output, and the way different error types affect different aspects of post-editing quality, process aspects as well as product aspects. Chapter 7 presents an analysis of the impact of machine translation quality on students' processing of multi-word units. We suggest a new parameter to be added to multi-word unit classifications dealing with post-editing, and we discuss the final post-edited quality in light of these categories, as well as the usage of external resources while processing multi-word units during post-editing.

The final chapter of this dissertation is dedicated to a summary of our most important findings and conclusions.

Chapter 1 Related research

In this chapter, we discuss the relevant literature for the main aspects of our research. First, we establish the differences in process between human translation and post-editing, focussing on speed, cognitive load, and the usage of external resources. Second, we look at the product, where we address ways of analysing translation quality, as well as the differences between human translation and post-editing. Third, we consider translators' attitude towards machine translation and post-editing, drawing from surveys as well as experimental research. In the last section, we discuss the importance of experience and its relationship to the three aspects under scrutiny: process, product, and attitude.

1.1 Process

> *Accessing the black box and gaining a better understanding of what goes on during translation will advance the field of study, open new areas of research, and improve the way translation is viewed and taught.*
>
> *- Sabine Lauffer, The translation process: An analysis of observational methodology*

As the translation process itself is a highly complex and holistic process, researchers should take a holistic and integrative approach when analysing it, combining elements from empirical science and liberal arts, i.e., rigorous and controlled experiments while allowing for philosophical interpretation and theoretical exploration (Hansen, 2010a). It should be studied from different angles, taking the human subjects and cognition into account, as well as the texts and their translated products (Hansen, 2010a). To do so, there has been a shift from rather intrusive methods such as think-aloud protocols (TAP) to new, more unobtrusive and more ecologically valid tools such as keystroke

logging tools and eye trackers. Whereas TAP have been used to elicit problem-solving strategies and other steps in the translation process (Bernardini, 2001; Jääskeläinen, 2002; Kußmaul & Tirkkonen-Condit, 1995), they have been shown to influence the translation process itself, leading to a decrease in productivity and a processing of the text in smaller segments (Jakobsen, 2003; Krings, 2001). To get the whole picture, researchers have suggested using different methods at the same time, a process that is referred to as 'triangulation' (Alves, 2003; Jakobsen, 2006; O'Brien, 2004). A comparable trend can be found in post-editing research, with Schütz (2008) suggesting a combination of keystroke logging and eye tracking. The first registers time, keystrokes, and pauses and can be used to study productivity, pause behaviour, and text production. The latter registers the location and duration of eye movements and fixations, and can be used to study mental workload and cognitive processing. According to Schütz (2008), the ultimate goal of the information gained with these tools is to "build systems that can better anticipate, learn and emulate effective human behavior [sic]" (para. 1).

1.1.1 Speed

From an industry perspective, translation speed is the most important process factor. With the increased level of globalisation and modern methods of sharing information, much more text needs to be translated into diverse languages, and information rapidly becomes outdated. An increased translation speed is paramount to a company's expansion - giving it access to new markets - and relationship with its customers - providing up-to-date information at all times. Early research into post-editing was therefore mainly interested in identifying whether or not post-editing was indeed faster than regular translation. Gibb (1985, as cited in Tirkkonen-Condit (1990)) stated that only 35% of a translator's work could be sped up by computerisation. Of course, computers back then were not nearly as advanced as they are today, and the shift in machine translation from rule-based to phrase-based statistical machine translation had not yet taken place. More recently, with more advanced computers at our disposal, researchers have looked into the actual process of so-called computer-assisted translation (CAT), of which the post-editing of machine translation is an aspect. Some of the tools discussed in the following paragraph integrate more functionalities than just machine translation, although our focus when discussing them will be on machine translation post-editing, as this aspect is most relevant to our research.

In her master's thesis, Guerra Martínez (2003) found post-editing to be much faster than human translation for the translation of marketing brochures - a genre not usually considered to be appropriate for translation by MT. She reported possible time-savings of five to six minutes for every 100 words. The benefits of using CAT tools, and, in particular, their machine translation functionalities, regarding processing speed have

been demonstrated repeatedly. Macklovitch (2006), for example, reported on a series of tests within the *TransType* project, using interactive machine translation in real translation agencies. He found that sessions during which the participants were allowed to use the interactive features of the tool were much faster than sessions where translators had no additional support. A comparable study was conducted by Koehn (2009), who observed the usage of his self-developed web-based translation tool *Caitra*. This tool combines prediction for auto-completion of a translation, a list with alternative translation options for each phrase, and machine translation output for post-editing. He discovered that computer-assisted translation almost always led to higher translation speed, with translators' speed increasing as they became more experienced with the tool. Post-editing was the fastest method of translation (39% faster than regular translation on average), presumably because between 74% and 91% of all characters of the MT output remained unchanged, decreasing the need for typing activity. Most pauses during post-editing lasted longer than 10 seconds, indicating that most time during post-editing was spent on thinking about changes. However, the actual changes themselves required relatively little time (Koehn, 2009). In the same year, Guerberof confirmed the advantages of using CAT tools, with the usage of translation memories leading to an approximate increase in speed of 5% compared to human translation, and the usage of post-editing leading to an increase in speed of approximately 16%.

The greatest productivity increase by using post-editing has been reported by industry users, with the productivity test conducted by Autodesk - a software company - being a commonly cited example. This test showed post-editing to be 42% to 131% faster than regular translation, depending on the language combination (Plitt & Masselot, 2010). More recently, these findings were confirmed by another test within the same company, reporting an increase in productivity of 37% to 92% depending on the language combination (Zhechev, 2014). It must be noted, however, that this excludes time spent looking at style guides or consulting terminology, elements that are often included in other research. As such, it is hard to compare these findings to other findings from the industry, for example, Microsoft reported productivity gains from 6% to 20% (Groves & Schmidtke, 2009), or to findings from studies with more general text types, such as the one conducted by Guerra (2003).

Not all studies report a convincing increase in translation speed when post-editing, however. Garcia (2011) as well as Carl, Dragsted, Elming, et al. (2011), for example, found post-editing to be sometimes faster than human translation, but not statistically significantly so. In the case of Garcia (2011), this might be explained by the fact that participants were not allowed to consult external resources during the post-editing process, although they could use them during the human translation process. Not being able to look things up could lead to increased insecurity, and thus a slower process. Garcia (2011) did find a significant effect when participants were post-editing into a

foreign language and they were allowed to use external resources. In the case of Carl, Dragsted, Elming, et al. (2011), the researchers indicated the small number of participants or lack of experience with the tools as potential factors that influenced the results. A study by Lee and Liao (2011) did not show an increase in speed for post-editing compared to human translation either, although the way post-editing was applied was rather peculiar. Since computers were not available to all participants, both the translation and post-editing tasks were conducted on paper, and students had to self-report the time they spent on each task.

As speed is such a crucial element to take into account, and findings for more general text types are mixed, we wish to establish whether there is a difference in speed between human translation and post-editing from English into Dutch for general text types. Moreover, we want to verify how this difference is influenced by experience. We use keystroke logging tools to get an accurate measurement of time.

1.1.2 Cognitive load

In addition to speed, it is important to study the cognitive aspects of both human translation and post-editing. Even if post-editing is found to be faster than human translation for general text types, if it is more cognitively demanding, it can cause exhaustion, which will lead to decreased productivity in the long run. Psychological research has shown that there are limitations to a person's working memory capacity, or the amount of information and processes that can be contained in the human mind at the same time (Miller, 1956). Macizo and Bajo (2006) discovered that reading a text for translation demands more of a person's working memory than reading a text for repetition. The question remains whether post-editing leads to a reduction or an increase of cognitive load. Looking at Figure 1 from a working memory perspective, we can expect the post-editing process to be cognitively more demanding than the human translation process (Krings, 2001), as post-editing provides a translator with an additional resource (the MT output), and thus an additional process to be contained in the translator's working memory. At the same time, it might also provide translators with useful information, making it easier for them to solve certain problems.

Building on the hypothesis that whatever a person is looking at is also what that person is thinking about - and thus cognitively processing - at a given time (Just & Carpenter, 1980), the most direct way to measure cognitive effort is through fixations: the longer a fixation lasts, the higher the level of cognitive processing is. Jakobsen and Jensen (2008) indeed found that when the complexity of a task increased (ranging from reading to actual translation), the average fixation duration became longer and the number of fixations increased. O'Brien (2007) compared human translation with post-editing and correcting translation memory (TM) matches and found post-editing to be

less cognitively demanding than human translation. As she worked with professional translators, this seems somewhat contradictory to the findings by Dragsted (2006) that professional translators perceived CAT as cognitively more demanding, but the key is in the word 'perceive': even though professional translators perceived CAT as being cognitively more demanding, the empirical data showed that it was actually cognitively less demanding. Both aspects (attitude/perception and empirical results) will be discussed in later chapters.

When looking at the cognitive processing of translation in more detail, there seems to be some difference in the processing of source and target text, with the target text requiring more attention during translation than the source text (Jakobsen & Jensen, 2008; Pavlović & Jensen, 2009). The differences become a bit less clear when comparing human translation with post-editing. Both Carl, Dragsted, Elming, et al. (2011) and Nitzke and Oster (2016) found longer total fixation times on the target texts than on the source texts for both translation methods. However, Carl, Dragsted, Elming, et al. (2011) found the effect to be stronger for post-editing, whereas Nitzke and Oster (2016) did not find differences in target text fixation between HT and PE. They did find differences in source text fixation, with shorter total fixation time when post-editing. In agreement with Carl, Dragsted, Elming, et al. (2011), Koglin (2015) found that, compared to human translation, fixation time on the target text was longer during post-editing. However, her findings for human translation indicate that fixation time on the source text was longer than on the target text, which is in contrast with the abovementioned works, where fixation time was always longer on the target text, regardless of translation method. Koglin (2015) suggested that the differences in experimental design could account for these different results, as participants in the Carl, Dragsted, Elming, et al. (2011) study had no previous post-editing experience, and there were time constraints imposed. The general trend in all three studies seems to be that fixations during post-editing are more target text-centred, and those during human translation more source text-centred, although the more detailed interactions seem to differ across studies.

An additional factor to take into account for post-editing is the machine translation output quality, which can also impact the cognitive effort involved. Doherty, O'Brien, and Carl (2010) used eye-tracking to identify MT comprehensibility and discovered that the total fixation time and number of fixations correlate well with MT quality. Stymne et al. (2012) looked at the relationship between different types of MT errors and fixation data. They found more and longer fixations on passages containing errors and determined that the average fixation time for word order errors as well as incorrect or missing words was the longest. A higher number of fixations for bad MT output compared to good MT output was also found by Doherty and O'Brien (2009), although they did not find a significant difference in fixation duration.

To gain a better understanding of cognitive processing during human translation and post-editing for general text types from English into Dutch, we use an eye tracker to

gather fixation data (the number of fixations and fixation duration). Fixations are mapped to the source and target text, allowing us to study the differences in focus between both areas for both translation methods (post-editing and human translation), and for both levels of experience (students and professionals). In later chapters, this fixation data allows us to study the relationship between machine translation errors and cognitive effort during post-editing.

1.1.3 External resources

A final aspect we are interested in regarding translation processes, is the usage of external resources, as they can provide insight into translators' problem-solving strategies (Göpferich, 2010) or uncertainty management (Angelone & Shreve, 2011). And these can, in turn, shed light on translation strategies as well as elements of translation competence (Hunziker Heeb, 2012). Indeed, the translation process could be considered to be a problem-solving process. Although the term 'problem-solving' is often used in a mathematical or scientific context, a translator is also trying to solve a problem, namely: how do I use the given information (i.e., the source text) and the tools I have at my disposal (i.e., external resources and translation strategies) to obtain the desired result (i.e., the currently non-existent target text). A translator can use knowledge (internal support) and research (external support) to solve problems and make decisions regarding the best solutions to these problems (PACTE, 2005). For post-editing in particular, the aspect of uncertainty gets an additional dimension: it is not just translators' uncertainty about their own knowledge, initiating search queries related to source text meaning, meaning transfer, or target text production, but also translators' uncertainty about the quality of the MT output. In fact, Pym (2013) lists 'learn to trust and mistrust data' as a key skill in the modern translation age, 'trust' indicating that translators sometimes do not trust technology enough, and 'mistrust' indicating that translators need to be aware of common MT errors, as well as be able to decide when it is better to translate from scratch.

Translators consult external resources when their 'internal' resources do not provide solutions to a problem. The external resource can either provide a direct solution, or provide alternative suggestions that spark the solution in the translators' mind (Pavlović, 2007). Pavlović (2007) studied collaborative translation processes and found that on average 17.48% of solutions to problems were arrived at after consulting external resources. Popular external resources in her study were bilingual dictionaries and texts, as well as *Google*. Chodkiewicz (2015) looked at the usage of external resources when revising translations and found that students mostly used *Google* queries and bilingual dictionaries, followed by *Wikipedia* and monolingual dictionaries. In a study with professional translators conducted by J. Gough (personal communication, April

2016), participants mostly used search engines, bilingual dictionaries, knowledge-based resources (such as Wikis), and webpages.

External resources are usually registered by means of screen capture software such as *Camtasia Studio* (Göpferich, 2010). The drawback of this software, however, is the fact that the data still need to be replayed and manually encoded before they can be analysed automatically, which can be quite time-consuming. Think-aloud protocols can provide some idea of the resources consulted, but participants' utterances are often incomplete and researchers still need to look at the screen recordings in parallel to make sense of their data (Ehrensberger-Dow & Künzli, 2010). Some previous research has made use of data gathered with the *TransSearch* tool to get a better insight in translators' queries (Macklovitch, Lapalme, & Gotti, 2008), but they were limited to one type of resource (*TransSearch*) and did not take other types of resources into account. There are tools, however, that are capable of logging multiple applications. *Inputlog* in particular is one of those tools. It is a keystroke logging tool that was originally intended for writing research, which logs all *Windows*-based applications (Leijten & Van Waes, 2013). In a recent study, *Inputlog* has been used to analyse the external resources used by a professional communication designer when creating a proposal (Leijten, Van Waes, Schriver, & Hayes, 2014).

We use *Inputlog* to gain a better understanding of the differences in usage of external resources during human translation and post-editing of general text types from English into Dutch, taking into account the number and type of resources, as well as the time spent in each external resource.

1.2 Product

> *The real danger is not that computers will begin to think like men, but that men will begin to think like computers.*
>
> *- Sydney Harris, H. Eves Return to Mathematical Circles*

Translation quality assessment (TQA) is a key task in the translation industry. It is necessary for the identification of translation problems, for the evaluation of the output of machine translation systems, and for the assessment of translators' work. For our research in particular, we needed a translation quality assessment approach allowing for a diagnostic and comparative analysis, and a method that was suitable for the evaluation of human translation, machine translation and post-editing.

We first discuss some of the existing methods of translation quality assessment, followed by an overview of studies related to human translation and post-editing quality.

1.2.1 Translation quality assessment methods

Ranking and scoring methods

Techniques often used for comparative analyses are – among others – ranking methods and scoring methods. Ranking consists of presenting reviewers with possible translations of a source text passage, and letting them rank those according to quality. This technique is commonly used in the evaluation campaigns of the workshops on statistical machine translation (Bojar et al., 2013) and has also been used to compare the quality of post-edited passages with that of human translated passages (Carl, Dragsted, Elming, et al., 2011).

Scoring methods can be implemented in various ways. Scoring methods often used for the quality assessment of machine translation output consist of quality scales for fluency (target language) and adequacy (meaning of the source text). The most common scale is a five-point scale, as used by the *Advanced Research Project Agency* (White, O'Connell, & O'Mara, 1994). Though originally designed for fluency and adequacy assessment, other parameters have been added, such as usability (Babych, Elliott, & Hartley, n.d.) or style (Fiederer & O'Brien, 2009). More recently, the scale method has also been applied to human translations or post-edited machine translations, as part of a holistic evaluation process (Colina, 2009; Fiederer & O'Brien, 2009; Waddington, 2001). Other scores are based on criteria for a specific purpose. Garcia (2010), for example, used experienced evaluators for the translation quality assessment of both human translations and post-edited machine translations. They scored the texts on the basis of criteria used by *NAATI*, the national standards and accreditation body for translators and interpreters in Australia. Texts are judged for accuracy (is the source text meaning and message accurately transferred?), quality of language (and the impact on accuracy), and application of good practices, with accuracy being the most important aspect. Points are deducted for each error, with accuracy errors weighing heavier than others.

Though the above-mentioned approaches have proven successful and valid for comparative analysis, they have their limitations. First, most scoring and ranking methods operate at the sentence level and do not take into account the text in context and the text as a whole. Second, ranking or scoring methods do not provide a fine-grained analysis needed for diagnostic research: they do not tell how translations differ. For example, a translation might receive a score of 3 and another translation for the same sentence might receive a score of 4, but there is no way of knowing what exactly makes the quality of the first translation worse than that of the second. Is it

grammatical problems? Lexical issues? A matter of coherence? Third, human annotators can use score scales in different ways (Wisniewski, Singh, & Yvon, 2013): what is a 'good' translation for one reviewer might be a 'very good' translation for another. If a sentence is very long, but contains one serious error, it might be unclear whether this should be considered a 'very bad translation' or a 'bad translation'. This evaluator subjectivity is one of the main points of criticism on translation quality assessment approaches (Williams, 2009).

Error analysis

A partial solution to the above-mentioned problems could be found in translation quality assessment through error analysis. This method consists of defining different error categories and marking or counting the errors of each category present in a certain text or sentence. Weightings can be added as well: either the evaluator decides on the weight for each instance of that error type - as is the case in the scheme used by the *ATA*, American Translators Association (2009) - or the error type as a whole receives a fixed weight - as is the case for BlackJack, a tool used by *ITR*, a British translation agency (ITR, 2002). Depending on the goal of the analysis, different error typologies may be adopted. Some error typologies have been specifically designed for the analysis of machine translation output (Farrús, Costa-jussà, & Mariño, 2011; Vilar, Xu, D'Haro, & Ney, 2006; Weiss & Ahrenberg, 2012), others for the analysis of human translations (Secara, 2005).

Although the process of error analysis is highly time-consuming, and although "[t]here is not always a direct relationship between the number and gravity of errors, the quality of the [translation] and the perceived acceptability and usability of the text" (Hansen, 2010b, p. 386), error analysis generates rich data, which is necessary for diagnostic and comparative evaluation of translations. Error analysis can help identify specific translation issues, and their effect on, for example, cohesion and readers' understanding (Weiss & Ahrenberg, 2012).

It is also crucial to the improvement of machine translation systems. While automatic evaluation metrics such as the widely used BLEU (Papineni, Roukos, Ward, & Zhu, 2002) can be used to compare the overall quality of different systems, a more detailed error analysis is necessary to identify specific strengths and weaknesses (Berka, Bojar, Fishel, Popovic, & Zeman, 2012; Stymne & Ahrenberg, 2012). Knowledge of the types of errors that an MT system makes has even been said to reduce post-editing time (Martínez, 2003). Identifying the main problems of automatic translation output can help create error profiles (Stymne & Ahrenberg, 2012) and focus research efforts (Vilar et al., 2006). Some errors can be integrated into automatic error detection, helping post-editors save time and notice more errors (Valotkaite & Asadullah, 2012). Linguistic information gained through error analysis can help improve statistical machine translation systems

(Farrús et al., 2011). Within the domain of quality estimation for machine translation, Wisniewski et al. (2013) argue for more fine-grained quality estimation scores operating at the sentence or word level. Post-editor training can also benefit from error analysis, since students need to be made aware of typical MT errors (Depraetere, 2010).

Error analysis is clearly a viable method of translation quality assessment, but here as well, some points of criticism and points of attention can be noted. Gouadec (1981) highlights the importance of clear error definitions as well as keeping track of the location of each error. Larose (1998) as well as O'Brien (2012) stress the importance of taking the translation situation into account, describing translation evaluation as a relative process. Often used metrics such as the *LISA QA* system for the localisation industry are criticized because of the fact that many categories overlap and that serious errors at macro-textual level are missing (Jimenez-Crespo, 2011). In her comparison of different error-based translation quality assessment approaches taken from the translation industry as well as from translation teaching, Secara (2005) notes some important conditions for good error-based evaluation methods: evaluation should not be limited to word and sentence levels; weightings for errors should be objective; error schemes should be hierarchical; a metric should be easy to use; both style and content should be taken into account; categories should be specific and clear, yet justifiable. One of the error annotation schemes under scrutiny is the scheme developed within the framework of the *MeLLANGE* project (2006), during which a *Learner Translator Corpus* was created, an aligned corpus enriched with linguistic information and error annotations. For a recent and more detailed discussion of metrics for translation quality assessment, see Mateo (2014).

Recent shifts in translation evaluation

O'Brien (2012) reported on a benchmarking exercise of different quality evaluation models currently in use in the translation industry. Arguing in favour of a more dynamic approach to translation quality evaluation, she was surprised to find that most models stick to a sentence-level error analysis and barely ever take text type or function into account. The dissatisfaction with current so-called 'one-size-fits-all' approaches, such as *SAE J2450* and the *LISA QA Model*, has led the translation industry to look for novel approaches that can cater to the needs of a rapidly evolving translation industry. Where *TAUS*, the Translation Automation User Society, has developed a *Dynamic Quality Framework* with corresponding guidelines for translation quality evaluation (TAUS, 2013a, 2013b), *QTLaunchPad*, a European Commission-funded collaborative research initiative, provides a *Multidimensional Quality Metric* which contains possible TQA issue types from which users can build custom metrics, tailored to their assessment task (QTLaunchPad, 2013). Both quality frameworks were developed almost simultaneously with the TQA approach used in the present study. We will discuss the differences and

likenesses in section 3.2. For future work, researchers suggest using fine-grained error typologies (Koponen et al., 2012; Stymne et al., 2012) and different languages (Koponen et al., 2012; Popovic, Lommel, Burchardt, Avramidis, & Uszkoreit, 2014; Stymne et al., 2012). In line with Temnikova (2010) and Lacruz et al. (2014), we believe that error categorisations need to incorporate a method for ranking the different errors according to severity.

1.2.2 Quality in human translation, post-editing, and machine translation

Errors are a part of the translation process. In human translation, errors can be caused by a lack of competence, a lack of understanding of the source language, or not knowing how to manipulate the target language (Séguinot, 1989).

Interestingly, post-editing has been found to be beneficial to translation's quality compared to human translation (Garcia, 2010; Plitt & Masselot, 2010), especially regarding accuracy (Fiederer & O'Brien, 2009; Lee & Liao, 2011). Carl, Dragsted, Elming, et al. (2011) reported that post-edited sentences were usually ranked as being better than sentences translated from scratch. Comparable results were obtained by Garcia (2011), who found that post-edited texts received better grades than texts translated from scratch.

Although the overall quality of post-editing can be better than that of human translation, studies with error analysis have shown a more nuanced picture. Guerberof (2009), for example, compared human translation with post-editing and translation from translation memory suggestions, and found that post-editing led to better final quality than translation from translation memory suggestions, although human translation still outperformed post-editing. The error types showed that human translation performs worse than post-editing regarding mistranslations, but post-editing performs worse than human translation for the other categories: accuracy, terminology, language, and consistency (Guerberof, 2009). In a study conducted by Lee and Liao (2011), post-editing led to fewer unintentional omissions, a better register, and more accurate translations, but human translation was found to be better for word use consistency.

Of course, post-editing effort (and, potentially, its quality) is influenced by machine translation output, which, in turn is influenced by source text characteristics. Long and complex sentences, short and ambiguous sentences, long noun phrases, and prepositional phrases can all be problematic for machine translation systems (Koponen, 2016b). In addition, post-edited texts stay closer to the source text structures than human translated texts (Depraetere, 2010).

We developed a fine-grained translation quality assessment approach to compare the quality of human translations and post-edited output in detail, as well as to observe machine translation quality and study its impact on post-editing effort, taking experience into account as an additional factor. The approach is discussed in more detail in section 3.2.

1.3 Attitude

> *Traditionally translators have had a reputation for being rather sceptical regarding the efficacy of MT, with views being reported that range from mild amusement through to open hostility and fear.*
>
> *- Heather Fulford, Freelance translators and machine translation*

In addition to final quality, translators' attitudes also matter. Even if post-editing is found to be faster, without having to compromise on quality, it is still important for translators to feel happy about their performance, seeing how users only accept technology they deem useful (Dillon, 2001). The common idea about translators' perception of machine translation is that they do not like it, as they fear it will take their jobs away (Krings, 2001). It is relatively hard to find studies focussing on translators' perception of and attitude towards machine translation and post-editing, although there have been some surveys throughout the years, and some researchers discuss attitude and perception as an aspect of a larger experiment.

One of the earlier surveys is the one conducted by Fulford (2002), which questioned freelance translators in the UK. Although only 23% of the respondents had received any type of MT training, 53% of the respondents had some experience with post-editing work. They found post-editing less rewarding than regular translation, although some of them seemed to derive some satisfaction from correcting MT errors. Though the translators in Fulford (2002) were mostly sceptical about MT, they were interested in learning more about it, and they particularly asked for ways of deciding when MT was appropriate and efficient. In a later paper on a more representative sample (Fulford's 2002 article was a report on 30 respondents filling out the exploratory survey), Fulford and Granell-Zafra (2004) found that 75% of participants were not familiar with MT, and that they, while happy with ICT, were not convinced of CAT tools, as the benefits of using them were not entirely clear.

A later survey conducted by Guerberof (2013) indicated that translators' attitudes towards MT were somewhat mixed. Hers was again a relatively small survey, with 24

translators and three reviewers as respondents. Twice as many translators felt that post-editing required more effort than human translation than vice versa. When asked whether they liked post-editing, nine participants said they disliked it, compared to seven participants who said they liked it, the others were indifferent.

Findings from empirical studies are mixed as well. Even though Koehn (2009) found post-editing to be as productive as other methods of translation, participants did not enjoy it and did not have the idea it was useful. Comparable to translators' desire to learn more about MT found in Fulford (2002), Tatsumi (2010) found that participants wanted to learn how MT worked and what the expected level of final quality was. There seems to be a trend of a more positive and flexible attitude towards MT and post-editing (de Almeida, 2013; Garcia, 2010; Lee & Liao, 2011; Tatsumi, 2010), with Garcia's and Lee & Liao's participants even thinking they would perform better when post-editing compared to when translating from scratch. Especially with customised MT systems, participants seem to feel that post-editing was faster, and that the MT output was useful (Green, Heer, & Manning, 2013). We are aware of only one study that looked into the English-Dutch language pair: Gaspari, Toral, Naskar, Groves, and Way (2014) compared actual with perceived post-editing effort and productivity. For translations into Dutch, post-editing was perceived as more effortful and more time-consuming than human translation, and participants preferred human translation over post-editing (Gaspari et al., 2014).

More recently, researchers have studied the attitude of final clients towards MT. Interestingly, most people seemed to prefer human translation, until they received information about the time and cost, after which they mostly prefered rapid post-editing (Bowker & Buitrago Ciro, 2015).

In order to gain a better understanding of the attitudes of students and professional translators towards post-editing, and to see whether their attitudes change after performing post-editing tasks themselves, we created two surveys: the first to be filled out before the experiment, the second to be filled out after. A discussion of the creation of these surveys can be seen in section 4.1.2.

1.4 Experience

> *...statistical-based MT, along with its many hybrids, is destined to turn most translators into posteditors one day, perhaps soon. And as that happens, as it is happening now, we will have to rethink, yet again, the basic configuration of our training programs. That is, we will have to revise our models of what some call translation competence.*
>
> *- Anthony Pym, Translation skill-sets in a machine-translation age*

Translation process research is necessary to learn about the qualities good translators possess (Hansen, 2010a). This knowledge can then be integrated into translation training, to make for a better future generation of translators. Longitudinal studies like the work conducted by the PACTE group or within the framework of TransComp set out to create models for translation competence as well as its acquisition. The translation competence model developed by PACTE (2003) is a model of characteristics that define professional translators and consists of several interacting sub-competences. Göpferich suggested another, though comparable, translation competence model in 2009. She assumed that the interaction between and coordination of the different sub-competences would improve with increased translation competence, and that beginning translators focused more on the surface level, whereas more advanced translators used more global and diversified strategies. Both models implicitly contain the assumption that professional translators are the more competent translators. Differences have indeed been found between professional and non-professional translators (A. Jensen, 1999; Séguinot, 1991; Tirkkonen-Condit, 1990), although there are studies that indicate that the differences are smaller than often thought (Jääskeläinen, 1996; Kiraly, 1995). An elaborate discussion of the issues related to experience and competence can be found in Jääskeläinen (2010). She addressed a few potential explanations for the seemingly incongruent findings listed above: professional translators might underperform in an experimental setup because they are not performing routine tasks, not all professionals can be expected to be experts - i.e., exhibit consistently superior performance - and specialisation might play an important role as well. Jääskeläinen (2010) concludes the chapter by stressing that future research needs to include clear definitions of expertise and professionalism, as well as relevant background information on subjects.

In this dissertation, we consider professionalism to mean 'having experience working as a professional translator'. We will compare this level of experience with that of student translators, who do not have any experience beyond their studies, for the aspects of translation and post-editing discussed above: process, product, and attitude.

We devote a separate section (section 1.5) to the discussion of teaching post-editing, as post-editing is a new yet crucial skill for the modern translator.

1.4.1 Experience and process

Regarding the translation task, inexperienced translators have been shown to treat it as a more lexical task, whereas professional translators pay more attention to higher-order concerns such as coherence and style (Séguinot, 1991; Tirkkonen-Condit, 1990).

A comparable trend is found in revision research. Sommers (1980), for example, found that experts adopt a non-linear strategy, focusing more on meaning and composition, whereas student revisers work on a sentence level and rarely ever reorder or add information. Hayes, Flower, Scriver, Stratman, and Carey (1987) as well reported that expert revisers first attend to the global structure of a text - the so-called higher-order concerns or HOC - whereas novice revisers attend to the surface level of a text - lower-order concerns or LOC. Broekkamp and van den Bergh (1996) found that students were heavily influenced by textual cues during the revision process. For example, if a text contained many grammatical errors, the reviser's focus switched to solving grammatical issues, and HOC were ignored.

Finding studies that compare the performance of more and less experienced translators regarding post-editing is much harder than finding studies related to regular translation or revision. The following sections therefore contain findings from a variety of backgrounds, and are not limited to the relationship between experience and post-editing.

Experience and speed

In a relatively small study with six professional translators (three of them were French and three were Spanish), de Almeida and O'Brien (2010) found that the translators with the most professional experience (expressed in number of years) were also the fastest post-editors. Regarding regular human translation, more experienced translators have been found to translate faster than less experienced translators, while making more translation decisions (i.e., situations where a choice is made between different ways to carry on the translation process) at the same time (Tirkkonen-Condit, 1990), but not necessarily always so (Guerberof, 2012; Jääskeläinen, 1996; Kiraly, 1995). It must be noted that even though Tirkkonen-Condit (1990) uses the word 'professional', she was working with translation students in their first and fifth years. Professionalism in this context has to be seen as 'level of experience' rather than actual professional translation experience. Jääskeläinen (1996) further suggest that time is related more to success of the process (with the more successful translators requiring more time) than to experience.

Experience and cognitive load

Sweller (1988) studied the cognitive load during problem solving and related it to expertise. He established that experts use so-called schemas to solve problems: a cognitive structure that allows them to identify and classify problems so that the required steps to solve them can easily be identified. Novices, however, do not have the knowledge required to classify problems as belonging to comparable categories, and so they have to resort to general problem-solving strategies, also known as means-end strategies. Using a computational model (computer models capable of testing different strategies depending on a problem type), these latter strategies were found to require more cognitive effort than the strategies adopted by experts (Sweller, 1988).

In addition to the psychological research related to cognitive load and working memory, the cognitive aspects of translation have been studied by translation process research as well. Dragsted (2006), for example, investigated the translation process of translators (six professionals and six students) working with a translation memory system. She found that translators usually work with clauses or phrases as translation units (presumably the maximum number of elements that can be contained in a translator's working memory), but that working with a TM forces a translator to work with a whole sentence as the translation unit. This was seen as a disadvantage by the professionals and as an advantage by the students. According to Dragsted (2006), this can be explained by the way professional translators cognitively deal with translation: they can handle larger chunks of information, taking the source text as a whole into account, and they are more aware of the potential problems that arise from sentence segmentation. Students have also been shown to require more fixations and more pauses than professional translators while translating (Dragsted, 2010), elements that are normally used as indicators of increased cognitive effort.

Experience and external resources

In general, professional translators are more confident about their own competences than student translators, resulting in an overall lower consultation of external resources (Prassl, 2010). They are said to show a greater tolerance towards ambiguity and uncertainty in the source text (Fraser, 2000), and to rely on dictionaries less than students (Fraser, 2000; A. Jensen, 1999; Tirkkonen-Condit, 1990). On the other hand, they are more aware of serious translation problems and tend to require more steps to solve these problems, by using various types of external resources (Hunziker Heeb, 2012; Jääskeläinen, 1990). Jääskeläinen (1990) found that more experienced students preferred monolingual sources, whereas less experienced students preferred bilingual sources. It is unsure whether this extrapolates to professional translators and other language combinations as, for example, Gough found professional translators to consult bilingual dictionaries more than five times more frequently than monolingual dictionaries (J.

Gough, personal communication, April 2016). Kiraly (1995), however, found no differences in the usage of bilingual and monolingual dictionaries between novice and professional translators.

The fact that translation trainees benefit more from post-editing than professional translators (Garcia, 2011) could be an indication that professional translators are more insecure about MT output quality, which could lead to a higher number of consulted resources in the post-editing condition, and which could, in turn, negatively affect productivity.

1.4.2 Experience and product

More and less experienced translators treat the translation process differently, although the particular differences are hard to establish. Translators with a higher degree of professionalism have been said to make more translation decisions (Tirkkonen-Condit, 1990) and to monitor the task on a higher level, taking aspects such as coherence, structure, and register into account, whereas novice translators are said to treat translation as a linguistic task (Séguinot, 1991; Tirkkonen-Condit, 1990). A. Jensen (1999), however, differed from Tirkkonen-Condit regarding problem solving: experienced translators were found to perform fewer editing events, and fewer problem-solving activities, not more.

Regarding the final quality, Kiraly (1995), established that there was no clear difference between a target text produced by professional and non-professional subjects. He suggested that 'translator confidence' could be a more important factor than actual translation experience, as previously proposed by Laukkanen (1993) in an unpublished study referred to by Jääskeläinen (1996). Jääskeläinen (1996) compared two studies, the first conducted by Gerloff in 1988, the second by herself in 1990, and came to conclusions similar to Kiraly's (1995): she found that professional translators did not necessarily perform better than novice translators.

1.4.3 Experience and attitude

Weaker students appreciated post-editing more than strong students, and found it less stressful than regular translation, whereas stronger students found MT output to contain too many dumb errors (Kliffer, 2005). Studies explicitly comparing the attitudes of professional translators and students are rare.

Moorkens and O'Brien (2015) found students to be somewhat more positive towards post-editing than professional translators. Post-editing was disliked because it was a tedious task, it was believed to be more time-consuming, or the MT had poor quality.

Carl, Gutermuth, and Hansen-Schirra (2015) compared students and professional translators who performed a regular translation task, a post-editing task, and a monolingual post-editing task (i.e., post-editing without being given the source text). After the experiment, 54.5% of the professional participants indicated they were 'somewhat' or 'highly' satisfied with their post-editing work, compared to 75% of the students. While this seems a positive trend, especially for students, the researchers added that 83% of the participants indicated that they would rather translate from scratch than post-edit. Participants felt that a lot had to be post-edited (approaching 100% for the professionals, and 75% for the students), but they still preferred regular post-editing over monolingual post-editing. When asked about MT quality, most participants rated it 'below average' or 'well below average'. Professionals judged MT output quality more harshly than students, especially when asked about its grammaticality and accuracy.

As there seem to be quite a few differences between more and less experienced translators, we include the factor experience (students vs. professional translators) in most of our analyses. Only the analyses from the pretests and the last chapter are conducted exclusively on students' data, as they were readily available.

1.5 Teaching post-editing

> *...the inclusion of post-editing in an MT course is appropriate for language learners. It can be shown to help their language learning and their translation skills at an appropriate level, and it also helps their awareness of the communicative aspects of language, and gives them a perspective on the use of foreign languages in the workplace.*
>
> *- Judith Belam, "Buying up to falling down": A deductive approach to teaching post-editing*

Several researchers have made a case for the integration of post-editing into the translation curriculum. O'Brien (2002), for example, listed four reasons why post-editing should be taught to translators: (1) to meet the increasing demand for translation; (2) because post-editing skills are different from translation skills; (3) to make future translators more comfortable with post-editing and thus more tolerant; and (4) to improve the uptake of machine translation technology. Both Fulford (2002) and Tatsumi (2010) noted that translators want to learn more about MT and its limitations, which is an additional reason to add post-editing and MT to translator training.

According to O'Brien (2002), the ideal post-editing module consists of a theoretical component and a practical one. During the theoretical component, students would learn about post-editing, machine translation technology, controlled language, terminology management, and programming. For the practical component, students need to be encouraged to practice post-editing as much as possible, both in and outside the classroom, to increase their level of comfort with post-editing. They need to become acquainted with post-editing different text types from different MT systems in different languages. In addition to post-editing practice, O'Brien (2002) suggests that students need to learn how to work with terminology management tools, controlled authoring tools, corpora, and even programming languages. Though O'Brien (2002) acknowledged that her proposed module outline is too extensive for a regular course, she did provide an overview of potential components of post-editing training.

Yet even with limited time available to teachers, post-editing has proved to be a useful addition to translation and language courses. Reporting on a post-editing workshop from a machine-assisted translation course, Belam (2003) described the benefits of introducing students to post-editing. Firstly, the post-editing task was beneficial to language learning, as students had to study the text, learn new vocabulary, and understand the text. Secondly, the task improved students' translation skills, as they had to discuss errors as well as translation strategies, and they learned that MT sometimes provides useful suggestions. Finally, discussing the degrees of post-editing in relationship to text function, content and style trained students' communication skills. Kliffer (2005) taught MT in a university translation course by letting students post-edit. The main benefits according to this study are that students become aware of semantic and functional accuracy between source and target text, that they learn about MT's capabilities and limitations, and that they gain extra translation training by revising MT output. Depraetere (2010) looked at what students intuitively do when post-editing, from which she derived the key post-editing skills students still need to learn, i.e. what they need to be taught. She found that students did not have a tendency to introduce preferential or stylistic changes when post-editing MT output - which is good post-editing practice - yet students sometimes trusted the MT output too much. To help students understand the post-editing process, "[t]here is a distinct need to raise the students' awareness of typical MT errors" (Depraetere, 2010, para. 7).

More recently, Doherty and Kenny (2014) have described the Statistical Machine Translation course given at Dublin City University. In this course, students received hands-on experience with SMT systems and quality evaluation. Overall feedback of the students was positive, and they became more aware of the complexity of machine translation as well as its limitations. The students were in direct contact with the tools' developers, a collaboration providing gains to both academia and industry.

Chapter 2 Hypotheses

Building on the related research, we can now formulate hypotheses to our research questions for the main analysis. As stated, the hypotheses for the more specific analyses from Chapter 6 and Chapter 7 will not be discussed here, but in the relevant chapter.

Is post-editing faster than human translation?

We hypothesise that post-editing is faster than human translation (Guerra Martínez, 2003; Zhechev, 2014), and that professional translators translate faster than students (Tirkkonen-Condit, 1990), although we tentatively assume the beneficial effect of post-editing on translation speed to be greatest for students, as less experienced translators seem to handle translation as a lexical task (Tirkkonen-Condit, 1990), and post-editing provides translators directly with lexical information.

Is post-editing cognitively more demanding than human translation?

We expect post-editing to be cognitively less demanding (O'Brien, 2007) because it provides translators with lexical information and it might help them make decisions in situations where multiple translation options are possible. Again seeing how students treat translation as a lexical task (Tirkkonen-Condit, 1990), we expect post-editing to be especially beneficial for them (Sweller, 1988).

Is the fixation behaviour different for post-editing and human translation?

Laukkanen (1993) found that insecurity leads to heavier reliance on the source text during human translation. As such, we expect students to rely more heavily on the source text than professional translators. We also expect less reliance on the source text when post-editing (Carl, Dragsted, Elming, et al., 2011; Carl et al., 2015).

Are more (or other) external resources consulted in human translation compared to post-editing?

Overall, we expect translators to look up in fewer resources or spend less time in external resources when post-editing compared to translation, since the MT output

should already provide some lexical elements to start from, whereas there is nothing to fall back on during human translation. We expect professional translators to consult fewer resources than students (Prassl, 2010), especially dictionaries (A. Jensen, 1999), although we do expect them to use a wider variety of resources (Hunziker Heeb, 2012).

Is there a difference in overall quality between the product of human translation and the product of post-edited machine translation output?

We expect overall quality to be comparable across methods and both groups of participants (Carl, Dragsted, Elming, et al., 2011; Kiraly, 1995).

Is there a difference in the most common error types in human translations and post-edited texts?

On a more fine-grained level, we expect post-editing to be better than human translation regarding accuracy and mistranslations (Lee & Liao, 2011), but human translation to be better regarding language and consistency (Guerberof, 2009).

We also expect to see differences in the types of errors between both participant groups, as professionals process texts on a higher level than students (Séguinot, 1991), and translation methods. We expect professionals to make fewer content and coherence errors, and we expect the translators who specialise in general text types - the domain under scrutiny in the present paper - to perform better (Jääskeläinen, 2010).

How rewarding is post-editing compared to human translation?

We expect translators to find post-editing less rewarding than human translation (Fulford, 2002), although we expect students to find it somewhat more rewarding than professional translators (Carl et al., 2015; Kliffer, 2005).

How useful is MT output according to translators?

As the system used in this study is not a specifically trained MT system (Green et al., 2013), we expect that participants do not find the MT output all that useful (Koehn, 2009).

Which translation method is perceived as being faster?

We expect participants to perceive post-editing as being more time-consuming than human translation (Gaspari et al., 2014).

How is the final quality of both methods of translation perceived?

It is hard to form hypotheses on the basis of the existing research. Translators have been asked about the quality of MT output, but - to the best of our knowledge - not about their perceived quality of the final product. Translators did say they are "more

prone to mistakes" when post-editing and that they found it hard to balance time and quality (Moorkens & O'Brien, 2015, p. 79), which could be an indication of lower expected final quality when post-editing compared to human translation. Here as well, we expect students to be somewhat more positive, as they regard the process as just "improv[ing] a few things" (Moorkens & O'Brien, 2015, p. 79).

Which translation method is the most preferred translation method?

We expect participants, especially professional translators to perceive post-editing as more effortful (Dragsted, 2006; Guerberof, 2013), and to prefer human translation (Carl et al., 2015; Gaspari et al., 2014). We expect students to have a more neutral or positive attitude towards post-editing (Moorkens & O'Brien, 2015).

Is there a difference in perception before and after the experiment?

We tentatively assume participants to be somewhat more positive towards post-editing after the experiment than before. There is not much to go by in the existing literature, but, for example, Garcia (2010) found that after post-editing a text, a few participants changed their preference in favour of post-editing. It must be said, however, that in Garcia's study, participants were already rather positive towards MT before participating.

Chapter 3 Pretests

As mentioned before, the main goal of this research is to get a better understanding of some of the differences between human translation and post-editing. To achieve this goal, we will look at three different aspects: (i) the translation process, (ii) the quality of the final product, and (iii) the attitude of translators towards machine translation and post-editing.

Before conducting the main experiments, we wanted to test possible experimental setups, and we needed to develop a translation quality assessment approach suitable for the assessment and comparative analysis of machine translation output, post-edited texts, and human translations. Although our focus is on general text types, we wanted to make sure the quality assessment approach could be used by others working on different text types as well. This is why we carried out two pretests, one[1] with general texts, and one with technical texts. In addition, the tests allowed us to get acquainted with a keystroke logging tool and to explore potential analyses for future studies.

Regarding the translation process, we were interested in finding out whether post-editing was indeed faster than human translation for general text types (Carl, Dragsted, Elming, et al., 2011; Garcia, 2011), and whether fewer resources were consulted during post-editing compared to human translation, as the presence of lexical info should aid translators with their uncertainty management (Göpferich, 2010). For the product analysis, we were interested in establishing differences in quality and specific error types between human translation and post-editing, as well as the relationship between machine translation output quality and the final quality after post-editing. The main goal of our attitude analysis was to get a first idea of students' familiarity with machine translation and the way they perceive its usefulness, and the speed and final quality of post-editing compared to human translation.

[1] The first pretest was conducted together with a master student as part of his master's thesis (Tondeleir, 2013).

In the following sections, we will first discuss the experimental setup of both pretests, the translation quality assessment approach developed on the basis of the pretests, and some preliminary process, product, and attitude analyses[2].

3.1 Method

3.1.1 Participants

Sixteen master's students of translation (four male) taking a general translation course participated in the first pretest, seventeen master's students of translation (six male) taking a technical translation course participated in the second test. Participants' mother tongue was Dutch, and they translated from English into Dutch during both courses. Participants had no previous experience with post-editing, and nine students participated in both experiments.

3.1.2 Materials

The corpus for the pretest on general text types consisted of four newspaper articles taken from the 'news articles' text subtype in the *Dutch Parallel Corpus* (Macken, De Clercq, & Paulussen, 2011). To make sure the texts were approximately of equal length (260-288 words), some sentences (additional examples or thoughts at the end of the article) were removed, but only when removing them did not affect the cohesion of the text. We used the *Style & Diction* demo on *editcentral.com*, which computes six of the most commonly used readability metrics, as also used by K. T. H. Jensen (2009), to evaluate the complexity of the four texts. Text complexity can have an impact on translation difficulty (Sun, 2015), and we wished to be able to control for as many factors as possible in our main experiments. Considering text complexity in these pretests allowed us to verify whether some of our findings could be attributed to differences in text complexity. If so, we would be able to perform our text selection for the main experiment on the basis of text complexity alone.

[2] Part of the work described here is published in Daems, Macken, and Vandepitte (2013) and Daems, Macken, and Vandepitte (2014).

According to these complexity metrics[3], two texts (text 1 and text 2) were relatively easy; two texts (text 3 and text 4) were relatively hard. Table 1 gives an overview of the different complexity scores provided by *editcentral.com*. Machine translations were obtained by using *Google Translate*.

Table 1 Complexity scores DPC texts.

	text 1	text 2	text 3	text 4
Flesch reading ease score	54.8	54.9	34.6	40.8
Automated readability index	11.8	14.6	20.1	17.6
Flesch-Kincaid grade level	9.8	12.4	16.5	14.9
Coleman-Liau index	13.5	11.1	14.9	13.4
Gunning fog index	12.2	16.1	21	18.4
SMOG index	11.3	13.2	17	15.1

For the pretest on technical texts, we selected two fragments (*Object Selection* and *Screen Capture*) of approximately 370 words each (368 and 382, respectively) from a manual for *CourseLab*, e-learning software (Websoft). As the main goal of this second experiment was to verify whether our translation quality assessment approach could be used for a different text type, text selection was not as rigorous as for the first pretest. We did select two fragments from the same manual to control for complexity, and we made sure both fragments had a comparable mix of titles and instructions. Machine translations were obtained by using both *Bing Translator* and *Google Translate*. The texts and the machine translation output used in the pretests are available in Appendix 1.

[3] The Flesch reading ease score is inversely proportional (lower scores are assigned to more difficult texts); the other scores are directly proportional (higher scores are assigned to more difficult texts). The different metrics take the following information into account:

Flesch reading ease score: average sentence length and average number of syllables per word
Automated readability index: average word length and average sentence length
Flesch-Kincaid grade level: average sentence length and average number of syllables per word
Coleman-Liau index: number of characters per 100 words and number of sentences per 100 words
Gunning fog index: average sentence length and percentage of complex words

3.1.3 Procedure

The conditions for both pretests were somewhat different, as we were looking at different possible ways of setting up the final experiment.

For the pretest with newspaper articles, participants were divided into four groups. Each group received one post-editing task and one regular translation task from English into Dutch. To assess the impact of text complexity, the first and second group received the easier texts, and participants in the first group post-edited text 2, whereas participants in the second group post-edited text 1. Groups three and four received the harder texts, and participants in group three post-edited text 4, whereas participants in group four post-edited text 3. There was no time restriction for the tasks. Students who finished early post-edited an extra text (text 3 for the first two groups, text 2 for the other two groups). The final number of recorded sessions per method can be seen in Table 2. Data for one participant could not be saved, and another participant accidentally performed the tasks assigned to another group.

Table 2 Number of recorded sessions per method.

	text 1	text 2	text 3	text 4
Human translation	5	4	4	4
Post-editing	3	8	6	4

For the pretest with technical texts, participants first watched a short presentation about the *CourseLab* tool and its interface, to better understand the sections of the manual they had to translate. They then received a post-editing task and a regular translation task from English into Dutch. Students were assigned to one of four groups; the order of tasks (post-editing and human translation) and texts was different in each group. We imposed a time limit of 40 minutes. As it was a technical translation task, students were instructed to adhere to the terminology in the *Microsoft Language Portal* for all terminology issues. The final dataset, after removing a corrupt data file, consisted of four recorded sessions per group. After the experiment, participants filled out a short survey, asking them about their experience with and attitude towards machine translation.

The instruction for both studies was to achieve a translation of publishable quality, and the target audience of the translations was said to be comparable to the target audience of the source text.

We used *PET*, a post-editing tool (Aziz, De Sousa, & Specia, 2012), to record the keystrokes and time for each task. In this tool, the texts to be translated were presented per segment (though previous and upcoming segments were visible as well). Segments

corresponded to sentences for the first pretest, and varied in length (ranging from menu names to section titles to instructions in a bulleted list to full sentences) for the pretest with technical texts. The source text was shown on the left-hand side of the screen, whereas the right-hand side was empty for the human translation task or contained the machine translation output for the post-editing task (Figure 2).

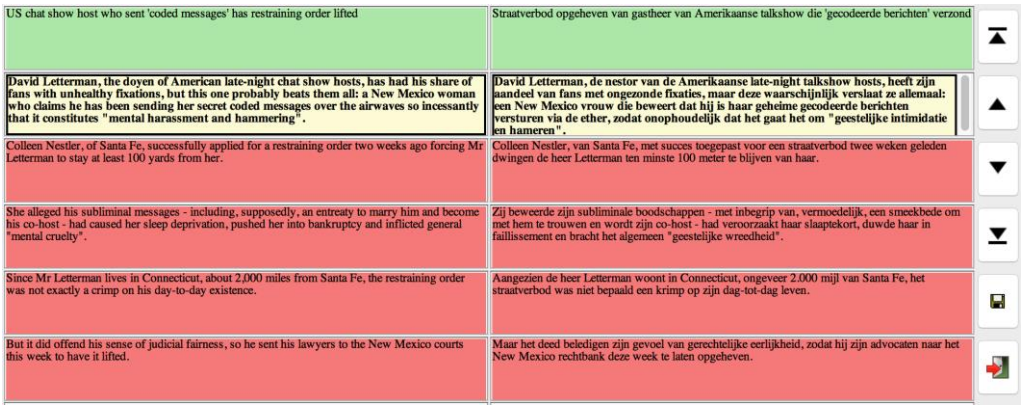


Figure 2 Screenshot of *PET* interface during post-editing task.

For the first pretest, the machine translation output was *Google Translate*'s output; for the second pretest, participants could choose between the output of *Google Translate* or Bing Translator for each sentence (Figure 3). This was one of the extra functionalities in *PET* that we wanted to evaluate, the availability of multiple sources giving students extra flexibility and a possible increased sense of control. Bing Translator was chosen as the second machine translation tool as students were asked to adhere to Microsoft Language Portal and Bing Translator is a Microsoft product as well, so we assumed it would sometimes outperform *Google Translate*.

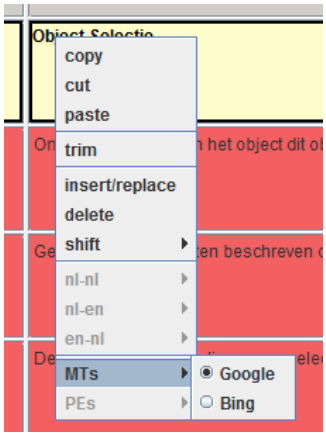


Figure 3 Screenshot of MT selection pane in *PET*.

After each segment in the regular translation task, an evaluation screen popped up, where students had to indicate how difficult they found the translation (very easy, easy, regular, difficult, very difficult) and select the types of issues they encountered (none, terminology/vocabulary, sentence structure, semantic ambiguity, other). They were

asked to list the types of resources they consulted in the comment section. The screen looked a little different for the post-editing task: students had to perform a comparable evaluation, but the judgment of the difficulty of the translation task was replaced by a judgement task of the quality of the MT output (Figure 4). Participants could select the following quality issues: none, untranslated words, bad word order, too literal, lexical issues, grammatical errors, other.

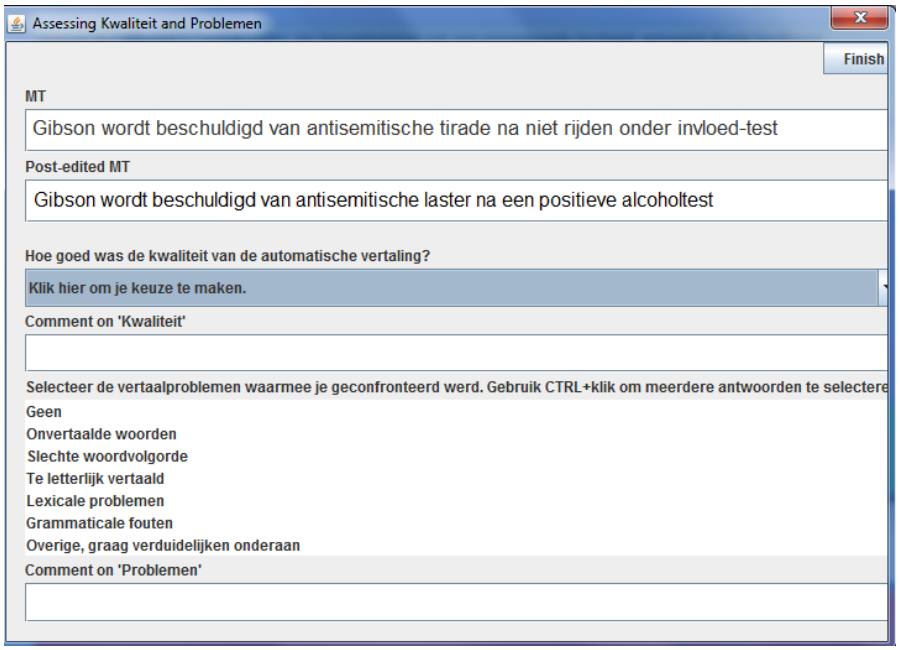


Figure 4 Screenshot of *PET* evaluation screen for post-editing task.

During a part of the first pretest, there were problems with the internet connection, and so the students had to translate and post-edit a few segments without being able to consult external resources. These segments were removed from the analysis of the usage of external resources.

3.2 Translation Quality Assessment Approach

As discussed in Chapter 1, existing translation quality assessment (TQA) approaches often suffer from low inter-rater reliability, lack of flexibility and limited use for diagnostic analyses. For the present investigation, the translation quality assessment approach needed to be suitable for diagnostic and comparative analysis, in order to compare machine translation output with post-edited texts and human translations. It also needed to be flexible and fine-grained, in order to gain a better understanding of specific error types and establish any relation to translation and post-editing effort.

Though the following approach was specifically designed for use within our project, it can be used by anyone interested in comparative translation quality analysis, the identification of specific problems for different translation methods, the improvement of post-editor training or machine translation environments, and the development of quality estimation systems based on human annotations (Wisniewski et al., 2013). The approach most suitable for the type of fine-grained analysis we needed, was an error typology. And, since our approach had to be applicable to both MT and human translations, we combined the dichotomy of adequacy and fluency issues often found for MT evaluation with the granularity of error analysis. An MT text is judged as being adequate when the meaning of the source text is preserved in the target text. It is considered fluent when it is grammatical and well-formed from a target-language perspective.

3.2.1 Classification

The concepts of adequacy and fluency correspond closely to the notions of adequacy and acceptability as defined by Toury (1995). A translation that fulfils the norms of adequacy is a translation that is true to the source (con)text and audience, whereas a translation that fulfils the norms of acceptability is a translation that is true to the target (con)text and audience. While a combination or compromise of both appear in almost any translation, the concepts should be regarded as having distinct theoretical statuses (Toury, 1995). When constructing our error typology, we therefore wanted to keep this clear dichotomy between both aspects. We looked at a few existing error typologies for inspiration, and soon found that, while elements of adequacy and acceptability can be found in most typologies, they are not always treated as distinct concepts. In the commonly used *LISA QA*, for example, the *Doc Language* category covers adequacy issues such as mistranslations, as well as acceptability issues such as terminology, style, and language. The same holds true for the *SAE J2450 QA*, which has no main categories and combines acceptability elements such as syntactic errors, word structure or agreement errors, and misspelling, with adequacy elements such as omissions. Two error typologies that do make comparable distinctions are the one developed by *MeLLANGE* (2006) and *QTLaunchPad*'s *MQM* (2013). The former makes the distinction between 'content-related' and 'language-related' errors, the latter between 'accuracy' and 'fluency'. Not all content issues are necessarily discrepancies between source and target text, and the *MeLLANGE* error typology contains references to issues caused by source language as well as target language in both main categories. *MQM*'s metric is closer to the type of categorisation we wished to develop, but, as *MQM* did not yet exist when we created our error typology, we could not use it as a starting point. It is interesting to see that, though developed independently, the general categories seem to

be comparable across typologies, with our adequacy issues corresponding closely to *MQM*'s accuracy issues, and our acceptability issues corresponding closely to *MQM*'s fluency issues. The main difference lies in the fact that the *MQM* metric requires annotators to choose between accuracy and fluency errors, whereas our translation quality assessment approach allows for more than one error type to be assigned to the same error, as the same error can have an impact on adequacy as well as acceptability. For example, a word sense disambiguation error (= adequacy issue) can lead to a logical problem when considering the target sentence in isolation (= acceptability issue). In addition, *MQM* does not have a category for lexical issues (e.g., awkward collocations), whereas these issues frequently occurred in our pretests, leading us to create a separate category for them.

In our TQA approach, the acceptability errors[4] are further subdivided into grammar and syntax, lexical issues, spelling and typos, style and register, and coherence issues. An overview of all subcategories can be found below. Taking into account criticism on previous error-based methodologies, we included extra categories to make sure the text as a whole and the text in context could be assessed during the evaluation: text type specific errors and coherence issues.

Overview of acceptability subcategories

grammar and syntax

- article
- comparative/superlative
- singular/plural
- verb form
- article-noun agreement
- noun-adjective agreement
- subject-verb agreement
- reference
- missing constituent/preposition
- superfluous word/constituent
- word order
- structure

[4] The error categories are *possible* error categories, in the sense that the categories are not necessarily erroneous for each text type. Depending on the goal of the assessment, different categories may or may not be regarded as errors. This can be expressed by assigning a zero weight to the category. The term 'error categories' is used throughout the text with the same intended meaning.

- other

lexicon

- wrong preposition
- wrong collocation
- named entity
- word non-existent

spelling and typos

- capitalization
- single-word spelling mistake
- compound
- punctuation
- typo

style and register

- register
- untranslated
- repetition
- disfluent sentence/construction
- short sentences
- long sentence
- text type
- other

coherence

- conjunction
- missing information
- logical problem
- paragraph
- inconsistency
- other

The adequacy category contains those errors that can only be identified when juxtaposing source and target text: differences in meaning, additions, deletions, explicitations, and terminology issues. The complete overview can be consulted below.

Overview of adequacy subcategories

- contradiction
- word sense disambiguation
- part of speech
- hyponymy
- hyperonymy
- terminology
- quantity
- time
- meaning shift caused by punctuation
- meaning shift caused by incorrect translation of function word
- meaning shift caused by misplaced word
- deletion
- addition
- explicitation
- coherence
- inconsisent terminology
- other meaning shift

A detailed overview of each subcategory with examples and guidelines can be found in Appendix 2, as well as online[5]. We tested preliminary categorisations on the output of the first pretest, and further fine-tuned the categorisation on the basis of the output of the second pretest, for example, by adding terminological issues. The proposed categorisation is the result of our fine-tuning efforts, still it does not claim to be exhaustive. It is primarily designed for the analysis of translations from English into

[5] http://users.ugent.be/~jvdaems/TQA_guidelines_2.0.html

Dutch, but this does not mean that the approach has a limited use. Depending on the goal of the assessment, more specific subcategories can be defined: languages requiring cases, such as Russian, can be analysed by adding a subcategory 'incorrect case' to the category of grammar and syntax. Likewise, complex verb morphology can be included so as to account for incorrect pronoun suffixes for imperative verbs in Italian. Just as we slightly altered the categorisation to analyse a different text type (technical texts), the hierarchical design of the proposed categorisation allows for language pair-specific customization, as has recently been done successfully by Fomicheva (2015) in her bachelor's thesis, comparing machine translations made with *Google Translate* from Swedish into Russian.

3.2.2 Annotation

When developing our TQA approach, we wanted to reduce the subjectivity often found in scoring methods (Wisniewski et al., 2013), and to facilitate the choice between acceptability and adequacy issues (Stymne & Ahrenberg, 2012). Rather than letting evaluators mark texts for both acceptability and adequacy at the same time, the annotation process was split up into two steps. In a first step, evaluators received the full target text and they marked translations for acceptability only. In a second step, evaluators received both source text and target text and they marked translations for adequacy only. The two-step approach is similar to the ones required in EN 15038, the European Standard for translation companies, which recommends two distinct phases: an error analysis for errors relating to acceptability (where the target text as a whole is taken into account, as well as the target text in context), and one for errors relating to adequacy (where source segments are compared to target segments). This approach had a few advantages: the assessment was no longer limited to the sentence level, there was less ambiguity between the different categories, and the approach was easier to use, thus effectively solving problems other evaluation methods faced (Vilar et al., 2006)

In addition, the annotator's task was limited to highlighting and identifying problems. It was not the annotator's task to assign severity weights to errors. Each error category was given an error weight beforehand, and the weights can be altered to suit different assessment goals or text types, as suggested by the TAUS error typology guidelines (2013b). As such, the error typology becomes more flexible than the 'one-size-fits-all' approaches.

Error weights ranged from 0 to 4, depending on the impact of the error. A weight of 0 was given to neutral problems: items that were either not problematic for the task at hand (such as cases of explicitation in general texts), or items that were not the translator's fault (such as typos that remained undetected due to a problem with the spell-checker of the translation tool). Minor problems received a weight of 1. These

problems were errors that did not affect the readability or the understanding of the text, such as capitalisation errors or misspelled compound nouns. Though the TAUS guidelines suggested limiting error weights to four severity levels (neutral, minor, major, and critical), this did not give an accurate representation of an error's impact when we tested it. With four severity levels, both incorrect prepositions and wrong collocations would have received the same weight, even though prepositions have a lower impact on readability and understanding than collocations. This is why we split TAUS' 'major' category up into 'medium' errors, receiving a weight of 2, and 'major' errors, receiving a weight of 3. The 'critical' category, with weight 4, was reserved for errors that dramatically affected the content or understanding of a text, such as contradictions or word sense errors, or for non-adherence to explicit guidelines, such as using a specified terminology for technical texts.

To facilitate the annotation process even further, we used the brat rapid annotation tool (Stenetorp et al., 2012). This tool allows users to add their own annotation schemes and texts. It provides a nice interface and user-friendly environment. Because errors can be highlighted and annotated directly in the text, we believe the tool is both faster and more accurate than other methodologies, such as letting reviewers highlight errors in a Word document and then transfer the errors to an Excel sheet (Guerberof, 2012). After highlighting a word or sentence, a pop-up screen appears where the reviewer can select the correct category. The tool allows for a hierarchical categorisation if needed. The pop-up screen also contains a notes' section, which allows reviewers to explain why they consider something to be an error or why they assigned a certain category to that item.

The main advantage of this approach is the fact that issues are identified and mapped to the text in the same step. Annotations are recorded with indices of the span (a range of numbers that correspond to the position of characters in the text), facilitating the analysis afterwards. During the pretests, two separate files were annotated for each task, one for acceptability, one for adequacy. An example of an annotated sentence from the acceptability stage can be seen in Figure 5. Figure 6 and Figure 7 show examples from the adequacy stage. Sentences were annotated by making use of the brat web-based annotation tool. Since reviewers received both target and source sentences for the latter activity, errors that are usually hard to map – such as deletions – could be highlighted directly in the source text. The tool could also be configured to allow for relationships between different words, which was useful for highlighting problem words that were split up by non-relevant words (Figure 6).

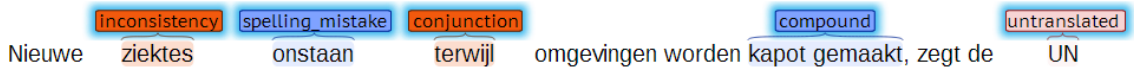


Figure 5 Example of an annotated sentence for the acceptability task.[6]

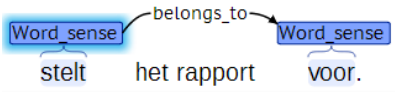


Figure 6 Example of a split-up adequacy annotation ('voorstellen' is a separable verb in Dutch).[7]

Figure 7 Example of an adequacy annotation in a source text segment.

This method still required us to merge the files afterwards. A later version of the tool allowed us to perform the acceptability and adequacy analysis on the same files. The updated brat tool had a function that allowed us to switch off visibility for previously made annotations. This function was created by a master's student as part of their thesis (Naert, 2013) and is not a part of the standard installation of brat. Because of the function, adequacy issues could be marked on the same text without the acceptability issues being visible. Afterwards, both annotation types were made visible so that relationships between both could be highlighted (Figure 8). This facilitated the processing and scoring of annotations afterwards: when looking at adequacy and acceptability in isolation, the error weight for both errors was counted, but when looking at overall quality, only the adequacy error weight was counted whenever the acceptability error was caused by the adequacy error (as was often the case for word sense issues and logical problems).

Figure 8 Example of an acceptability and adequacy annotation in the same file.[8]

[6] translation: New diseases arise while environments are destroyed, says UN.

[7] translation: the report suggests

[8] translation: The show is billed as the museum's largest ever.

3.2.3 Testing the TQA approach

One of the disadvantages of using translation quality assessment methods is low inter-annotator reliability. In order to establish the validity of our approach, two aspects had to be investigated: whether or not annotators highlighted the same passages, and whether or not they labelled the items with the same error category. Stymne & Ahrenberg (2012) provided an overview of inter-annotator agreement research in the field of machine translation evaluation. The researchers could find some inter-annotator agreement scores for fluency, adequacy and ranking assessment (0.25, 0.23, and 0.37 kappa figures, respectively), but no data on inter-annotator agreement for error analysis. Their own error analysis experiment led to agreement from 27% for less detailed error analysis to 77% for more detailed error analysis with severity scores. In a second phase of Stymne & Ahrenberg's analysis, after guidelines had been created, agreement rose from 40% to 80%. In their study, the most difficult levels of annotation seemed to be determining the distinction between fluency and adequacy errors and the determination of the severity level of each error, two annotation issues that annotators no longer have to concern themselves with in our approach. While we still distinguish between adequacy and acceptability, the annotators' task is made easier by the fact that both error types are dealt with in a separate step, and the fact that annotators do not have access to the source text when annotating for acceptability.

Our annotators were two translation and language specialists, one with a Master's degree in Translation (English-Dutch), and one with a Master's degree in English and Dutch Linguistics. Inter-annotator agreement was calculated on the annotated corpus of translations from the above-mentioned pretests. The annotators were told to adhere to the online guidelines (see Appendix 2). For the first study, we had four source texts, for which we collected 16 human-translated texts (HT), 22 post-edited texts (PE) and 4 machine-translated texts (MT) in total. For the second experiment, we had two source texts, for which we collected 17 human-translated texts, 17 post-edited texts and 4 machine-translated texts (as we had two different machine translation systems) in total.

Based on the annotations and the comments provided by the annotators, we could identify the errors that were highlighted by both annotators or only by one annotator. Table 3 provides an overview of the agreement scores in different stages. Initially, only 341 of the 796 acceptability annotations made on the HT and PE texts of the first experiment were made by both annotators, leading to 39% agreement and a corresponding kappa score of 0.32. For adequacy, only 134 of the 291 cases were

highlighted by both annotators, leading to 42% agreement and a kappa score of 0.31[9]. This seems to be comparable to the numbers obtained by Stymne & Ahrenberg (2012), where only 186 of 473 errors were highlighted by both annotators. For the second experiment, initial agreement was slightly higher for acceptability, where 574 of 1155 annotations were made by both annotators, leading to an agreement of 50% and a kappa score of 0.44. For adequacy, annotators initially agreed on 227 of 497 annotations, which led to an agreement of 46% and a corresponding kappa score of 0.30.

Following the suggestion by Stymne and Ahrenberg (2012) that inter-annotator agreement could benefit from joint discussion of examples, we introduced a consolidation phase during which the evaluators discussed each other's annotations to see whether or not they agreed. The comments from the note section provided extra information which helped the annotators understand each other's motives. Many errors were only highlighted by one annotator because the other annotator had not observed the error, not because she did not agree. There were also errors that recurred in different translations, so if the annotators disagreed on one conceptual item, this could result in a larger difference when looking at the total number of items. After the consolidation phase, agreement was much higher: acceptability agreement went up to 67% (κ=0.65) and 81% (κ=0.80) for the first and second experiments respectively, while adequacy agreement went up to 82% (κ=0.79) and 94% (κ=0.92) for the first and second experiments respectively.

Further, it was hypothesised that a lenient annotator would be lenient across all texts, whereas a stricter annotator would be equally strict across all texts. This hypothesis was confirmed by the significant correlation found between both annotators after fitting a linear regression[10]: $r = 0.67$, $n = 38$, $p < 0.001$ and $r = 0.95$, $n = 34$, $p < 0.001$ for the acceptability scores of the first and second experiments respectively, and $r = 0.87$, $n = 38$, $p < 0.001$ and $r = 0.86$, $n = 34$, $p < 0.001$ for the adequacy scores of the first and second experiments respectively. Moreover, when looking at the items that were highlighted by both annotators before the consolidation phase, it seemed that agreement on the categories was quite high (between 83% and 90%), indicating that the categorisation as explained in the online guidelines was clear. The scores for the MT annotations were slightly higher, since the MT annotation was carried out after the HT and PE annotation and consolidation, so the annotators were already more acquainted with the methodology.

[9] Kappa is influenced by the number of categories and takes chance agreement into account, which is why an increased agreement does not necessarily correspond to an increase in kappa score, and there are differences between acceptability and adequacy kappa scores.

[10] This could not be done for the MT-annotations due to the small number of data points (i.e., only 4 machine-translated texts per experiment).

Table 3 Comparison of initial inter-annotator agreement, agreement after consolidation phase, correlation between annotators, and agreement on categories before consolidation. Scores are given for acceptability and adequacy, for human translation and post-editing together as well as machine translation.

	HT&PE acceptability		HT&PE adequacy		MT acceptability		MT adequacy	
	Exp1	Exp2	Exp1	Exp2	Exp1	Exp2	Exp1	Exp2
Initial agreement	39% (κ=0.32)	50% (κ=0.44)	42% (κ=0.31)	46% (κ=0.30)	53% (κ=0.49)	79% (κ=0.77)	57% (κ=0.46)	51% (κ=0.41)
Agreement after consolidation	67% (κ=0.65)	81% (κ=0.80)	82% (κ=0.79)	94% (κ=0.92)	84% (κ=0.83)	95% (κ=0.94)	94% (κ=0.92)	86% (κ=0.83)
Correlation between annotators	r=0.67, n=38, p<0.001	r=0.95, n=34, p<0.001	r=0.87, n=38, p<0.001	r=0.86, n=34, p<0.001	n/a	n/a	n/a	n/a
Agreement on categories	90% (κ=0.89)	89% (κ=0.88)	89% (κ=0.87)	88% (κ=0.83)	83% (κ=0.81)	93% (κ=0.93)	86% (κ=0.79)	86% (κ=0.82)

After the consolidation phase, the annotations that both annotators agreed on were used to quantify and analyse the errors that occurred. The results of this analysis will be discussed in section 3.3 Results.

3.2.4 Error sets

The error analysis allowed us to identify the main problem categories for each method of translation, or for each translator. By calculating the average error score per word, the overall quality of different translations could be compared, as well as the acceptability or adequacy quality. However, in order to better compare the different translation methods, we wanted to gather information about specific source text passages as well. This would enable us to identify passages that were problematic for more than one post-editor or translator, passages that were problematic in the MT output but no longer in the post-edited version, and passages that were problematic for human translators but not for post-editors, or vice versa. To be able to enrich our data with this information, the concept of source text-related error sets was introduced.

A source text-related error set consists of a source text passage and the corresponding passages that reveal errors (both adequacy and acceptability errors)

made by MT, post-editors or human translators. The following example illustrates the idea of source text-related error sets:

(1) ST: Changes in the environment that are sweeping the planet...
MT: Veranderingen in de omgeving die het vegen van de planeet tot stand brengen... (wrong word sense) *"Changes in the environment that bring about the brushing of the planet..."*
HT1: Veranderingen in de omgeving die het evenwicht op de planeet verstoren... (other type of meaning shift) *"Changes in the environment that disturb the balance on the planet..."*
PE1: Veranderingen in de omgeving die over de planeet rasen... (wrong collocation + spelling mistake) *"Changes in the environment that raige over the planet..."*

The word 'sweeping' was mistranslated by *Google Translate* (MT), one translator (HT1) and one post-editor (PE1). Though the translations and types of errors are different, the errors are all related to the same ST passage.

For each text in the pretests, we manually grouped all errors together with their corresponding source text passages. In order to quantify error sets, each translation method received a uniform weight (1) per ST-passage. This weight was divided over the different translators that could have made the error. For example, if a text was translated by two different MT-systems, post-edited by two different post-editors and translated by four human translators, the MT-systems and post-editors would each have a weight of 0.5, whereas each human translator would have a weight of 0.25 (see Figure 9 for a visualisation of this principle). If, for example, three human translators made an error within the specified ST-passage, it could be said that the passage was problematic for 75% of human translators. As such, we could identify the passages that were more problematic for one type of translation than for another. We could also identify the passages that were problematic for MT, but that did not show up after post-editing, or those passages that were still problematic in the post-edited texts. This allowed us to draw conclusions on the types of passages that post-editors could handle, or the types of translation difficulties that post-editors or MT-systems needed to be trained on.

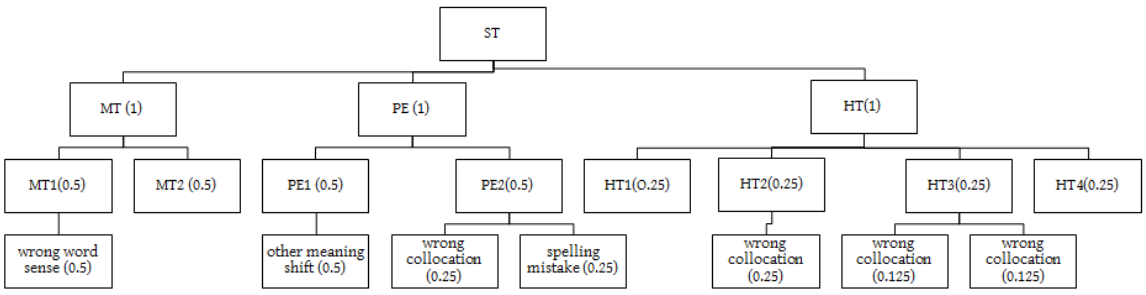


Figure 9 Example of the quantification within a source text-related error set. Each translation method (MT, PE and HT) receives a uniform weight of 1 per ST-passage. The weight is equally divided over all possible translators for that translation method, and the corresponding error categories.

Identifying the difficulty level of each passage was one step; a second step was identifying the main problem categories for each passage. This was done by looking at the problem categories for each translator within that passage. The translator's weight was divided over the different categories. Figure 9 shows a visualisation of this division. Taking the above example as a starting point, this would mean that the 0.5 weight of MT1 would go to the category 'wrong word sense'. The 0.5 weight of PE1 would go to 'other meaning shift' and the 0.5 weight of PE2 would be evenly divided over 'wrong collocation' and 'spelling mistake'. If two of the four human translators also made an error within that passage, their weight of 0.25 would be divided over the corresponding error categories. If the same ST-passage occurred twice in the same text and HT3 made the same error twice, for example a 'wrong collocation' error, then each 'wrong collocation' error received a weight of 0.125 for that translator. This method ensured that results did not get skewed when only one translator made the same mistake over and over again since each translator made up an equal part of the total weight. Afterwards, the totals were determined for each translation method. The weights thus obtained are called 'proportional weights'. For the example ST-passage, this resulted in the following proportional weights for each error category per translation method:

(2) MT = 0.5 wrong word sense
PE = 0.5 other meaning shift + 0.25 wrong collocation + 0.25 spelling mistake
HT = 0.5 wrong collocation

By summing up the totals per error category over all ST-passages in the text, we got an overview of the most problematic error types for each translation method. This was a more balanced overview than the overview that was purely based on the overall number of errors in a text. Another advantage of defining source text-related error sets was the fact that we could analyse specific subsets. For example, we could filter our data to only contain elements that were problematic for MT, but not for any other translation method. Or we could focus on those passages that were problematic for MT, but not for post-editing; or the passages that were only problematic for humans (PE and HT), but not for MT systems; or those passages that were problematic for all three translation methods; etc. See Figure 10 for a representation of the different possible intersections.

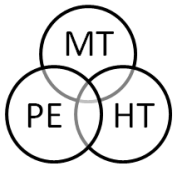


Figure 10 Possible intersections of interest for comparative analysis after identifying source text-related error sets.

3.3 Results

3.3.1 Process

Regarding the translation process, we were mainly interested in determining whether or not a productivity gain could be found for general text types (newspaper articles) when post-editing machine translation output. The results for the first pretest can be found in Table 4. This indeed seems to confirm that post-editing can lead to faster translations, even when using *Google Translate* on general text types. There do seem to be strong differences between the different texts, differences that, at first sight, cannot be explained by the above-mentioned complexity indicators.

Table 4 Average time (in seconds) per source text token, and the productivity gain.

	HT	PE	Productivity gain
Text 1	8.4295	6.5723	22.03 %
Text 2	8.4752	5.1836	38.84 %
Text 3	8.0855	4.9993	38.17 %
Text 4	7.6988	7.4034	3.84 %

To statistically verify this relationship, we fit a linear mixed effects model[11] in *R* with the average time in seconds per source text token as dependent variable, translation method (HT/PE) as independent variable, and participant as random effect (to account for individual differences across participants). We tested this model against a null model without independent variable and found a significant improvement in model fit ($p<0.001$, AIC reduction from 7236 to 7223). The model summary showed post-editing to have a significantly lower average time per source text token than human translation (2414 ms, standard error = 604, $p<0.001$). The effect plot can be seen in Figure 11. Surprisingly, adding the text number as a possible predictor to the model did not improve it (AIC 7231, $p = 0.74$).

[11] Linear mixed effects models are explained in more detail in 5.1 Process analysis.

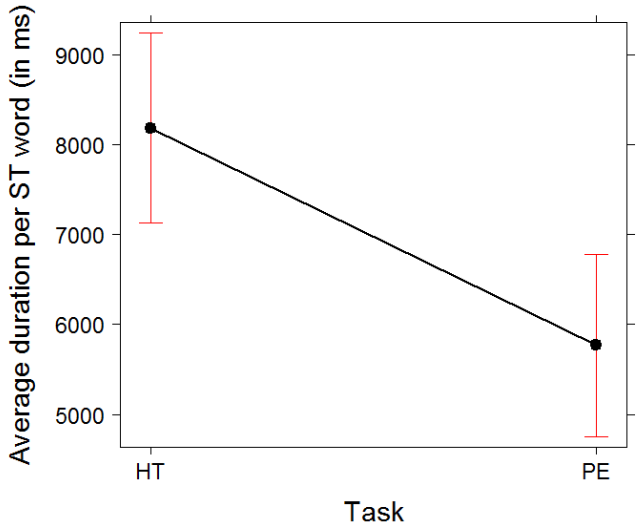


Figure 11 Effect plot of the impact of Task (= translation method) on the average duration per ST token (in ms).

In addition to speed, we were also interested in the usage of external resources for both tasks. As *PET* only recorded what happened inside the tool itself, participants were asked to list the external resources they used after each translated segment via a pop-up screen. Three options were predefined: monolingual dictionaries, bilingual dictionaries, and *Google Search*es. We also added an option 'other', where participants could add other types of resources they used.

After filtering out the data for the segments during which participants lost internet access, there was data for 339 segments in 36 sessions (16 HT, 20 PE). For each session, we counted the number of sentences without external resources, the number of monolingual dictionaries used, the number of bilingual dictionaries used, and the number of times *Google Search* was used. These values were divided by the number of sentences in the session (as we did not always have reliable data for an entire session, this was done to make the numbers more comparable). We then performed two-sample t-tests to determine whether there was a statistically significant difference in mean between human translation and post-editing for each of these variables. This was indeed the case for the average number of sentences where no resources were consulted and the number of bilingual dictionaries, and almost for the average total number of external resources. The results of the t-tests are summarised in Table 5.

Table 5 Summary of two-sample t-tests analysing differences in mean between HT and PE for the usage of external resources.

	HT: μ(σ²)	PE: μ(σ²)	df	t	p
Sentences without external resources	0.45 (0.02)	0.63 (0.07)	31	2.62	0.01
Avg # bilingual dictionaries	0.38 (0.05)	0.15 (0.02)	23	3.76	0.001
Avg # monolingual dictionaries	0.12 (0.05)	0.07 (0.02)	25	0.81	0.43
Avg # *Google Searches*	0.13 (0.03)	0.17 (0.04)	34	0.49	0.63
Avg # other sources	0.09 (0.02)	0.09 (0.05)	32	0.05	0.96
Avg total # external resources	0.72 (0.08)	0.47 (0.19)	33	2.03	0.051

These findings show that there are significantly more sentences for which students do not consult external resources during post-editing than during human translation. Most types of external resources seem to be consulted with a comparable frequency across task types, with the exception of bilingual dictionaries, which are used significantly more often during regular translation than during post-editing.

3.3.2 Product

The hierarchical structure of the translation quality assessment approach allows us to analyse translation quality for different levels of granularity. The following analyses are mostly exploratory, as they were a way of comparing different possible angles for our later experiments.

The topmost level distinguishes between acceptability and adequacy issues. The average error weight per word for all pretests, and for all three translation methods (human translation, machine translation, and post-edited machine translation) can be seen in Figure 12 (acceptability) and Figure 13 (adequacy).

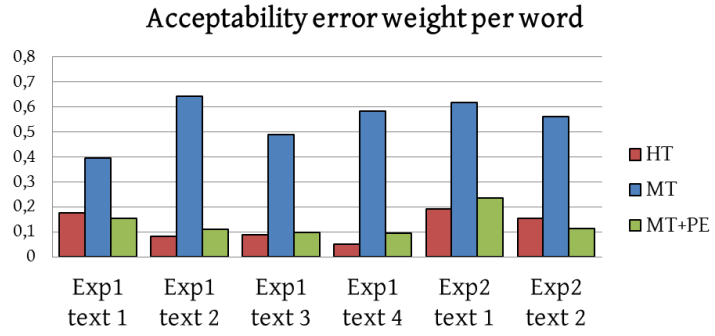


Figure 12 Acceptability error weight per word for all pretests and translation methods.

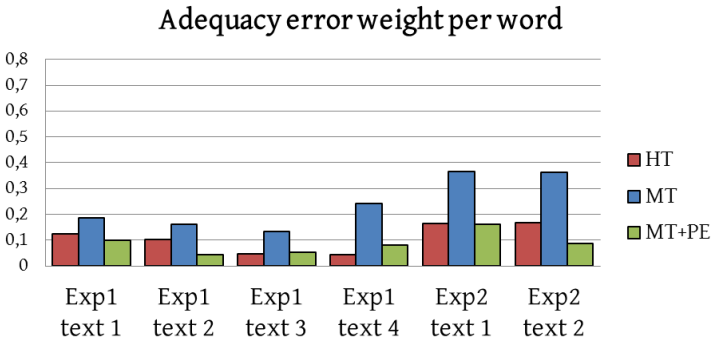


Figure 13 Adequacy error weight per word for all pretests and translation methods.

These graphs show that all three translation methods (HT, MT, and PE) score better for adequacy than for acceptability errors, though this difference is smallest for HT (with only four out of six texts where this is the case). MT scores worse on adequacy for technical texts than for general texts, which is probably caused by the abundance of terminology issues. Students' post-editing leads to a marked increase in quality when compared to the initial MT quality. The difference is most notable for acceptability errors, with error reductions of up to 83%. For four out of six texts, post-editing even leads to better adequacy results than manual student translation, whereas this is only the case for two out of six texts when looking at acceptability. It can also be noted that there is a large difference between the different texts.

On a more fine-grained level, acceptability was further subdivided into grammar, lexicon, coherence, style & register, and spelling & typos, whereas adequacy has no further subdivision on this level. Figure 14 shows the proportion of each error category for each translation method. As not all students managed to translate the entire text for the second experiment, we limited this analysis to the data of the general translation study (newspaper articles).

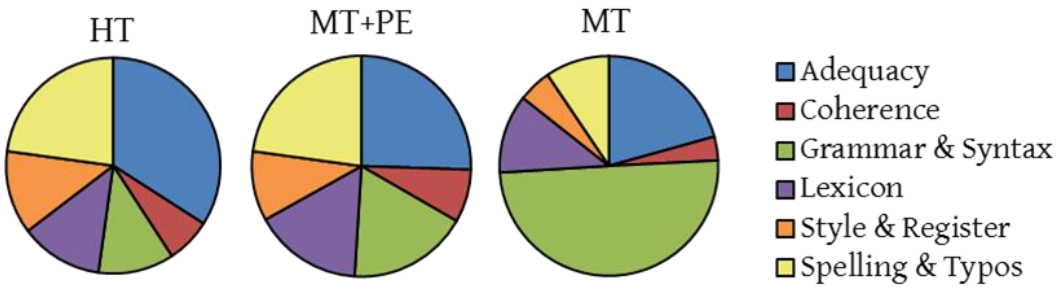


Figure 14 Proportion of main error categories in newspaper articles for all translation methods.

Around 50% of all MT errors consist of grammatical errors. PE seems to suffer most from adequacy errors, grammatical errors and spelling mistakes, whereas adequacy errors are clearly the most common error type for HT.

For the third and most fine-grained level of error analysis, a proportional representation makes less sense, as there are so many subcategories. Figure 15 and Figure 16 show the most common human translation problems (those accounting for at least 5% of all errors) and problems after post-editing for the pretest with newspaper articles.

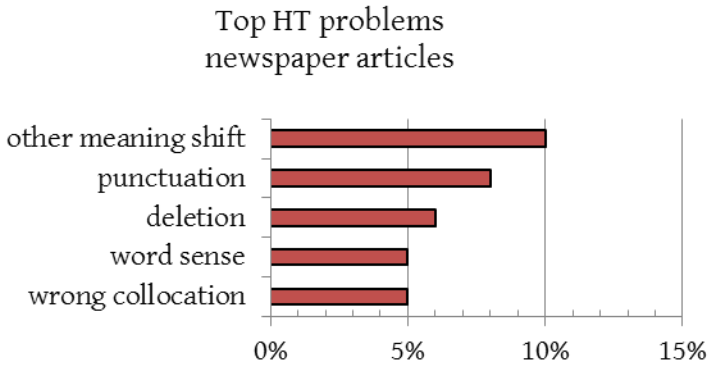


Figure 15 Overview of HT errors accounting for at least 5% of all errors made during the newspaper article study.

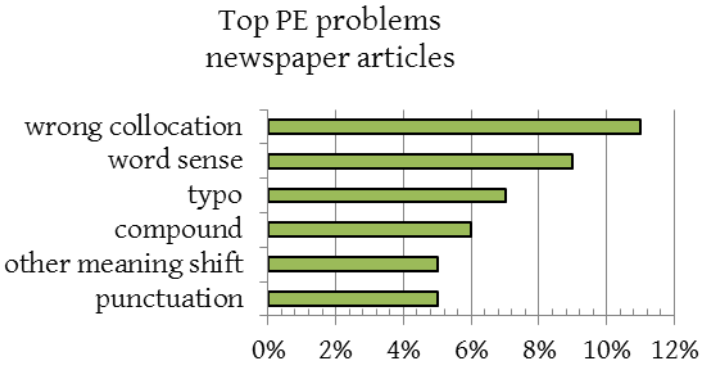


Figure 16 Overview of PE errors accounting for at least 5% of all errors made during the newspaper article study.

Human translations and post-edited texts have four frequent problem categories in common (meaning shifts, punctuation, word sense problems, and wrong collocations), but there is a difference in occurrence. Wrong collocations and word sense errors make up a larger portion of the PE errors than of the HT errors, whereas the opposite is true for meaning shifts and punctuation errors. There are also a few translation method-specific problems, such as typos and misspelling of compounds for post-editing and deletions for human translations.

An example of a word sense disambiguation problem and a wrong collocation can be seen below.

(3) ST: The frequency of this phenomenon should be appreciated so that claims of apparent cure by novel treatment strategies can be seen in an appropriate context.
MT (Google): ...zodat vorderingen van schijnbare genezing... (= word sense error, judicial meaning of 'claim')
PE (x3): ...zodat vorderingen van schijnbare genezing... (= word sense error, judicial meaning of 'claim')

(4) ST: The issue of environmental degradation (...) is causing increasing concern among scientists.
MT (Google): ...steeds meer zorgen baart tussen wetenschappers. (= wrong collocation, 'among' should not be translated as 'tussen' in this context. Correct collocation: 'baart wetenschappers steeds meer zorgen')
PE: ...baart steeds meer zorgen tussen wetenschappers. (= wrong collocation. 'among' should not be translated as 'tussen' in this context. Correct collocation: 'baart wetenschappers steeds meer zorgen')

Figure 17 and Figure 18 show comparable results for technical texts: many of the top problem categories overlap for human translations and post-edited texts (terminology, logical problems, compound misspellings, meaning shifts, and untranslated text), yet there is a difference in frequency. Logical problems are slightly more common for human translations, compound misspellings make up a larger portion of PE errors than of HT errors, whereas the opposite is true for meaning shifts.

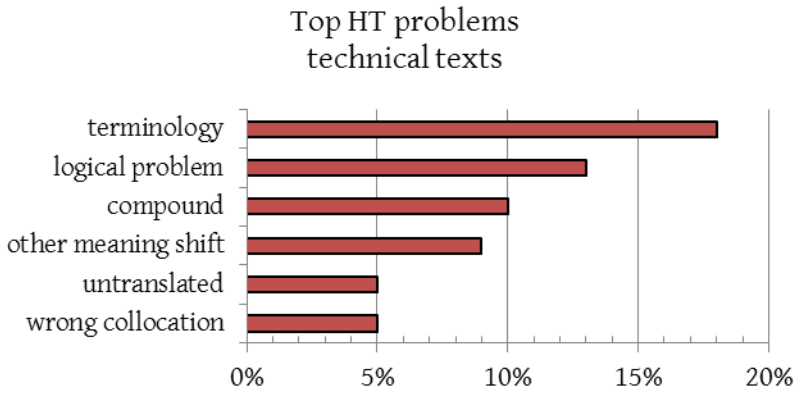


Figure 17 Overview of HT errors accounting for at least 5% of all errors made during the technical texts study.

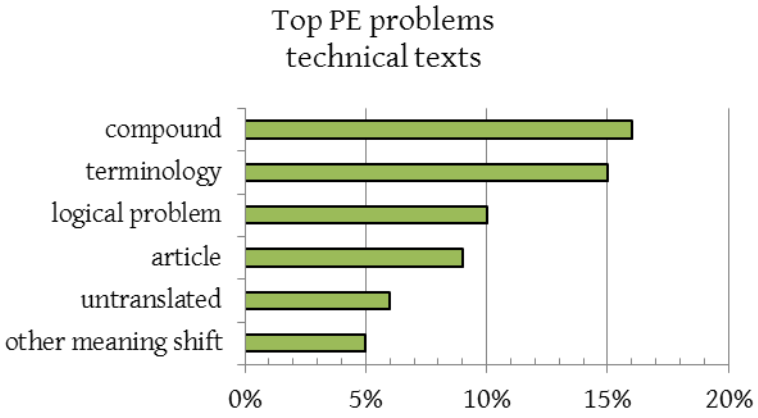


Figure 18 Overview of PE errors accounting for at least 5% of all errors made during the technical texts study.

Here as well, there are some method-specific errors: wrong collocations for human translations and missing or superfluous articles for post-editing. The high number of article, compound and terminology errors found for technical texts can be explained by the text type. In technical texts, terminology abounds. Since many English terms consist of two words which, in Dutch, usually have to be written as a one word compound, terminology issues also lead to compound problems. Example 5 illustrates this point.

(5) ST: Object Selection
Correct translation according to glossary: objectselectie
MT (Google): Object Selectie (= compound problem)
MT (Bing): Een selectie van objecten (= terminology problem)
PE (x3): Object Selectie (= compound problem)

Missing or superfluous articles can also be related to terminology issues, since most terms are nouns and nouns often require articles. However, the abundance of article errors for the second study seems to be mostly caused by missing articles in the source text. Some examples can be seen below:

(6) ST: The processing information regarding selected object is displayed in the status field.
Expected: ...regarding the selected object...
(7) ST: To select such objects they need to be opened within Master-Slide.
Expected: ...within the Master-Slide.
(8) ST: From the drop-down menu select program to record the simulation from.
Expected: ...select the program to record the simulation from.

One of the main advantages of introducing error sets is the possibility of extending our data with diagnostic information. To gain an ever better understanding of post-editing, we need to understand its relationship to machine translation quality. Since post-editing takes place after machine translation, we can look at source text passages that were problematic for a post-editor and check whether or not these passages were already problematic in the MT output. We can also try to understand which aspects of

MT output are most problematic for post-editors by comparing the source text passages that were problematic for MT with those that were still problematic after post-editing or no longer problematic after post-editing. These two aspects will be discussed in the following paragraphs. The source text-related error sets connect all MT and PE errors to the corresponding source text segments, allowing us to study this relationship in more detail. In the following charts, the relevant (sub)sets of the error sets are visualized by the Venn diagrams next to each bar chart.

We selected all source text segments that were problematic for at least one post-editor, and calculated the proportional post-editing and machine translation error weight for each of the problem types occurring in those segments. The proportionally most common PE errors for both pretests can be seen in Figure 19 and Figure 20. If a passage was already problematic in the MT, the error is considered to be caused by the MT output (lower part of the bars), if a passage was not problematic for MT, but only in PE, it is deemed to be caused by the student post-editor himself (upper part of the bars).

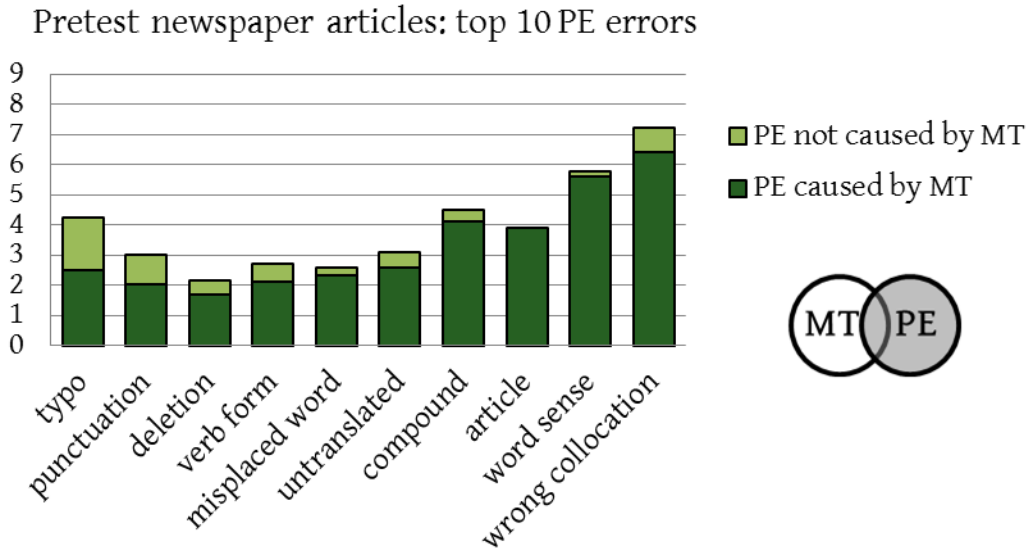


Figure 19 Most common PE errors and their origin in MT for newspaper articles. Values expressed in total proportional weight. Categories sorted from smallest to largest difference between proportional weights of both origin types.

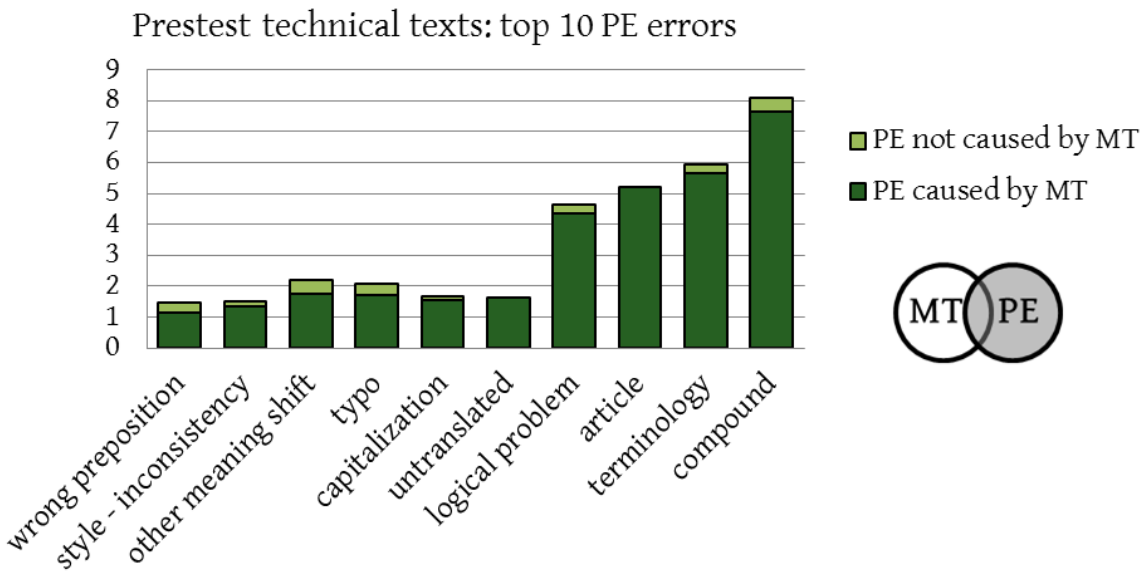


Figure 20 Most common PE errors and their origin in MT for technical texts. Values expressed in total proportional weight. Categories sorted from smallest to largest difference between proportional weights of both origin types.

It is clear from these graphs that most student PE errors do find their origin in MT errors. The high number of wrong collocation errors, word sense errors, article and compound errors (the four categories where post-editing scored worse than human translation) for the first experiment can all be explained by a high number of errors in the MT output for the same passages. The same holds true for the article and compound errors found in the second experiment. It would seem that student post-editors for general texts would benefit most from extra training in spotting and solving wrong collocations as well as word sense errors.

Figure 21 and Figure 22 take the most common MT errors on the basis of the proportional error weight as a starting point. The bars depicted in these graphs differ from the bars in Figure 19 and Figure 20 in the sense that items are no longer mutually exclusive. In Figure 19 and Figure 20, an error was either caused by MT, or it was not. In the following bar charts, however, the lower parts of the bar visualize a subset of the total bar. The bars' total length visualizes the total proportional error weight for all MT errors found. The bar minus the top section visualizes all MT errors that occurred in source text passages that were problematic for at least one post-editor after PE (= the subset of MT and PE errors). The lower part of the bar represents the actual impact of the error on PE: the total proportional weight for all PE errors found in the subset, reflecting the number of post-editors that failed to solve the MT error.

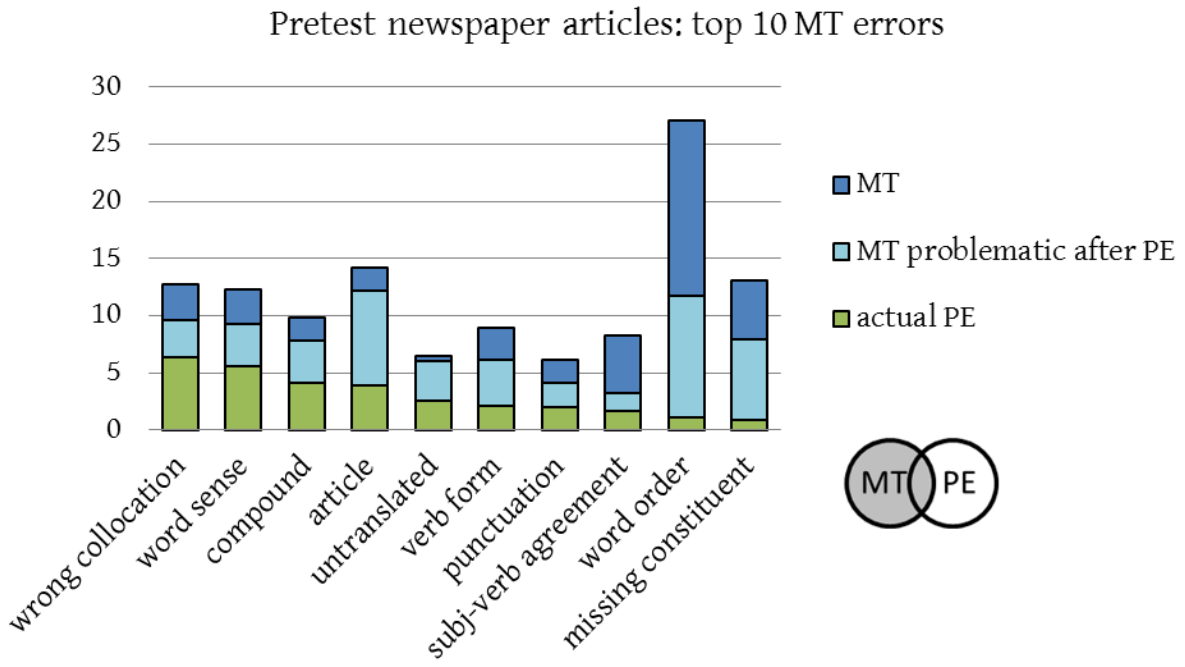


Figure 21 Most common MT errors for newspaper articles, proportion of these errors problematic for at least one post-editor and the errors' actual impact on PE. Values expressed in total proportional weight. Categories sorted from highest to lowest actual impact on PE.

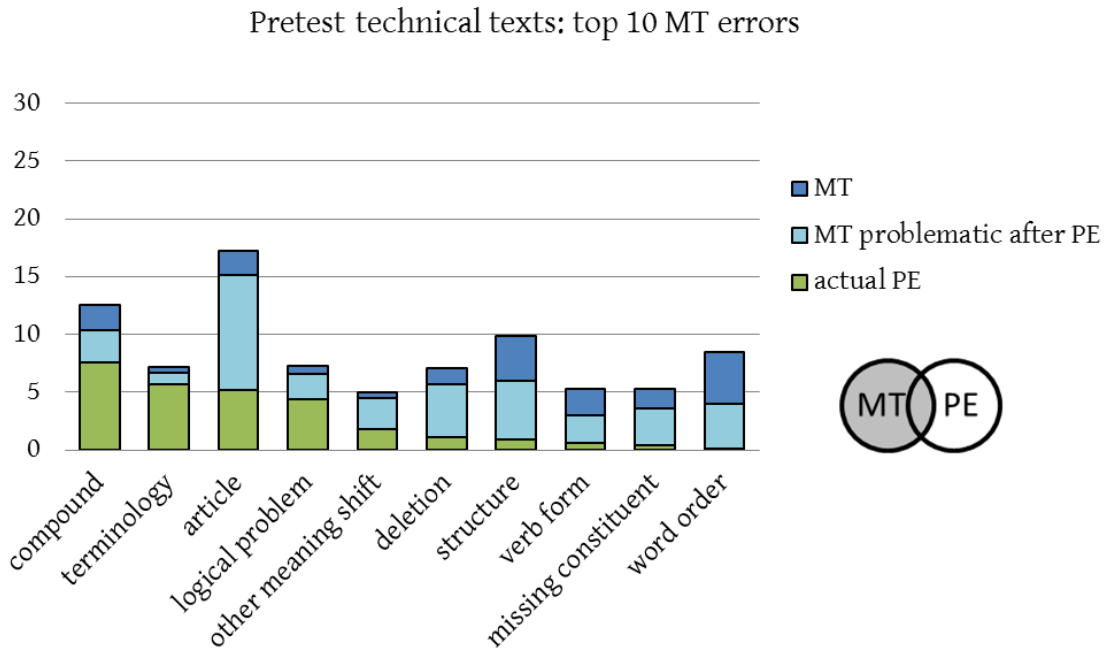


Figure 22 Most common MT errors for technical texts, proportion of these errors problematic for at least one post-editor and the errors' actual impact on PE. Values expressed in total proportional weight. Categories sorted from highest to lowest actual impact on PE.

For both text types, most MT errors seem to be problematic for at least one student post-editor (bar minus the top section), yet there is a lot of individual variation. Word order errors seem to be least problematic in comparison, along with subject-verb agreement errors for newspaper articles.

Looking at how problematic the errors actually are for student post-editors, the most problematic error categories for general texts concern the spelling of compounds, word sense disambiguation errors and wrong collocations. For technical texts, the most problematic categories are compounds, terminology issues and logical problems.

Another way of learning more about the nature of post-editing is by comparing it with a more familiar process, namely, regular human translation. We selected all segments that were problematic for at least one post-editor and one regular translator, and calculated the proportional error weight for all problem categories. The ten error categories for which the difference between post-editing and human translation was the largest, can be seen in Figure 23 and Figure 24.

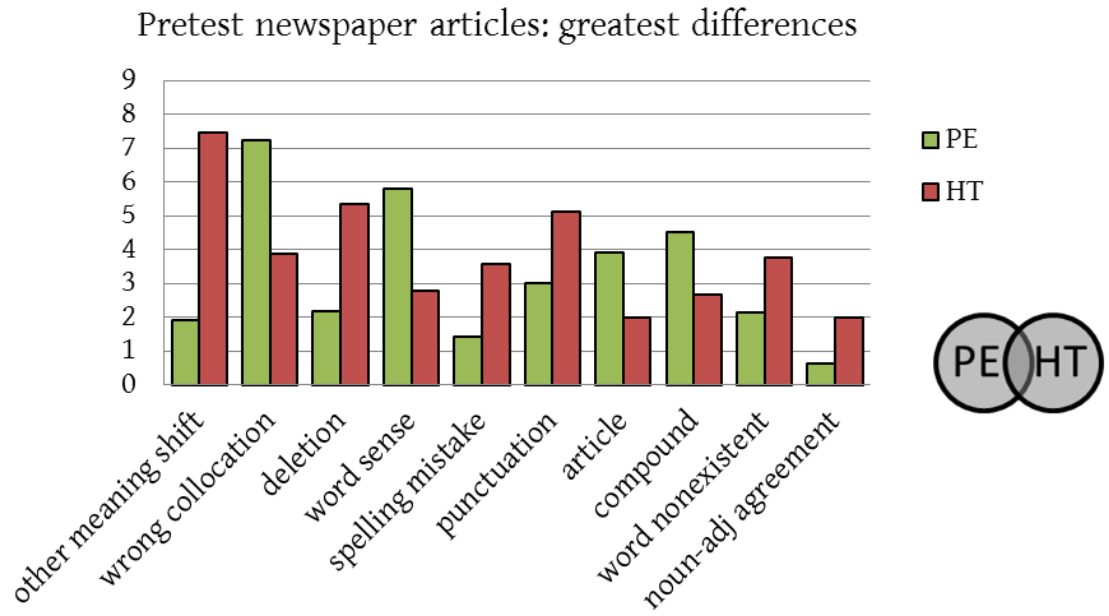


Figure 23 Top 10 error categories with greatest differences in total proportional weight between PE and HT for newspaper articles. Categories are sorted from largest to smallest absolute difference.

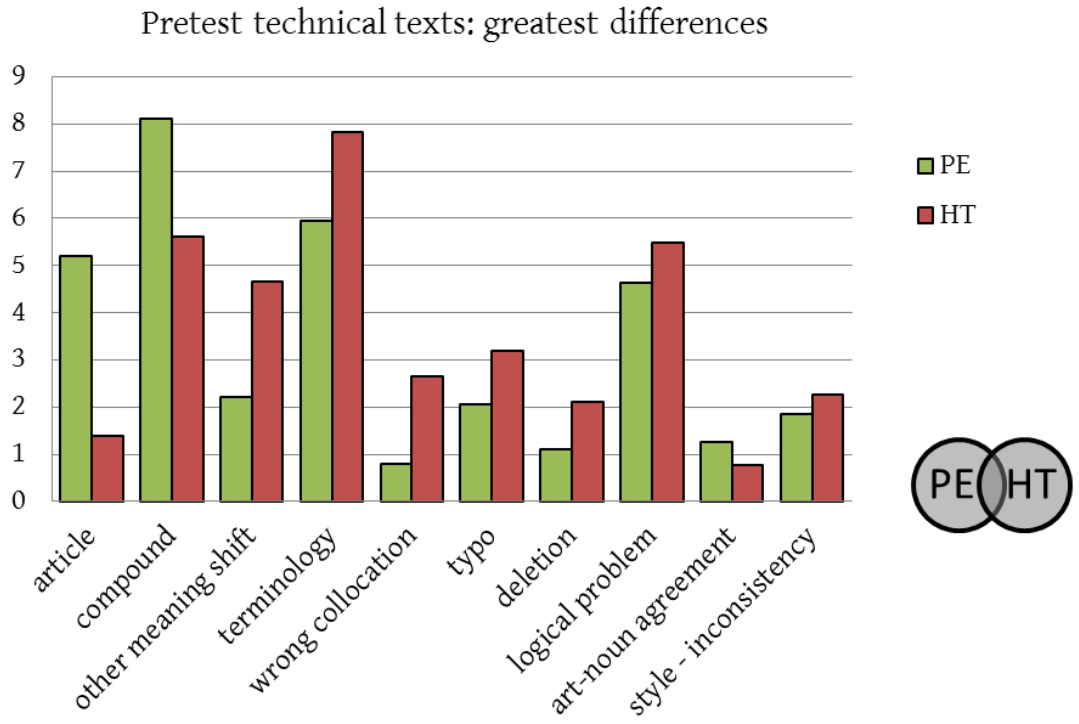


Figure 24 Top 10 error categories with greatest differences in total proportional weight between PE and HT for technical texts. Categories are sorted from largest to smallest absolute difference

When looking at these graphs, we can establish that PE only scores worse for four out of ten error categories for general texts (wrong collocation, word sense, missing or superfluous articles, and compounds), and only for two out of ten error categories for technical texts (articles and compounds). For both text types, other types of meaning shift and deletions are much more problematic for HT than for PE, whereas the opposite is true of missing or superfluous articles and the spelling of compound nouns. This can be explained in light of Figure 14, which showed adequacy issues to be the most common problem category for HT, whereas PE suffered most from spelling issues and grammatical issues

3.3.3 Attitude

All sixteen participants of the second pretest filled out a short survey after the experiment. The surveys were filled out on paper and the answers were manually transferred to a spreadsheet.

Only three participants said they never used MT, the others indicated that they used it sometimes (11 participants) or often (4 participants). When asked about the types of MT they used, most answered *Google Translate* (12 participants), while only one participant listed SDL Trados (which is one of the tools students learn to work with as part of the curriculum), and one listed Systran. The concept of machine translation

itself did not seem to be entirely clear to the participants, as some of them listed concordancer tools and dictionaries instead of machine translation systems.

In the following tables, we look at the answers to different survey questions in relationship to participants' familiarity with MT, and preference of HT compared to PE.

Table 6 Preference of translation method.

	Prefers HT over PE	Prefers HT, doesn't mind PE	Likes HT & PE equally
Often uses MT	1	3	0
Sometimes uses MT	4	4	2
Never uses MT	2	1	0

Table 6 shows that most participants still prefer human translation, although around half of them do not mind post-editing, regardless of the frequency of their MT use. No one indicated that they preferred PE over HT.

Table 7 Perceived usefulness of MT.

	MT is often useful	MT is sometimes useful
Prefers HT	0	7
Prefers HT, doesn't mind PE	4	4
Likes HT & PE equally	3	0

All participants seem to agree that MT has its uses, be it sometimes or often (Table 7). None of the participants selected the option 'MT is never useful' or 'MT is rarely useful'. The table seems to contain a relationship between preference and perceived usefulness: those who prefer HT never thought of MT as often useful, half of those who preferred HT but did not mind post-editing thought MT was often useful, the other half thought it was sometimes useful, and all of the participants that liked HT and PE equally thought MT was often useful.

Table 8 Perceived speed of HT and PE.

	HT is faster than PE	PE is as fast as HT	PE is faster than HT
Prefers HT	3	1	3
Prefers HT, doesn't mind PE	1	2	5
Likes HT & PE equally	1	1	1

When it comes to the perceived speed of each translation method, opinions are a bit more diverse (Table 8). Those who prefer HT do not necessarily think HT is also faster than PE, and those who do not mind post-editing seem to be convinced that PE is at least as fast as HT, if not faster. Those who like HT and PE equally selected all three possible options, but as there are only three participants in this group, it is hard to draw any valid conclusions from this.

Table 9 Perceived quality of HT and PE.

	HT quality is better than PE quality	PE quality is equal to HT quality
Prefers HT	6	1
Prefers HT, doesn't mind PE	5	3
Likes HT & PE equally	1	2

Although most participants indicated that MT has its uses (Table 7) and most participants thought that PE could be faster than HT (Table 8), they seemed to be less convinced about PE's final quality (Table 9). Six participants felt that PE could lead to a final quality comparable to that of HT, but twice as many participants thought that HT quality was better than PE quality.

Table 10 Attitude change towards PE after participating in the experiment.

	more positive	same	more negative
Prefers HT	3	4	0
Prefers HT, doesn't mind PE	0	7	1
Likes HT & PE equally	1	2	0

Most participants' attitude towards PE did not change after participating in the experiment (Table 10). Participants who changed their mind usually felt more positive about PE after the experiment (4 participants) than more negative (1 participant).

3.4 Discussion

3.4.1 Discussion of results

Regarding the translation process, we found post-editing to be faster than human translation for general text types. In line with our expectations, the time gains were not as big as those reported in specialised contexts with technical texts (Plitt & Masselot, 2010; Zhechev, 2014). Whereas the time needed to translate from scratch was comparable across texts, we found great differences between the different texts for post-editing. These differences could not be explained by the complexity scores alone. The low number of participants and unequal division of participants across texts could of course also be an additional contributing factor, as one weak or strong participant could seriously skew the data.

We further found that students used external resources more often when translating from scratch compared to when post-editing. The number of sentences where no external resources were consulted was significantly higher when post-editing. When looking at specific source types, there was a statistically significant difference in the usage of bilingual dictionaries, but not for the other types of external resources. This might be because students generally treat translation as a lexical task (Tirkkonen-Condit, 1990), and the MT output already provided them with most lexical information, whereas when translating from scratch, they needed to consult dictionaries to verify lexical items.

From the general quality analysis, we learned that post-editing quality is not necessarily worse than human translation quality, corresponding to the findings by Carl, Dragsted, Elming, et al. (2011) and Garcia (2011). Just as Guerberof (2009) found more mistranslations in HT output compared to post-edited texts, PE adequacy often outperformed HT adequacy in our pretests. This "might indicate that using MT helps translators clarify possibly difficult aspects of the source texts thus improving general comprehension of the text" (Guerberof, 2009, para. 4.2).

The fine-grained error analysis showed that some problems were more common in post-editing (such as wrong collocations and word sense errors for general texts and

misspelling of compounds for technical texts), and some occurred equally frequent across both translation methods (such as terminology issues for technical texts). The abundance of wrong collocations in PE can be explained by the fact that student translators are often not critical enough of literal MT translations (Depraetere, 2010), although wrong collocations were slightly more problematic for human translation during the technical translation task. The abundance of terminology issues in both methods of translation is in contrast with Guerberof (2009), who found that post-edited segments contained more terminology issues than segments that were translated from scratch. The fact that many of the most common error types overlap between human translation and post-editing (four error categories for newspaper articles and six for technical texts) could indicate that these translation methods are not as different from one another as sometimes thought (O'Brien, 2002). Knowledge of error types can be integrated into post-editor training and translation tool development. Perhaps an extra warning could be integrated into a tool whenever certain polysemous words or awkward collocations could occur in the MT output. Such a tool would also benefit from a spell-checker, since this could reduce the large number of misspelled compound nouns, typos and punctuation errors found in the pretests. Regarding post-editor training, the post-editing of different text types could be used to make students aware of text type specific issues. For example, terminology, logical problems, missing or superfluous articles and untranslated text only belonged to the most common problem categories for technical texts, but not for newspaper articles. The misspelling of compound nouns, which was relatively problematic in the post-editing of newspaper articles (accounting for 6% of all PE errors), was a far more common issue for technical texts (accounting for 10% of all HT errors and 16% of all PE errors made). Regarding technical texts specifically, it is striking that, even when given a terminology list to adhere to, one of the most common problems in students' translations and post-edited texts is incorrect terminology. Depraetere (2010), like Guerberof (2009), suggested that "it is also necessary to provide a terminology list" (para. 6), but it would seem from our experiment that students need additional terminology management training on top of the terminology list.

Regarding MT evaluation, we can conclude that SMT systems could greatly benefit from some kind of rule-based post-processing step, in order to minimize the number of syntactical and grammatical errors. This idea has already been proven successful for English-Czech translations (Mareček, Rosa, Galuščáková, & Bojar, 2011). The introduction of such a step allows post-editors to focus more on lexical issues, logical problems and adequacy errors. For technical texts in particular, it is clear that terminology must be integrated into the translation tool, to minimize the number of terminology errors and compound misspellings.

Error sets helped us gather data on the most problematic text passages for each method of translation. This information can in turn be used to identify linguistic

checkpoints: elements in the source language that can present problems for translation (Naskar, Toral, Gaspari, & Way, 2011). Linguistic checkpoints help facilitate evaluation, from MT systems to student translations. Developers can use them to test the quality of an MT system when dealing with a specific type of problem, and teachers can use them to test students' translation skills on passages that are problematic for most translation students. A similar process is the Calibration of Dichotomous Items (Eyckmans, Anckaert, & Segers, 2009), where problematic ST passages are identified and then used as quality checkpoints for the assessment of student translations. While the most common MT errors, such as grammatical errors, were easily corrected by most students during post-editing, other error types proved more difficult to correct, and could make for interesting linguistic checkpoints: wrong collocations and word sense errors for newspaper articles, and terminology issues, compound spelling, and missing articles for technical texts. On the MT end, it could be interesting to integrate the error set information into MT confidence information. A good post-editing tool would perhaps benefit from warnings whenever certain awkward collocations or polysemous words could occur.

3.4.2 Discussion of methods for upcoming experiments

While section 3.4.1 showed some of the possible conclusions we can draw from the pretests' data, the pretests' main goal was establishing which aspects of the research design worked well, and which aspects we needed to tweak for our main experiment. In addition, the pretests' data was mainly used to create and evaluate our translation quality assessment approach.

We found differences between the different texts that could not be explained on the basis of Editcentral's complexity scores alone. There are two possible explanations for this. First, complexity scores might not correspond to translatability, so additional factors needed to be taken into account when selecting texts for our main experiment. Second, the number of participants differed greatly across texts and tasks, making it possible for individual students to severely skew the data. This made us develop a more balanced design for our main experiment.4

Both pretests had a different setup. For the pretest on technical texts, participants could choose the output of two different MT systems (Bing Translators and *Google Translate*), and we had to impose a time-limit because the experiment needed to take place during the students' Technical Translation class. For the main experiment, we decided to only use output from one MT system, as including an additional system only adds more parameters, and we needed the experiment to be as controlled as possible. We further did not want to impose a time-limit, as time pressure has been shown to

influence the revision process during translation, and to force translators to accept translation memory system suggestions more easily (Alves & Campos, 2009).

Although *PET* was a decent tool for our pretests, it had a few drawbacks for our main experiment. External resources were registered through participants' self-assessment after each segment. Participants complained that this was extremely time-consuming and repetitive. It is possible that participants forgot to mention certain resources (either deliberately or accidentally), making it hard to truly evaluate the usage of external resources. In addition, it was time-consuming to convert these individual assessments to an analysable format. Furthermore, *PET* could not be integrated with an eye tracker, and cognitive load - which can be measured via fixation data - was one of the aspects we wanted to evaluate with our main experiment.

Regarding the translation quality assessment approach, the inter-annotator agreement showed that the proposed categorisation was clear and that it was necessary to include a consolidation phase. Annotations after consolidation were successfully used for a variety of quality assessments with varying degrees of granularity. Though the approach required much time and human effort (the annotation process in itself cost around 45 minutes for 150 words of an unseen MT text, with acceptability annotations requiring the most time: 30 minutes), it did provide rich data. As familiarity with a text increased, the annotation time rapidly decreased, and the annotation time for HT or PE was also lower than for MT.

Although the concept of source text-related error sets seemed worth investigating, the application of the concept left something to be desired. Creating the error sets was extremely time-consuming, and the proportional values are highly influenced by the number of participants, especially with such small numbers of participants. While we do believe that using multiple translators' product analyses to pinpoint specific problematic passages or error types is a promising avenue for future research, we would need to be able to automatically create source text-related error sets to make it feasible, which was beyond the scope of this PhD project. For our main experiment, we turned to advanced statistical analyses rather than source text-related error sets for our comparative analyses.

Using a paper survey for the attitude aspect of our pretests was necessary to ensure that all participants handed in their surveys, but digitising the answers was relatively time-consuming. In addition, participants were only asked about their attitudes after they had participated in the experiment. For our main experiment, we therefore decided to conduct two surveys (one before the experiment, one after the experiment) with an online survey tool, to facilitate data processing.

Chapter 4 Method

The pretests helped us develop a suitable method for the main experiment, which will be explained in detail in the present chapter. Important aspects include the more rigorous text selection process, more advanced logging tools, two different types of participants (students and professional translators), and more advanced analyses. All of the following chapters contain findings of studies conducted on the basis of this main experiment. Chapter 5 consists of a general comparative analysis between human translation and post-editing as well as students and professional translators. Chapter 6 and Chapter 7 zoom in on two more specific aspects: the impact of MT output on post-editing, and the usage of external resources.

4.1 Materials

4.1.1 Source text selection

As the pretests showed some text-specific differences that could not be explained by complexity scores alone, we decided to apply a more rigorous text selection for our main experiment. We tried to control for as many factors as possible, so that our findings could be attributed to differences in experience and translation method, and not to between-text differences. Rather than testing for comprehensibility after text selection, we started by selecting fifteen different English newspaper articles from *newsela.com*, a website providing newspaper articles at different levels of complexity, as indicated by a Lexile score. We selected these articles on the basis of having comparable

Lexile scores, between 1160L and 1190L[1], and we made sure to select articles with different non-specialised topics. *Lexile®* measures are a scientifically established standard for text complexity and comprehension levels, giving a more accurate representation of how challenging a text is than existing readability measures. The scores are usually used in classrooms to provide students with texts of their appropriate reading levels. Our study is - to the best of our knowledge - the first one to apply these measures to translation research. Each article was reduced to its first 150-160 words, as participants would have to translate multiple texts during each session and eye-tracking data is more accurate for shorter sessions. Although the Lexile measure provided a baseline complexity score, we did select only a part of each article, so we further analysed each text for additional readability measures and potential translation problems. Two of the texts were discarded as they had become either too complex or too specific (too many proper nouns, political discourse and sports-related idioms) after reducing them to the first 150-160 words. We compared the remaining texts for sentence length (discarding texts with an average sentence length above 20 words and below 15 words), machine translation quality on the basis of our translation quality assessment approach (output obtained via *Google Translate* on January 24, 2014, analysis with a focus on word sense errors and wrong collocations as our pretest on newspaper articles showed these to account for 20% of all errors after post-editing), word frequency (based on the frequency information in the *Corpus of Contemporary American English, COCA*), and number of proper nouns. The final selection consisted of eight texts of seven to ten sentences long. The selected texts and their Dutch MT translations can be found in Appendix 3.

4.1.2 Survey creation

Surveys are the best way to gain insight into translators' experience with and attitude towards machine translation and post-editing. We already incorporated a survey into our second pretest, but it was a paper one - making it harder to process the responses - and it was only a short survey after the experiment. Ideally, we wanted to get some information on participants' experience and attitude before taking part in the experiment, as well as their experiences after taking part in the experiment. We therefore decided to create two surveys: one to be filled out before the experiment, one

[1] The authors would like to thank MetaMetrics® for their permission to publish Lexile scores. https://www.metametricsinc.com/lexile-framework-reading

to be filled out after (see Appendices 4 and 5). Both surveys were created online with *Qualtrics*[2].

The survey to be taken before the experiment contained some personal questions (age, gender) followed by questions about participants' professional background, tailored to the participant type (students received questions about their studies, professional translators about their education, work experience, translation throughput, and text types), and linguistic background (native language and - for the professional translators - working languages). The rest of the survey was devoted to questions about participants' awareness of existing MT systems and their experience with using them (whether they used them (and how), whether they found post-editing rewarding and useful, whether they found it as fast as human translation, and whether they felt the final quality was comparable to that of a human translation).

The survey taken after the experiment was a little shorter. It asked participants about their preferred method of translation for the texts translated during the experiment, the method they experienced as being the fastest, the method they found the least tiring, and the method they found the most useful. This was followed by some open questions where participants could elaborate on possible frustrations with the MT output and interface of the tool, as well as their general experience during the experiment.

Both surveys were tested by colleagues from the Language and Translation Technology Team (LT³) to make sure all questions were unambiguous and the conditional structure of the survey worked the way it was supposed to for different scenarios.

4.1.3 Tool selection

As discussed in section 3.4.2, the *PET* tool could not be integrated with an eye tracker and it was not capable of automatically registering external resources. We therefore looked for other possible registration tools. After attending a training school on keystroke logging organised by the developers of *Inputlog* and a PhD course in Translation Process Research by the developers of *CASMACAT*, we realised the potential of these tools for our own research.

The *CASMACAT* translator's workbench (Alabau et al., 2013) looks like an actual translation environment, which greatly improved the ecological validity of our experiment. In this study, we used a simplified version of *CASMACAT*, without

[2] https://www.qualtrics.com/

interactive translation. The tool contains keystroke logging and mouse tracking software to be able to observe the translation and post-editing process in detail. In addition, there is a plugin to connect an *EyeLink 1000* eye tracker with *CASMACAT*. The EyeLink registered participants' eye movements and this fixation data was automatically added to the *CASMACAT* logging data, to get the full picture of the translation process. The data contained, for example, the number of fixations on source and target text, the average fixation duration on source and target text, the number of production units (instances of continuous typing activity separated by pauses), the number of words within each production unit, time spent on parallel reading and writing activity, and the time needed to translate a segment (with or without pauses longer than 5 seconds). In our setup, we used the EyeLink with head support (consisting of chin and forehead rest, see Figure 25). While this arguably somewhat negatively influenced the ecological validity of our experiment, it made the fixation data much more accurate than it would have been without head support. Most participants remarked that they were less aware of the eye tracker during the experiment than they had expected to be.

An additional advantage of *CASMACAT* is that its data is compatible with the *Translation Process Research Database (TPR-DB)*, making it easier to share data and compare findings with other researchers. The *TPR-DB* comes with a suite of scripts to process the raw *CASMACAT* logging files and turn them into analysable tables.

Figure 25 Setup with the EyeLink eye tracker used in the experiment.

Using *CASMACAT* in isolation, however, would not be sufficient, as it only logs what happens within the *CASMACAT* interface itself, and we wanted to gain insight into the usage of external resources as well. As screen recordings have the drawback of having to be manually analysed and recoded (Göpferich, Jakobsen, & Mees, 2009) and online search reports have the drawback of being incomprehensive and time-consuming, we chose *Inputlog* (Leijten & Waes, 2008), a different keystroke logging tool, to help record

the usage of external resources. Originally intended for writing research within the *Microsoft Word* environment, *Inputlog* is capable of logging all applications and browser tab information, enabling us to automatically extract information on the usage of external resources. Once activated, the program disappears into the toolbar and runs in the background, so it did not interrupt the translation process itself at any time.

4.2 Participants

Participants were 10 master's students of translation (2 males) at Ghent University who had passed their final English Translation examination, and 13 professional translators (3 males). All participants were native speakers of Dutch and had English as one of their main working languages (as part of their studies for the student participants, and as part of their work for the professional translators). When asked about the language participants felt most confident translating from, most participants (7 students and 7 professional translators) selected English. Other languages were Italian, French, and Spanish for the students, and French for the professional translators. When English was not the first choice, it was always selected as the second best working language. With the exception of one translator, who had two years of experience, all translators had a minimum of 5 years and a maximum of 18 years of experience working as a full-time professional translator. Median age of students was 23 years (range 21-25). Median age of professional translators was 37 (range 25-51). All participants had normal or corrected to normal vision. Two students wore contact lenses and one student wore glasses, yet the calibration with the eye tracker was successful for all three. Two professional translators wore lenses. Calibration was problematic for one of the professionals. Sessions with problematic calibration were removed from the data. By way of compensation for participating, students received gift cards worth 100 euros in total, professional translators were paid 300 euros and their travel costs were refunded.

Students reported that they sometimes used MT systems as an additional resource during translation, but they had received no explicit post-editing training. Some professional translators had basic experience with post-editing, although none of the translators had ever post-edited an entire text. Their personal experience with post-editing – if any – was limited to MT output offered by a translation tool whenever the TM did not contain a good match.

To assess English proficiency - an important aspect of translation competence (Göpferich, 2009; PACTE, 2003) - all participants performed a LexTALE test (Lemhöfer & Broersma, 2012), which is a word recognition task that is also an indicator of vocabulary knowledge. We expected to see a clear difference in proficiency between students and

professionals, yet no statistically significant difference in LexTALE scores was found: $t(21)=0.089$, $p=0.47$; $\mu(\sigma^2)$ professionals = 88.27(90.24), $\mu(\sigma^2)$ students = 88(60.01).

4.3 Procedure

The experiment consisted of two sessions for each participant in the periods of June/July 2014 for students and April/May 2015 for professional translators. Most participants completed both sessions on two consecutive days. If they did complete both sessions on the same day, there was a minimum one-hour lunch break between both sessions. We used a combination of surveys, logging tools and a retrospection session to be able to triangulate data from different sources.

The first session started with the first survey and a LexTALE test. This was followed by a copy task (participants had to copy a text in *Microsoft Word* to get used to the keyboard, the EyeLink's head support and the screen) and a warm-up task in *CASMACAT*, combining post-editing and human translation, so participants could get used to the environment, the tools, and the different types of tasks. The actual experiment consisted of two texts that they translated from scratch, and two texts that they post-edited. Participants only received one text at a time, and the text was subdivided into editable segments, corresponding to sentences in the source text, with the exception of titles, which were presented as one segment even when they consisted of more than one sentence. The entire text was visible to the participants, and they were able to move backwards and forwards throughout the text, although only one segment could be edited at a time. For the human translation task, the left hand side of the screen contained the source text, the right hand side of the screen was empty. For the post-editing task, the left hand side of the screen also contained the source text, but the right hand side of the screen contained the machine translation output (see Figure 26 for an example of the *CASMACAT* interface). Before each task, the eye tracker was calibrated.

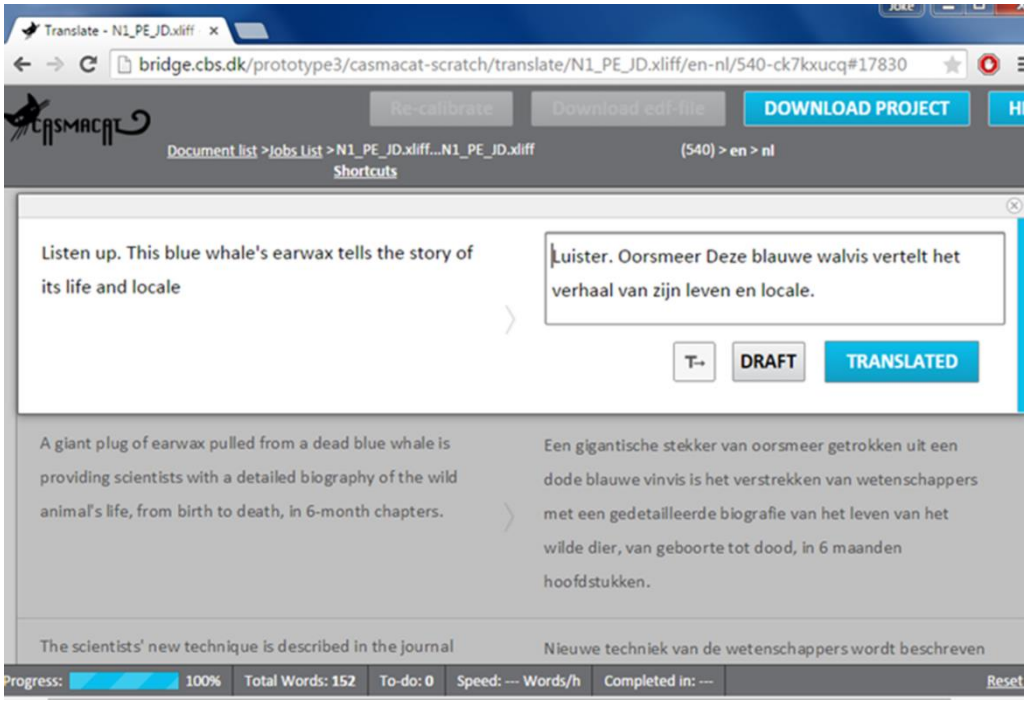


Figure 26 Example of the *CASMACAT* interface for a post-editing task.

The second session started with a warm-up task as well, followed by post-editing two texts and translating two texts from scratch. The order of texts and tasks was balanced across participants within each group in a Latin square design (see Table 11). In total, there were sixteen possible versions (orders) of the experiment. Students received versions one to five and nine to thirteen, professional translators received versions one to seven and nine to fourteen. The final part of the second session consisted of unsupervised retrospection (participants received the texts which they just translated in Microsoft Word and were requested to highlight and comment on elements they found particularly difficult to translate) and the second survey.

Table 11 Latin square design, mixed text order and task order. Columns are labelled with version codes, cells contain codes for the task type (PE=post-editing, HT=human translation) and text (ranging from 1 to 8).

	Version	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Session1	task1	PE1	PE8	PE7	PE6	PE5	PE4	PE3	PE2	HT1	HT8	HT7	HT6	HT5	HT4	HT3	HT2
	task2	PE2	PE1	PE8	PE7	PE6	PE5	PE4	PE4	HT2	HT1	HT8	HT7	HT6	HT5	HT4	HT3
	task3	HT3	HT2	HT1	HT8	HT7	HT6	HT5	HT4	PE3	PE2	PE1	PE8	PE7	PE6	PE5	PE4
	task4	HT4	HT3	HT2	HT1	HT8	HT7	HT6	HT5	PE4	PE3	PE2	PE1	PE8	PE7	PE6	PE5
Session2	task5	HT5	HT4	HT3	HT2	HT1	HT8	HT7	HT6	PE5	PE4	PE3	PE2	PE1	PE8	PE7	PE6
	task6	HT6	HT5	HT4	HT3	HT2	HT1	HT8	HT7	PE6	PE5	PE4	PE3	PE2	PE1	PE8	PE7
	task7	PE7	PE6	PE5	PE4	PE3	PE2	PE1	PE8	HT7	HT6	HT5	HT4	HT3	HT2	HT1	HT8
	task8	PE8	PE7	PE6	PE5	PE4	P23	PE2	PE1	HT8	HT7	HT6	HT5	HT4	HT3	HT2	HT1

There was no time limit, and participants were allowed to take breaks after each task or continue immediately with the next task according to their own preferences. We wanted to keep the experiment as ecologically valid as possible, and we assumed that especially students were not used to working on translations for more than an hour, and that professional translators also needed frequent breaks during their regular working hours.

4.4 Data exclusion and preparation

For each participant, we collected logging data for four post-editing tasks and four regular translation tasks, leading to a total number of 92 post-editing tasks and 92 regular translation tasks. The data for each participant consisted of the answers to the surveys before and after the experiment, *CASMACAT* logging files including keystrokes and fixations for the two warm-up tasks and the eight translation and post-editing sessions, *Inputlog* logging files containing keystrokes and external resources for the eight translation and post-editing tasks, and *Microsoft Word* documents containing the problematic passages participants highlighted.

All student sessions could be used for further analysis, but some of the professional translators' data had to be discarded due to technical problems. Either something went wrong with the logging files, there was an issue with calibration, or translators

accidentally closed the *CASMACAT* interface, leading to a disruption in the logging files. Rather than work with potentially problematic data, we discarded those recordings altogether. In total, five human translation and five post-editing tasks were discarded, leading to an overall total of 87 post-editing tasks and 87 human translation tasks.

Using the scripts provided with the *TPR-DB*, the *CASMACAT* xml-files were prepared for word alignment. A first, automatic, alignment was done with Giza++ (Och & Ney, 2003), which we then manually corrected with the *YAWAT* tool (Germann, 2008). Data from the aligned files was extracted and converted to more manageable table formats with another *TPR-DB* script.

4.5 Enriching the *TPR-DB* with external resources

From the *Inputlog* data, we extracted the focus events with the provided software (focus events contain information on the opened application or screen, time spent in the application, and keystrokes). We then manually grouped the different events into categories: dictionary, web search, concordancer, forum, news website, encyclopedia, MT, synonym search, spelling, term bank, and conversion. Most types of external resources were only sporadically used, with the exception of search engines, concordancers, dictionaries, and encyclopedias.

A next step was to combine the *CASMACAT* and *Inputlog* data for subsequent analysis. Since this is the very first study where data from both tools are combined, the *TPR-DB* had to be updated to accommodate for the new data. An InjectIDFX-script was developed to merge *Inputlog* data with the *CASMACAT* xml-files. *CASMACAT* only logs the keystrokes and events within the *CASMACAT* interface. The xml-files themselves contain a 'blur'-event whenever a person leaves the *CASMACAT* interface and a 'focus'-event whenever they return to the *CASMACAT* interface, but whatever happens between the blur and the focus-event is unknown. By adding the *Inputlog* data to the xml-files, we can analyse what happens when a person leaves the *CASMACAT* interface as well. We added an extra table to the *TPR-DB*: the EX-table, containing information on external resources consulted, the time spent in the resource, and keystrokes made within the external resource. We added an extra column to the EX-file where we added the categories we had assigned to the various *Inputlog* events. An extract from an EX-file can be seen in Table 12 below.

Table 12 Excerpt from EX-file[3].

EX id	Focus	Time	Dur	ST segN	ST segL	STidN	STidL	KD idN	KD idL	edit	category
...											
3	Translate - T1_T5_PE_P9.xlf - 204 - Google Chrome	-53975	0	9629	-1	5	-1	0	-1	---	MAIN
4	Nieuw tabblad - Google Chrome	81778	3360	9630	9629	15+16	12+13+14	122	121	woorden[..]nlij	NAVIGATION
5	Woordenlijst Nederlandse Taal - Officiële Spelling - Google Chrome	85138	3937	9630	9629	15+16	12+13+14	122	121	groot-bri	SPELLING
6	Translate - T1_T5_PE_P9.xlf - 204 - Google Chrome	89075	123512	9630	9629	15+16	12+13+14	122	121		MAIN
7	Nieuw tabblad - Google Chrome	212587	3548	9633	9632	75	70	193	192	linguee	NAVIGATION
8	Linguee | Nederlands-Engels woordenboek (en andere talen) - Google Chrome	216135	2718	9633	9632	75	70	193	192	n fact	CONCORDANCER
9	in fact - Nederlandse vertaling – Linguee woordenboek - Google Chrome	218853	4765	9633	9632	75	70	193	192		CONCORDANCER
10	Translate - T1_T5_PE_P9.xlf - 204 - Google Chrome	223618	264006	9633	9632	75	70	193	192	eed	MAIN
...											

Looking at the 'Focus' column and corresponding category label in Table 12, we see the participant moving from the main document (*CASMACAT*, EXid 3) to a new tab in *Google Chrome* (EXid 4), where he types 'woorden...' (see 'edit'), leading him to the Dutch spelling website *Woordenlijst* (EXid 5). He then types 'groot-bri' to look up the Dutch

[3] Each time the participant switches to another screen or application, a focus event is recorded, with code EXid and a label found in column 'Focus'. Time is time in ms since the beginning of the session, Dur is the time in ms spent in a particular focus event. STsegL represents the last segment opened in *CASMACAT* before leaving the tool, STsegN is the next segment opened after returning to the *CASMACAT* tool. STidL and STidN represent the last source token before leaving *CASMACAT* and the next token after returning to *CASMACAT*. KDidL and KDidN contain the ID of the last keystroke before leaving *CASMACAT* and the next keystroke after returning to *CASMACAT*. The actual characters typed within a focus event are shown in the column 'edit'. Each focus event is given a corresponding category.

spelling of Britain (*Groot-Brittannië*). After this search, he returns to the *CASMACAT* interface (EXid 6) for two minutes, after which he again opens a new tab in *Google Chrome* (EXid 7) for the next search: 'linguee', allowing him to go to the *Linguee* concordancer (EXid 8), where he looks up the translation of 'in fact' (EXid 9) before returning to the *CASMACAT* document once more (EXid 10).

It is currently impossible to automatically map external resources to the corresponding segment. In the data file, there is a column for the last segment that was open before the *CASMACAT* interface was left, and the first segment to be opened after returning to the *CASMACAT* interface, but the search itself could be related to either one, or even an entirely different segment. For example, a person can look up a word in a dictionary while translating the first segment of a text. If the person goes back to the *CASMACAT* interface without closing the screen with the search query on it, the next time that person opens the search query, this will show up exactly like the search made during the first segment in the data. It would require a lot of extra manual work to label each external resource with the correct segment. In the future, we will try to better map the *CASMACAT* and *Inputlog* data by looking at keystrokes or by filtering on the time spent on certain pages. At the moment, however, we decided to link resources to the last segment that was open at the time of the consultation, as this is most likely the most relevant segment. As can be seen in Table 12, categories were added to each line in the external resource file. These were added manually. In the final data file containing all segments, we added the number of times and the total duration for each type of external resource. To be able to better compare the data across all segments, we normalised the counts and durations by dividing them by the number of source text tokens.

Chapter 5 General analysis[1]

We carried out a general comparative analysis of the two translation methods (human translation and post-editing) for the two participant groups (student translators and professional translators). We were interested in the following main aspects: (i) the differences in process, as recorded by the logging tools during the experiment, (ii) the quality of the final product, as established by means of our translation quality assessment approach, and (iii) translators' general attitude towards post-editing and their experience with it, as recorded by the surveys before and after the experiment. In addition to these aspects, we also wanted to take a closer look at the impact of the usage of external resources on overall speed and quality. These analyses have been included in sections 5.1.4 and 5.3, respectively. A summary of the models discussed in the following sections can be found in Appendix 6, with exception of the models with three-way interactions, which have been included directly in the text to facilitate comprehension.

5.1 Process analysis

The data for the process analysis consisted of the concatenated SG-files as obtained by processing the *CASMACAT* data (Carl, Schaeffer, & Bangalore, 2016). We normalised a few variables and included some additional variables (which will be discussed where relevant) before loading the data file into *R*, a statistical software package (R Core Team, 2014). In total, the data file consisted of 1444 observations - i.e., segments. For each analysis, we excluded the segments with incomplete data (due to minor problems with

[1] Parts of this chapter will be published as Daems, Vandepitte, Hartsuiker, and Macken (2016b).

the eye tracker or keystroke logger). The number of segments retained was never lower than 1412.

All analyses discussed below were performed with *R*. We used the *lme4* package (Bates, Maechler, Bolker, & Walker, 2014) and the *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2014) to perform linear mixed effects analyses on our data. Mixed effects models contain random effects in addition to fixed effects (= independent variables). In our case, the random factors were always the participant (since we expect individual differences across participants to potentially influence the data) and the sentence code (an identifier of the text and the exact sentence in that text, since sentence-inherent aspects may also influence the data). A mixed model is constructed in such a way that it can identify the effect of independent variables on dependent variables while taking these random factors into account.

Whenever we discuss mixed models below, the first step is to build a null model, which contains only the dependent variable and random factors. In the next step, the predictor (or independent) variables are added to the model and tested against the null model, to see if the predictor variable is actually capable of predicting the dependent variable. The predictor variables in the following models are always translation method (human translation or post-editing) and experience (student or professional) with interaction. To compare and select models we calculated Akaike's Information Criterion (AIC) value (Akaike, 1974). The actual value itself has no meaning, only the difference between values for different models predicting the same dependent variable can be compared. According to Burnham and Anderson (2004), the better of the models being compared is the one with the lowest AIC value. Their rule of thumb states that if the difference between models is less than 2, there is still substantial support for the weaker model. If the difference is between 4 and 7, there is far less support for the weaker model, and if the difference is greater than 10, there is hardly any support for the weaker model.

5.1.1 Speed

The first aspect we investigated is translation speed. We built a mixed model with the average duration per word as a dependent variable. The model with predictors performed significantly better than the null model, yet only the translation method had a significant effect, with post-editing reducing the time needed per word with almost a second compared to human translation. The effect is plotted out in Figure 27. Students seem to require somewhat more time than professionals, although this effect was not significant.

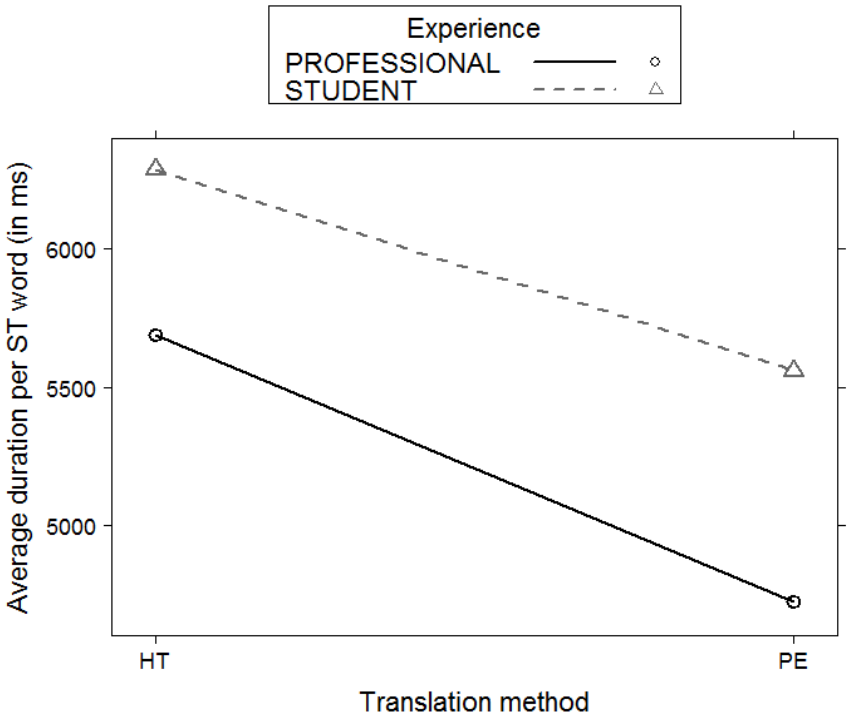


Figure 27 Effect plot of interaction effect between method (human translation and post-editing, HT and PE, respectively) and experience (professional and student) on translation speed (=average duration per word in ms).

5.1.2 Fixations

In addition to speed, we also calculated average fixation durations and total number of fixations to get an indication of cognitive effort. Average fixation duration was calculated by dividing the total fixation time within a segment by the number of fixations for that segment. Table 13 contains an overview of the average fixation duration on source and target segments for human translation and post-editing for both levels of experience, giving a first impression of the expected differences. The average fixation duration seems to be longer on target text segments than on source text segments in all cases.

Table 13 Average fixation duration across all segments.

	students		professionals	
	source	target	source	target
HT	229	273	218	266
PE	220	257	216	252

We built a mixed model with the average fixation duration as dependent variable. As was the case for speed, only method was a significant predictor, with the average

fixation duration being 5 milliseconds shorter when post-editing compared to human translation. The effect is plotted in Figure 28. Again, there seems to be a trend for fixation duration to be longer for students compared to professional translators, but this effect was not found to be significant either.

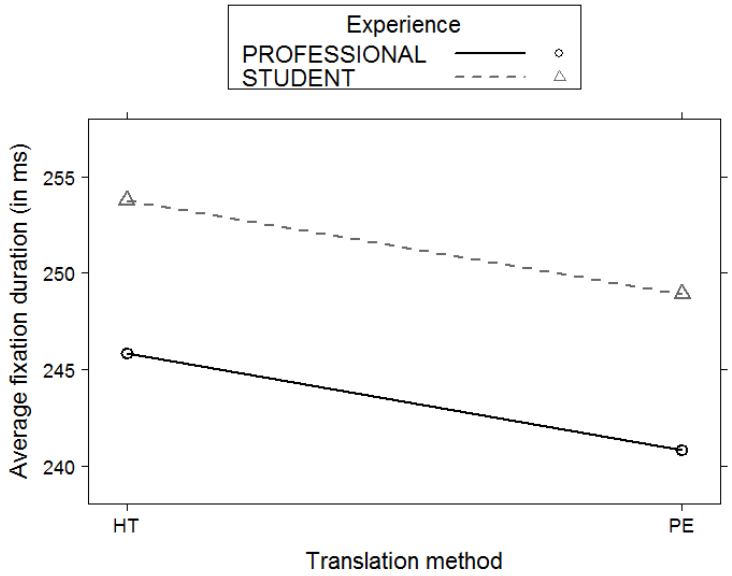


Figure 28 Effect plot of interaction effect between method and experience for the average fixation duration (in ms) across the whole text.

While overall average fixation duration gives us some indication of cognitive load, we also investigated fixations on source and target texts separately. For the analysis of the number of fixations on the source text, the fitted model again performed better than the null model, but only method was found to be significant. Processing of the source text during post-editing required fewer fixations per word than for human translation. In the analysis on the average fixation duration on the source text, only the interaction between method and experience is significant, showing that - for students only - the average fixation duration on the source text during post-editing is significantly shorter than during human translation (Figure 29).

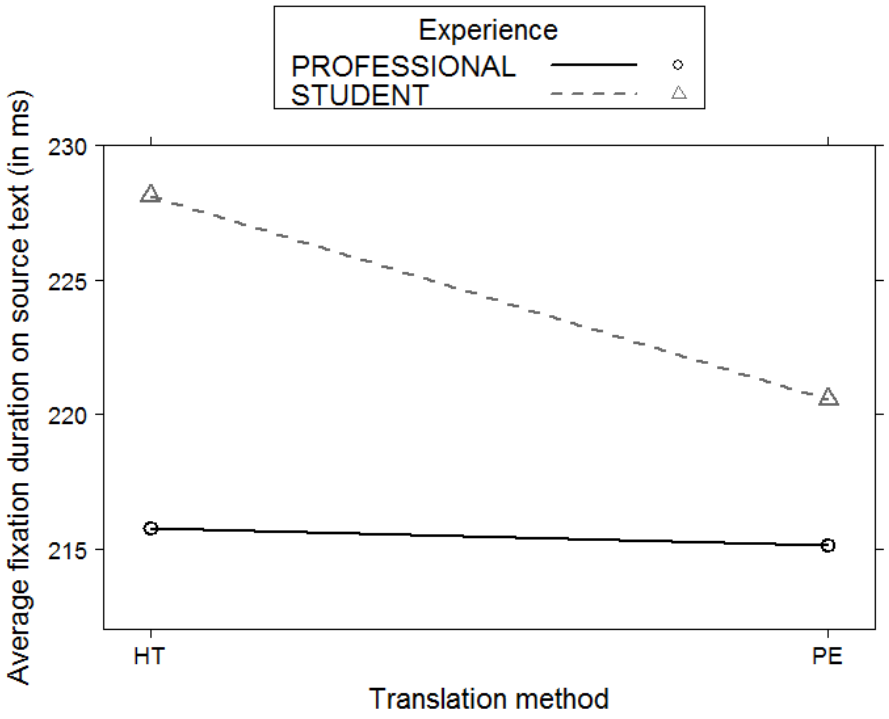


Figure 29 Effect plot of interaction effect between method and experience for the average fixation duration (in ms) on the source text.

For the average number of fixations on the target text, the summary of the fitted model showed only the interaction effect to be significant. The effect is plotted in Figure 30.

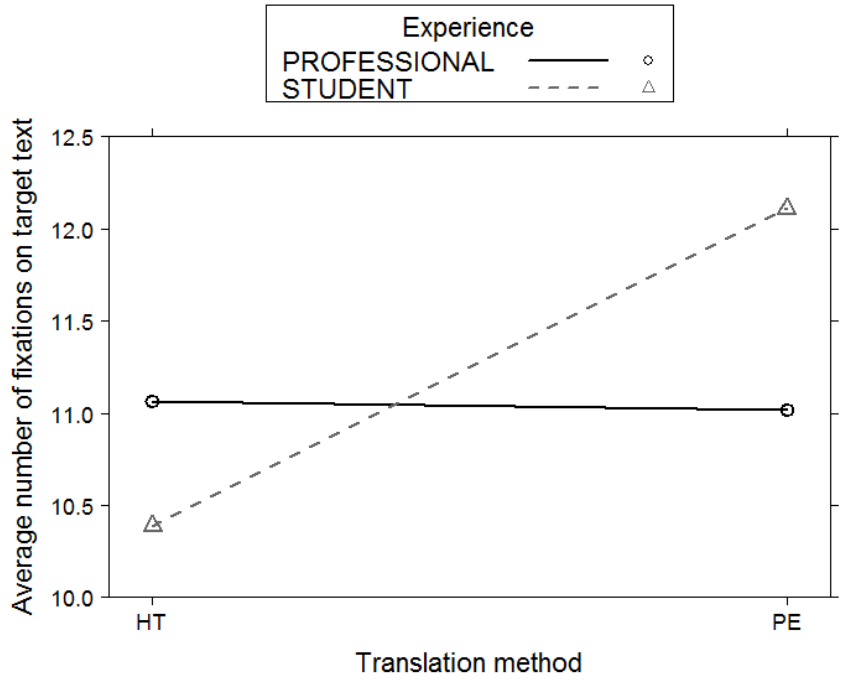


Figure 30 Effect plot of interaction effect between method and experience for the average number of fixations on the target text.

There is a higher number of fixations on the target text when post-editing compared to human translation, but only for the students (Figure 30). The number of fixations on the target text for professional translators seems to be comparable for both methods of translation. We also looked at the average duration of fixations on the target text. The model with fixed effects performed better than the null model, yet only method was found to be a significant predictor, with average fixation duration being 5 milliseconds shorter when post-editing compared to human translation.

5.1.3 External resources

On average, participants spent 17 per cent of their total translation or post-editing time in external resources (students HT=0.19; students PE=0.16; professionals HT=0.17; professionals PE=0.16), the full distribution can be seen in Figure 31.

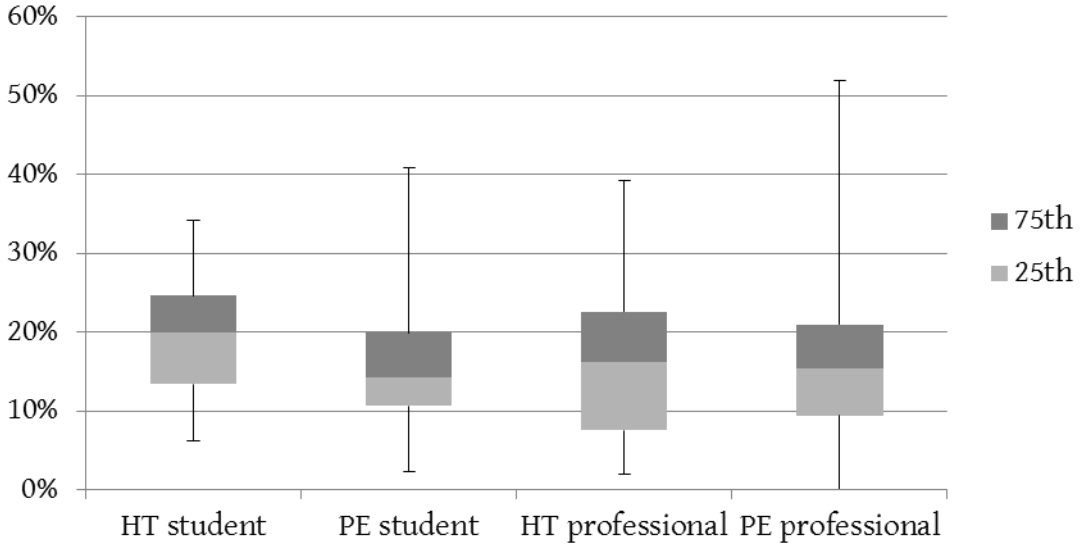


Figure 31 Distribution of percentage of time spent in external resources.

To observe external resource behaviour of the translators, we coded the information from *Inputlog*. Each consultation was labelled with the relevant category: dictionary, concordancer, search, encyclopedia, MT or 'other' (grammar or spelling websites, fora, news sites, term banks and synonym sites). We added the numbers of times each type of resource was consulted as well as the time spent in each type of resource to the SG-data file and calculated the average number of external resources consulted per source token, as well as the average time spent in external resources per source token. We fitted a mixed effect model with total time spent in external resources as dependent variable, but this model did not outperform the null model. Using only translation method or experience as a predictor did not lead to better results. The model predicting the total number of source hits (number of times an external resource was consulted)

seemed somewhat more promising. However, in the full model, as in the pretest, none of the predictors was shown to have a significant impact. The impact of method was almost significant, but not sufficiently so as to justify the model. Removing experience as a predictor further improved the model. In this model, method was a significant predictor, with post-editing reducing the number of external resources consulted per word.

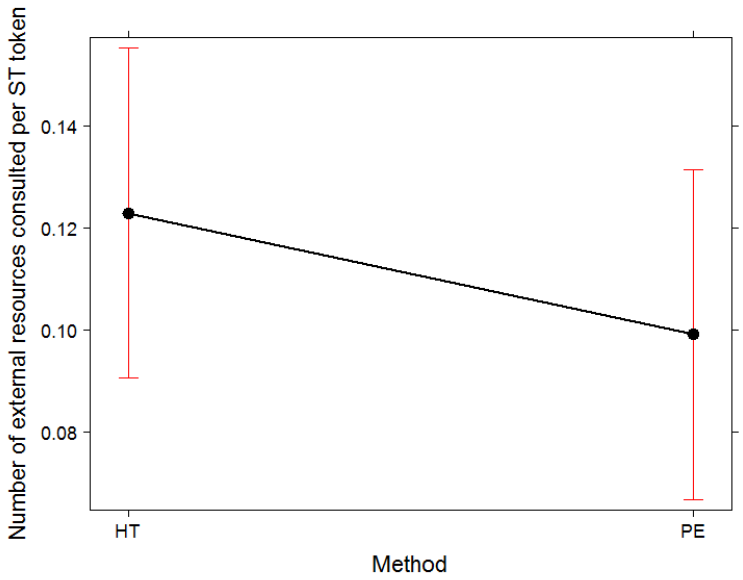


Figure 32 Effect plot of method on the average number of external resources consulted per ST token.

As the total time spent in external resources did not significantly differ between groups or translation methods, we looked at the usage of external resources in more detail. Figure 33 gives an overview of the percentage of overall time spent in external resources for each type of resource and reveals that for both groups of participants and both methods, *Google Search*, concordancers and dictionaries are the most common resources. It can be seen, however, that, everything taken together, students rely more heavily on dictionaries than professional translators. Professional translators seem to spend somewhat more time in MT than students, even when post-editing, which seems counterintuitive at first. From the surveys, however, we learned that *Google Translate* is often used to check the translation of a single word and to get alternative translations. Students consulted synonym websites rather than *Google Translate* when looking for alternative translations.

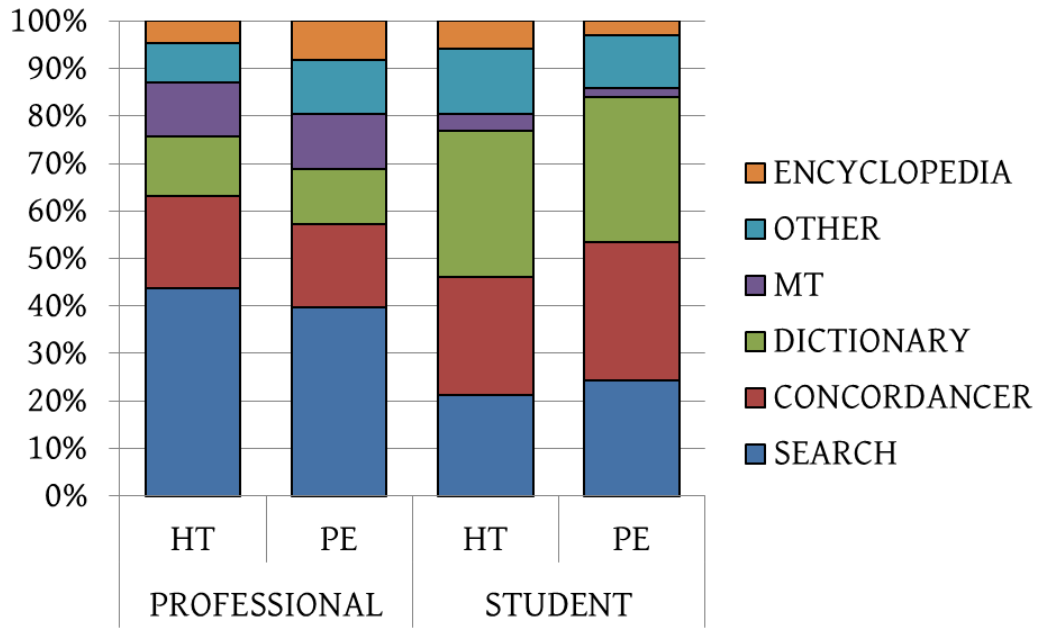


Figure 33 Percentage of total time spent in external resources per resource type for both methods and levels of experience.

To statistically verify these assumptions, we built a mixed model for each of the types of external resources (search, concordancer, dictionary, MT, encyclopedia, and other), with the average time spent in the external resource per source text word as dependent variable, experience and method plus interaction effect as independent variables, and - as usual - participant and sentence identification codes as random factors. None of the models with independent variables outperformed the null models, with the exception of the model with the time spent in dictionaries as dependent variable. While there was no statistically significant effect of task, there was a statistically significant effect for experience, with students spending more time in dictionary searches than professional translators. The effect plot, which can be seen in Figure 34, further shows a trend for students to spend less time in dictionaries during post-editing, but this effect was not significant.

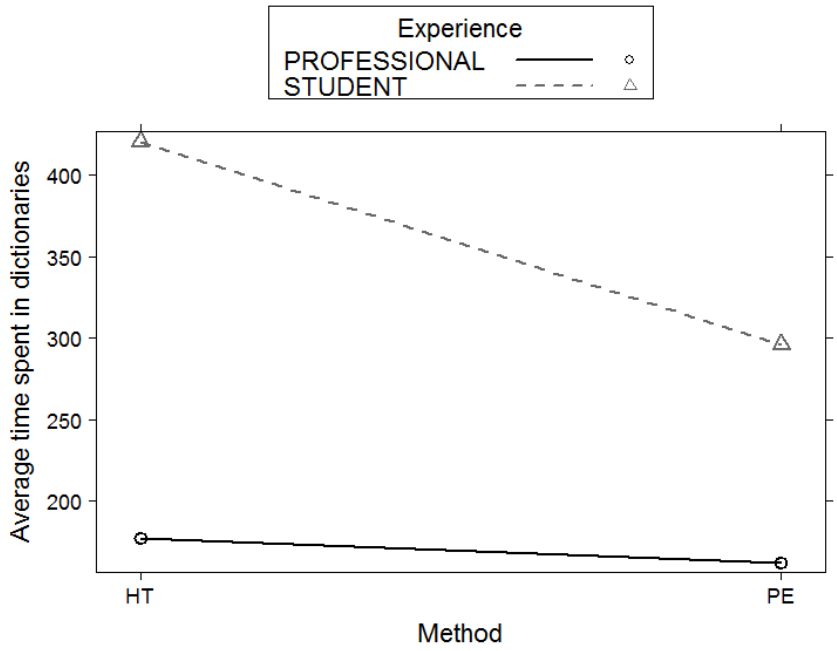


Figure 34 Effect plot of interaction effect between method and experience for the average time spent in dictionaries (in ms).

An investigation into the types of resources used within each category revealed that students used both the *Glosbe* concordancer and *Linguee*, whereas professional translators only used *Linguee*. In total, twenty-two different types of dictionaries were consulted across all participants. Six of those were consulted only by students, whereas nine of those were only consulted by professional translators. The dictionary most commonly used by all participants is *Van Dale*, a classic dictionary for the Dutch language. *Van Dale* was used more frequently than all other dictionaries combined. We also know the language of the search queries, but this seems fairly comparable across groups: 76% of the professional translators' queries in *Van Dale* were English (the source language), the others were Dutch (the target language), compared to 82% of search queries within the student group.

5.1.4 Impact of external resources on productivity

There are two conceivable ways in which the usage of external resources affects productivity. On the one hand, we can expect the total translation time to increase when a person spends more time in external resources, on the other hand, it is possible that the time spent in external resources decreases the overall time needed to translate a text, as a translator looks up external resources to solve problems. Previous research

has shown the first option to be true, with the use of reference materials slowing down the translation process (Luukkainen (1996), as cited in Raído (2014)), which we also confirmed in a previous analysis on the student subset of our data[2]. We extended our earlier work with the professional data, which of course introduced an additional factor. Where we only had a two-way interaction before, we now have a three-way interaction between the time spent in external resources, the translation method (human translation or post-editing) and participant experience (professional or student).

Regarding translation speed, we found significant differences between human translation and post-editing, but no additional effect of experience (section 5.1.1); regarding the total time spent in external resources we found no significant effects for either method or experience (section 5.1.3). In this section, we study the relationship between time spent in external resources and the total time needed to translate the entire text.

As in previous analyses, we began by building a null model with participant codes and sentence codes as random factors. In this case, the average total time per source text word was the dependent variable. We then built a second model containing the total time in external resources, translation method, and experience as possible predictor variables, plus interaction effect. The second model provided a significantly better fit than the null model (reducing AIC from 26959 to 26455). The model summary shows almost all predictors and effects to be significant, with the exception of experience and translation method in isolation (Table 14).

Table 14 Model summary of time in external resources, method, and experience, plus interaction effect predicting total time.

fixed effects	estimate	standard error	p
time in ext resource (DurER)	0.73	0.05	<0.001
experience-student (XPStud)	-57.85	509.64	0.91
method - post-editing (MPE)	-382.04	207.75	0.066
DurER:XPStud	0.54	0.1	<0.001
DurER:MPE	-0.35	0.05	<0.001
XPStud:MPE	869.81	326.81	0.008
DurER:XPStud:MPE	-0.44	0.13	<0.001

This shows that spending more time in external resources indeed leads to spending more time overall (model with DurER as predictor, first row). This effect is greater for

[2] The analyses presented in 5.1.4 and 5.3 are an extension of some of the analyses published as Daems, Carl, Vandepitte, Hartsuiker, and Macken (2016).

students than for professionals (model with DurER:XPStud as predictor, fourth row), and smaller during post-editing compared to human translation (model with DurER:MPE as predictor, fifth row). The three-way interaction effect (model with DurER:XPStud:MPE, last row) shows that, while the impact of spending more time in external resources on total time is greater for students (DurER:XPStud), this is less so the case when post-editing compared to when translating. All effects are visualised in Figure 35.

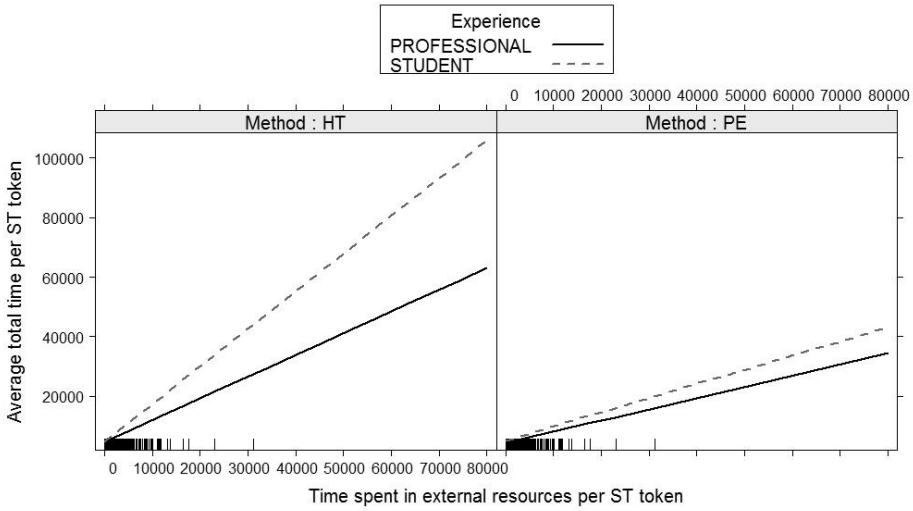


Figure 35 Effect plot of relationship between time spent in external resources normalised per ST token and total time normalised per ST token (both in ms).

5.2 Product analysis

We used the fine-grained translation quality assessment approach developed in the pretests to determine the final quality of the product. All final texts (human translations and post-edited machine translations) were annotated for acceptability (target text, language and audience) and adequacy (correspondence to the source text) issues by two annotators (the same annotators from the pretests), using the *brat rapid annotation tool* (Stenetorp et al., 2012). As the pretest showed a consolidation step to be a crucial step in the analysis, the error classifications were discussed by both annotators, and only the annotations both annotators agreed on were retained for the final analysis. Each error type received an error weight, corresponding to the severity of the error.

5.2.1 Overall quality

We fitted a linear mixed effects model with average total error weight per word as dependent variable, but the model with predictor variable did not outperform the null model, although the predictor experience was almost significant, with students performing somewhat worse than professional translators.

Specifically for the professional translators, we wanted to verify the assumption that translators specialised in the translation of general texts outperformed translators who did not specialise in general text translation. We looked at the number of errors each professional translator made, and compared that with the survey data: their years of professional experience and their level of specialisation for the current text type, i.e., percentage of their time translating general texts (Figure 36).

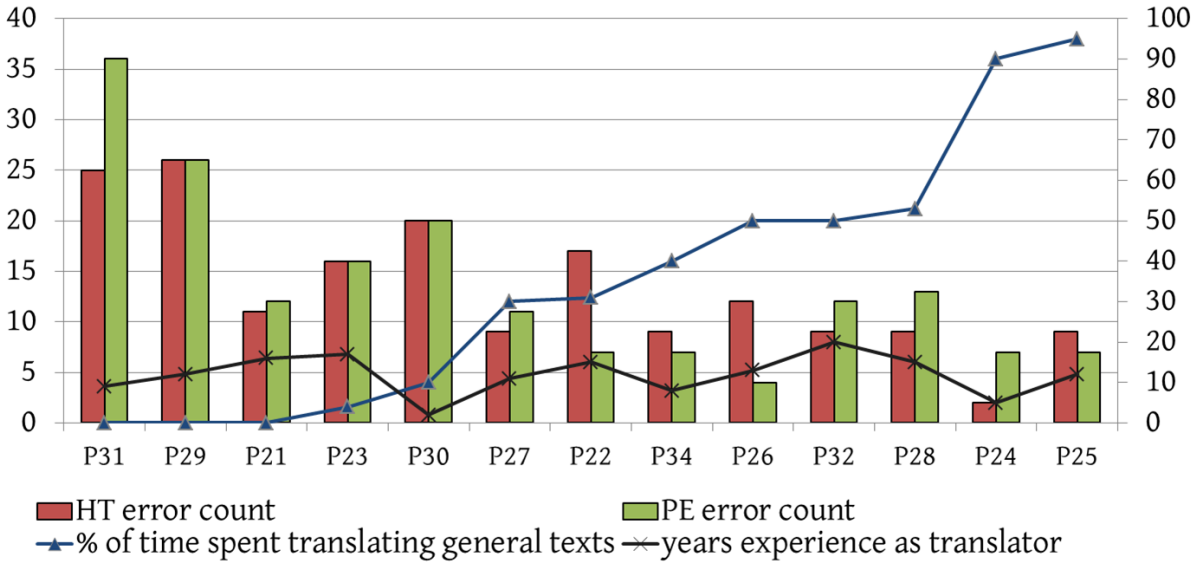


Figure 36 Relationship between professional translators' level of specialisation (percentage of time spent translating general text types, plotted on secondary axis), their translation experience (years, plotted on secondary axis), and the total error count for their human translation and post-editing tasks (plotted on primary axis). Labels on x-axis are participant codes

While the number of years of professional experience is not correlated with quality (r=-0.08, p=0.79 for HT; r=-0.17, p=0.57 for PE), there is a negative correlation between level of specialisation and number of errors (r=-0.76, p=0.003 for HT; r=-0.66, p=0.01 for PE), with participants 24 and 25 - spending respectively 90% and 95% of their time translating general texts - producing the highest quality translations.

On a more fine-grained level, we fit a model with average acceptability error weight per word as dependent variable, and one with average adequacy error weight per word as dependent variable. For acceptability, the results were comparable to those of the total error weight: the predictor model did not outperform the null model. For adequacy, however, the predictor model outperformed the null model when only

experience was taken as a predictor variable, not when both task and experience were used as predictor variables. The effect plot can be seen in Figure 37, showing that students' translations suffered significantly more from adequacy issues than professional translators (as represented by the higher average adequacy error weight).

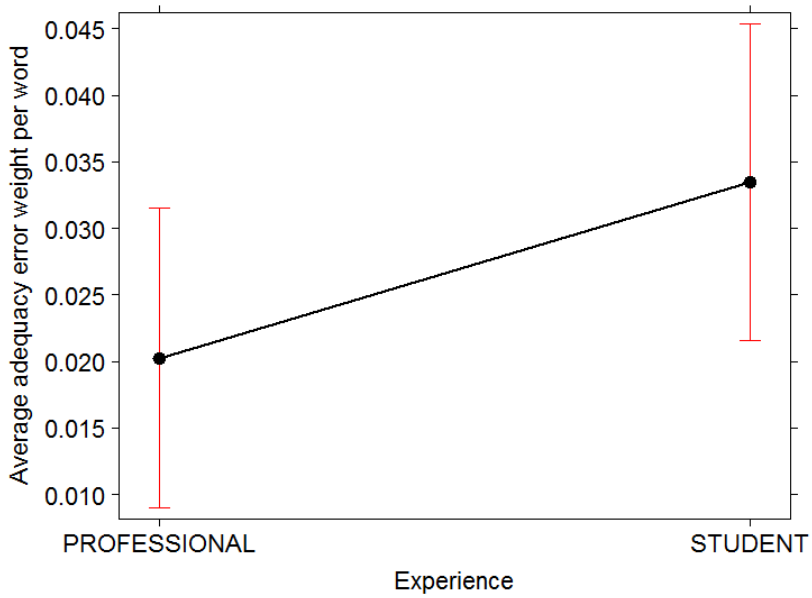


Figure 37 Effect plot of experience on the average adequacy error weight per ST token.

5.2.2 Main categories

On an even more fine-grained level, we distinguish between adequacy issues and various types of acceptability issues (grammar & syntax, coherence, lexicon, style & register, and spelling & typos). Figure 38 shows the percentage of all errors made for the main error categories, for both methods of translation and both groups of participants.

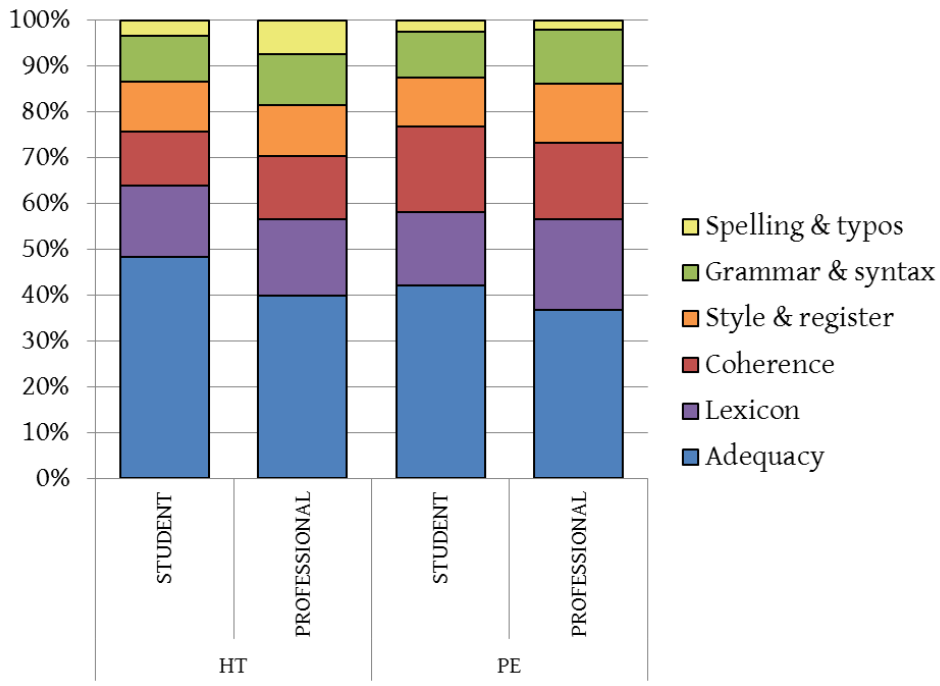


Figure 38 Occurrence of main error types for both methods and levels of experience.

What can be derived from Figure 38 is that, as was the case for the pretests, the most common error category are adequacy issues, and they are less common for post-editing than for human translation. For professional translators, spelling and typos are far less common in post-editing than for human translation. Coherence is somewhat more problematic for post-editing compared to human translation in both groups.

5.2.3 Common errors

On the most fine-grained level, as we did in the pretests, we can compare the most common error categories, in particular the ones that account for minimum 5% of all errors made.

Figure 39 shows the most common error types for human translations. As in the pretest on newspaper articles, meaning shifts, deletions, word sense issues, and wrong collocations are common problems; although deletions are much more common for students than for professional translators, whereas the opposite holds true for word sense issues. In contrast with the pretest on newspaper articles, logical problems are very common for students and professional translators alike. It was the second most common error in the pretest on technical texts, but that does not explain its abundance in the present experiment on newspaper articles. Also in contrast with the pretest, punctuation errors, which, in the pretest, accounted for 8% of all errors made, were not found in the final products of either students or professional translators.

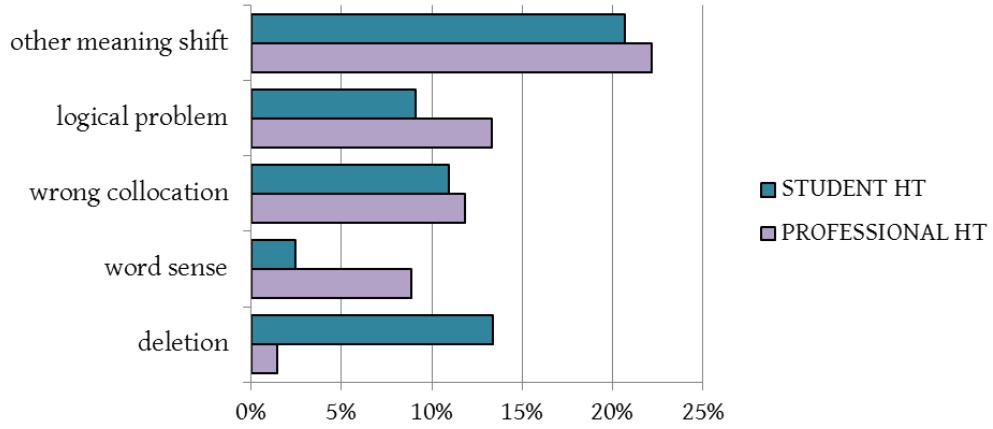


Figure 39 Overview of HT errors accounting for at least 5% of all errors made by either students or professional translators.

Some examples of the different problems can be seen below.

(9) ST: Major League Baseball
HT: de hoogste Amerikaanse basketbalklasse (other meaning shift, 'baseball' was translated as 'basketball')

(10) ST: A new exhibit of Hockney's work, including many iPad images...
HT: een nieuwe tentoonstelling ... met werken van Hockney, waaronder veel iPad-tekeningen ... (= logical problem; 'images' are described as paintings in the text, not drawings)

(11) ST: ... rejecting candidates who fail the tests.
HT: ...afwijst die falen voor de test (= wrong collocation, correct version would be 'niet voor de test slagen')

(12) ST: Cheering families ... donned shades as the sun crept from behind a cloud.
HT: Vrolijke gezinnen ... maakten schaduwen van zodra de zon van achter de wolken kwam. (= word sense error, 'shades' is used as 'sunglasses', not 'shadow')

(13) ST: Residents have to catch a cable car to the top of a nearby precipice...
HT: Inwoners moeten een kabellift nemen naar een nabije top... (= deletion, 'precipice' is not present in the translation).

For post-editing (Figure 40), three of the most common error categories (meaning shifts, wrong collocations, and word sense problems) are the same as those from the pretest on newspaper articles. The other common errors from the pretest (typos, compounds, and punctuation) have been replaced by logical problems, misplaced words, and deletions, the latter in particular for the students.

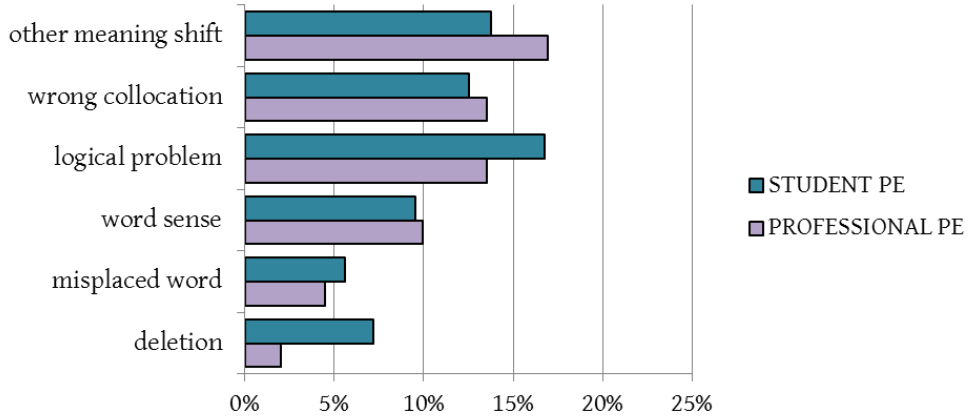


Figure 40 Overview of PE errors accounting for at least 5% of all errors made by either students or professional translators.

Some examples to illustrate the most common error types:

(14) ST: The volunteers keep us company.
MT: De vrijwilligers houden ons bedrijf
PE: De vrijwilligers houden ons bezig. (= other meaning shift, 'keep company' means 'to stay with someone', not necessarily 'keep them entertained')

(15) ST: They have pulled hundreds of billions of dollars out of the stock market...
MT: Ze hebben honderden miljarden dollars getrokken uit de aandelenmarkt...
PE: Ze hebben honderden miljarden dollars uit de aandelenmarkt getrokken (= wrong collocation, 'getrokken' in this context is a too literal translation in Dutch, and you cannot use it in combination with stock markets. Possible alternative is 'weggehaald'.)

(16) ST: ...until he began printing his digital images a few years ago.
MT: ... totdat hij begon te drukken zijn digitale foto's een paar jaar geleden
PE: ...totdat hij zijn digitale foto's een paar jaar geleden begon af te drukken (= logical problem, the 'images' are described as paintings, not pictures)

(17) ST: ...episodes of personal violence...
MT: ... afleveringen van persoonlijk geweld...
PE: ... afleveringen van persoonlijk geweld (= wrong word sense, 'episode' is not meant as an 'episode of a TV series', but as 'occurrences of an event')

(18) ST: Lie-detector test comes under fire as FBI hiring tool
MT: Leugendetectortest komt onder vuur als FBI inhuren hulpmiddel
PE: Leugendetectortest komt onder vuur te staan als ingehuurd hulpmiddel door de FBI (= meaning shift caused by misplaced word: 'hiring' has been translated as 'a tool hired by the FBI', whereas it is a tool that is used in the hiring process, so the word 'hiring' modifies the wrong word in the translation)

(19) ST: This blue whale's earwax tells the story of its life and locale
MT: Oorsmeer Deze blauwe walvis vertelt het verhaal van zijn leven en locale
PE: Het oorsmeer van deze blauwe vinvis vertelt ons zijn levensverhaal (= deletion, 'locale' is an important aspect as the text focuses on the environment rather than the whale's personal history)

When comparing human translation with post-editing, meaning shifts make up a smaller portion of all errors made after post-editing than during human translation, although it is still one of the most common errors. Logical problems become more abundant after post-editing for students, but remain the same for professionals. There is a lower number of word sense errors in human translations than in post-edited translations, especially for the students. Deletions are less problematic for post-editing than for human translation, whereas misplaced word issues were only problematic for post-editing, possibly because these errors often still make sense from an acceptability perspective, causing the translators to read over them, although they do contain a shift in meaning from the source text.

5.3 Impact of external resources on quality

So far, we have looked at process and product variables in isolation. However, a high quality product is presumably the effect of a successful process, and, in particular, successful problem-solving strategies (Göpferich, 2009). Spending more time in external resources (and thus increasing the overall time needed, as we saw in section 5.1.4) can be justified if this extra time also brings about an increase in quality. As we found no significant differences in overall quality between human translation and post-editing or students and professionals (section 5.2.1), it will be interesting to see whether spending time in external resources has a positive or negative impact on final quality, a debated issue (Raído, 2014). Gerloff (1988) and Jääskeläinen (1996) found that participants with the more intense research strategies also produced the highest quality products, but, for example, Dancette (2007) did not find a significant correlation between dictionary usage and translation quality.

We fit a linear mixed effects model to analyse the relationship between overall error weight normalised per ST token and the normalised total time spent in external resources. Normalised total error score was the dependent variable, method, experience, and time spent in external resources with interaction were added as predictor variables and participant and sentence codes were added as random effects. This model performed significantly better than the null model without predictors, reducing AIC from -2805.6 to -2827.9. The model summary, however, shows only the three-way interaction effect (DurER:XPStud:MPE) to be significant (Table 15).

Table 15 Model summary of time in external resources, method, and experience, plus interaction effect predicting total error weight.

fixed effects	estimate	standard error	p
time in ext resource (DurER)	4.31E-08	1.58E-06	0.98
experience-student (XPStud)	1.35E-02	1.06E-02	0.21
method - post-editing (MPE)	3.71E-03	6.7E-03	0.58
DurER:XPStud	3.7E-06	3.4E-06	0.28
DurER:MPE	5.62E-08	1.76E-06	0.97
XPStud:MPE	-1.42E-02	1.06E-02	0.18
DurER:XPStud:MPE	1.12E-05	4.32E-06	<0.01

The main finding here is that, although the impact of spending more time in external resources on the total error weight for students compared to professionals was not significant (model with DurER:XPStud as predictor, fourth row), the impact of this effect for post-editing is significantly higher than that for human translation (model with DurER:XPStud:MPE as predictor, last row). This effect is visualised in Figure 41.

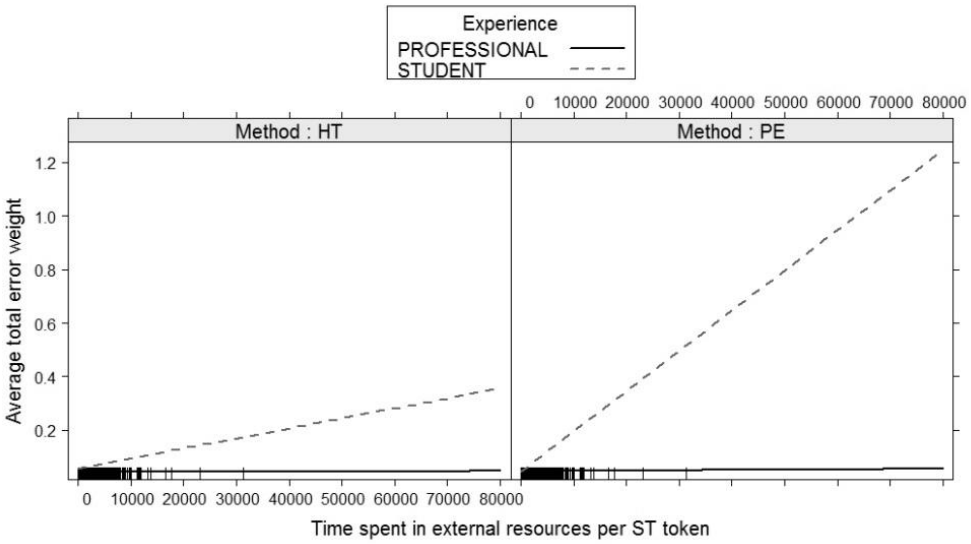


Figure 41 Effect plot of the predicted relationship between time spent in external resources normalised per ST token and overall error weight normalised per ST token, for both translation methods and levels of experience.

The differences in slope indicate a difference in the effect of consulting external resources for both types of tasks for students in particular. In the case of post-editing, spending a longer time in external resources does not lead to an increase in quality, but rather a decrease (visualised by higher average total error weight), indicating that the resource consulting strategies are not successful. Findings for professionals for both tasks, and students translating from scratch were not significant, indicating that here, consulting more external resources does not negatively or positively impact the final quality of the product.

On a more fine-grained level, we repeated the analysis in two different analyses, the first with average total acceptability error weight as dependent variable, and the second with average total adequacy error weight as dependent variable. For acceptability, the model with predictors did not outperform the null model (increasing AIC from -3501 to -3492). For adequacy, the model with predictors did perform significantly better (decreasing AIC from -3579 to -3606). The model summary was comparable to that of the total average error weight as dependent variable, with only the three-way interaction effect being significant (in this model, p=0.001). The effect plot can be seen in Figure 42.

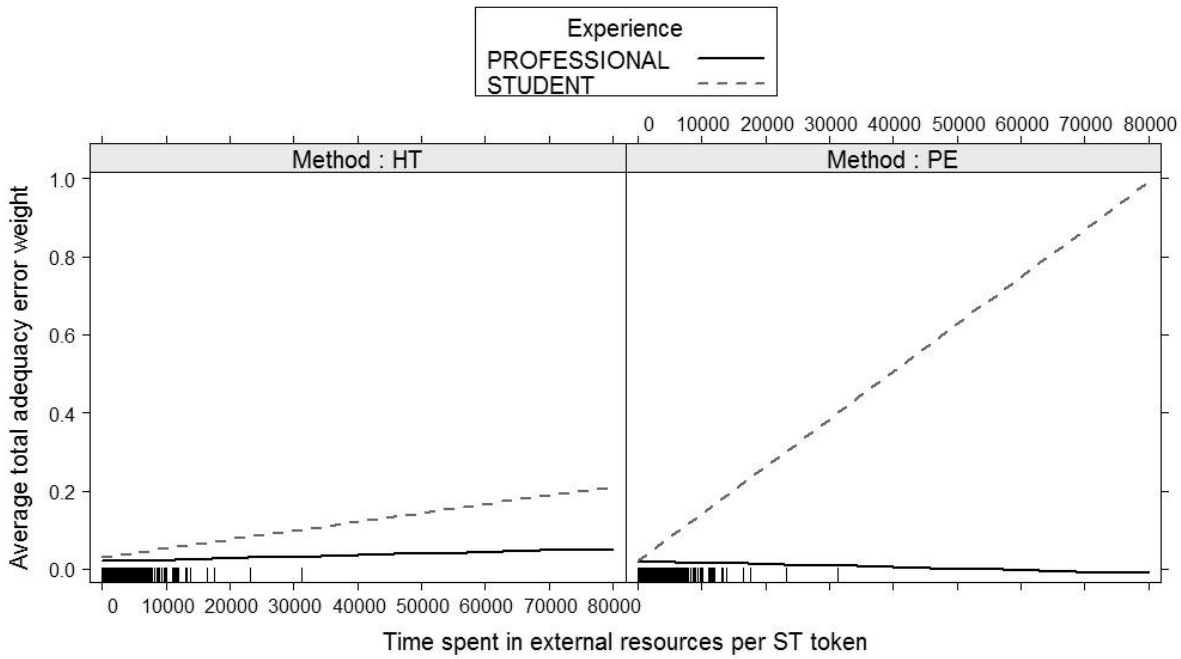


Figure 42 Effect plot of the predicted relationship between time spent in external resources normalised per ST token and adequacy error score normalised per ST token.

As we found a significant difference between students and professional translators for the usage of dictionaries (section 5.1.3) and a high usage of dictionaries has - in some contexts - been shown to correspond to higher quality products (Raído, 2014), we performed an additional analysis with time spent in dictionaries as a predictor variable, rather than total time spent in external resources. However, these models never outperformed the null models, with AIC increasing from -2805 to -2797 when predicting the total error weight, from -3500 to -3491 when predicting acceptability error weight, and from -3579 to -3570 when predicting adequacy error weight.

5.4 Attitude analysis

All participants (10 students, 13 professional translators) filled out a survey before and after the experiment, using *Qualtrics*. In contrast with the pretests (where surveys were filled out on paper), this made it easier to filter our data and create reports afterwards.

About half of the students as well as the professionals indicated that they had some experience with post-editing (question: 'I ... make use of MT systems while translating'; options: 'never, sometimes, often, always'). Their additional comments, however, showed that they often considered post-editing to be 'working with a translation tool', including editing TM matches as well as MT output. Therefore, the opinions on post-editing taken from the pre-test survey may encompass issues related to the usage of translation tools in general, in addition to post-editing. Some example answers to the open question 'how and why do you use MT for your translations' can be seen below (participants answered in Dutch, translations are our own).

(20) I mostly use it to make sure the terminology remains consistent and to make it easier to find previously translated segments.
(21) To translate faster, to make a translation memory, and to save common terminology.
(22) The translation memory helps me to save time and to work more accurately. The spellchecker is also very useful.

We first wanted to find out how rewarding participants perceived post-editing to be compared to human translation. Participants who never used MT could choose from hypothetical options ('I think I would find...'), participants who had used MT could choose from actual answers ('I find...').

Table 16 Most rewarding translation method.

	HT & PE equally rewarding	HT more rewarding, doesn't mind PE	HT most rewarding
professionals			
always uses MT	0	1	0
often uses MT	0	2	1
sometimes uses MT	2	2	2
never uses MT	0	1	2
students			
often uses MT	2	0	0
sometimes uses MT	1	4	1
never uses MT	0	1	1

Table 16 shows that, like in the pretest, most students who claimed to have some experience with post-editing found it equally rewarding as human translation, or

preferred human translation to a small degree. None of the translators found post-editing more rewarding. Of the professional translators with knowledge of post-editing, eleven translators found human translation more rewarding, although six of them did not mind post-editing. The open question that followed this question showed that professionals had different feelings about post-editing: those who enjoyed it mostly enjoyed not having to start from scratch, and noted that it could save them some time, provided the output was of sufficient quality. Those who preferred regular translation mentioned creativity and freedom as an important factor, and they did not believe that post-editing would necessarily save time. One translator explicitly mentioned the reduced per-word fee as a reason not to prefer post-editing. The examples below show some of the answers.

(23) If the MT output is good, I don't mind post-editing. I assume translation systems will only improve, and I may as well be prepared to accept that post-editing will become an increasingly important part of my job.

(24) Translation is a form of art, a creation. Post-editing looks like boring and unrewarding work.

Next, we asked the participants who had used MT before how useful they found MT output. A summary of their answers in relationship to how rewarding they found post-editing can be seen in Table 17.

Table 17 Usefulness of MT output.

	MT often useful	MT sometimes useful
professionals		
HT & PE equally rewarding	0	2
HT more rewarding	1	4
HT most rewarding	1	2
students		
HT & PE equally rewarding	2	1
HT more rewarding	3	1
HT most rewarding	0	1

Both students and professionals found MT output 'often' or 'sometimes' useful, with students being more inclined to choose 'often'. For students, the selection of 'often' seems to correspond somewhat with how rewarding they perceive post-editing to be, but this is not the case for professionals. Most participants added some clarifications to their answers, of which a few examples can be seen below.

(25) The quality of the output often surprises me. On the other hand, it's somewhat bizarre that translation tools don't seem to 'learn' anything as you go along and they keep making the same mistakes.

(26) It depends greatly on the text type and language combination, and I can only speak for *Google Translate*, as that is the only system I know. I can imagine the output of other systems could be much better.

Table 18 Perceived speed of both translation methods.

	HT is faster	HT is as fast as PE	PE is faster
professionals			
HT & PE equally rewarding	0	0	2
HT more rewarding	2	3	1
HT most rewarding	3	2	0
students			
HT & PE equally rewarding	0	0	3
HT more rewarding	1	2	2
HT most rewarding	1	1	0

Regarding speed (Table 18), half the number of students expected post-editing to be faster than regular translation, compared to only three out of thirteen professionals. There seems to be some correlation between finding post-editing rewarding and believing it is also faster, with none of the participants who found post-editing as rewarding as human translation indicating that they believe human translation to be faster, and none of the participants who found human translation most rewarding indicating that post-editing is faster. In the comments, the participants with no post-editing experience clarified that their answers were based on guesses; the others provided some different feedback.

(27) Post-editing is sometimes faster, but the result is not as pretty. I therefore only use it for business texts, not for editorial texts. (...) it inhibits creativity.

(28) I need to change too many things, compare the target text to the source text, check where the MT system went wrong, etc. I prefer translating from scratch.

(29) I think the speed is comparable, as you still need to check everything, reorder and rewrite things, which does not necessarily make it faster, but also not slower. I do expect it to depend on the text type.

In addition to speed, we are of course also interested in the perceived quality of the post-edited texts. The students in the pretest did not seem convinced of post-editing

quality in comparison to human translation quality. Table 19 shows the results for the main experiment.

Table 19 Perceived quality of both translation methods.

	HT quality is better	PE quality is equal to HT quality	PE quality is better
professionals			
HT & PE equally rewarding	0	1	1
HT more rewarding	0	6	0
HT most rewarding	3	1	1
students			
HT & PE equally rewarding	1	1	1
HT more rewarding	3	2	0
HT most rewarding	1	1	0

From the students and professionals who claimed to have some knowledge of post-editing, only two (one student and one professional) believed they produced better quality with post-editing. From the participants without post-editing experience, only one professional translator expected this to be the case. Expectations of PE quality seem to improve somewhat when participants do not consider HT to be most rewarding, especially with the professional translators. In both groups, only one participant expected PE quality to be better. The quality concerns listed explicitly were comparable among groups: a product can contain non-idiomatic expressions because of post-editing, and it is harder to control for consistency when post-editing than translating. The following examples illustrate some of the concerns and quality remarks.

(30) Being influenced by the output has two consequences: on the one hand, the terms can be very useful, on the other, it can cause you to translate source text constructions too literally.

(31) The sentences would be less fluent.

(32) Translation is, and I think it will always be, a craft. I always need to revise a final text by rereading it on paper. I don't think it will matter much whether I did the preparatory work manually or via post-editing.

In the survey taken after the experiment, we asked participants about their preferred translation method for the text type, their perceived speed, and what they thought was the least tiring translation method.

Most participants, students and professionals alike, preferred human translation over post-editing (Table 20) for this particular text type. Four professionals and one student

preferred post-editing. Of these five participants, four had indicated in the pretest that they often used machine translation.

Table 20 Preference of translation method after experience.

	prefers HT	no preference	prefers PE
professionals	7	2	4
students	8	1	1

Some additional feedback examples to clarify some of the preferences follow below.

(33) I prefer being able to choose my own structures and I find it hard to let go of the structures offered by post-editing.

(34) I like receiving a suggestion to start working from. I would not have come up with some of the words myself, but they do work well.

Regarding speed, six students and five professionals were convinced that post-editing was faster, compared to only one student and two professionals who believed human translation was faster. The remaining three students and six professionals did not perceive a difference in speed between both methods of translation.

Table 21 Perceived speed of both methods before and after the experiment.

	HT fastest after	HT & PE equally fast after	PE fastest after
professionals			
HT fastest	1	2	2
HT & PE equally fast	1	3	1
PE fastest	0	1	2
students			
HT fastest	0	0	2
HT & PE equally fast	1	1	1
PE fastest	0	2	3

For six professional translators and four students, the perception of speed did not change after the experiment (Table 21). In total, five participants changed their minds in favour of human translation, and eight participants believed post-editing to be faster than they thought before the experiment. Some arguments the participants added can be seen below.

(35) Sometimes I only needed to change a few words or the word order. I also needed to look up fewer words. Even if I know a translation, I tend to look up the translation to be sure. This is not necessary with post-editing because I can check my translation against the presented output. Of course I still look words up when in doubt.
(36) It is more motivating if part of the work is already done.

The question about which translation method participants considered to be the most tiring was included since we are interested in the perceived cognitive load. Responses for the professional translators varied, with a comparable number of participants choosing each of the three options (HT less tiring, PE less tiring, both equally tiring). The result is slightly different for the students, with only one student considering HT to be less tiring, and the others selecting PE or 'equally tiring'.

Table 22 Perception of how tiring both translation methods are.

	HT least tiring	HT & PE equally tiring	PE least tiring
professionals	5	4	4
students	1	5	4

It is interesting to see that the students' perceptions correspond to the fixation analysis which showed that post-editing was cognitively less demanding than human translation. In a follow-up question, we asked participants what they felt made both methods of translation tiring. Some of the responses can be seen below. While most participants focussed on the translation methods, some commented on the experimental environment.

(37) Human translation: having to start from scratch, post-editing: filtering out the errors, I tended to keep incorrect constructions, while it was often better to start from scratch
(38) During post-editing: looking for small inconspicuous errors (agreement issues, singular/plural, two different translations for the same term, etc.). Rereading the text two or even three times was no unnecessary luxury.
(39) Having to sit still for the eye tracker. Not having a decent (electronic) dictionary.

In some final follow-up questions, we asked participants what they found useful about the machine translation output and what could have been better. Most participants agreed that MT output was especially useful for terminology and vocabulary, and for short, unambiguous sentences. Things that could have gone better, according to the

participants, were too literal translations, incorrect translations of ambiguous words or sentences, and grammatical issues.

The very last question of the survey was reserved for general comments and feedback on the experiment in its entirety (see examples below).

(40) Post-editing was better than anticipated.

(41) Before the experiment, I was very sceptic about all forms of automated translation. I thought it wouldn't help much, that it would push me in a certain direction, that it would be more of a hindrance than a help. But it was better than anticipated. There were some useful, well found suggestions, and in the end, I still did what I wanted with the text. I don't think it was faster than regular translation, automatic translation is of too low quality to be faster, but it is useful for suggestions...

(42) I had no previous experience with post-editing and I have to conclude that it is useful and could save time.

(43) In general, everything went smoothly, and it felt like a translator's average working day, so it was not unrealistic.

5.5 Discussion

Overall, we found more significant differences between human translation and post-editing than between students and professional translators. This was the case for speed, average total fixation duration, number of fixations on the source text, average fixation duration on the target text, number of external resources consulted, and the number of dictionary searches. The effect of experience did become significant for the average fixation duration on the source text, the number of fixations on the target text, adequacy error weight, and the impact of external resources on speed and quality.

While post-editing has been shown to be faster than human-translation for technical texts (Plitt & Masselot, 2010), we now also found it to be statistically significantly faster for general text types. There was no significant difference in processing speed between students and professionals. Though in contrast with Tirkkonen-Condit (1990), this finding is in line with Jääskeläinen's observation (1996) that professional translators do not necessarily translate faster than students.

The fixation analysis has shown that post-editing overall is less cognitively demanding than human translation (which is in line with O'Brien (2007)) for professional translators and students alike. When processing the source text, students benefit more from the post-editing condition than professional translators, whereas when processing the target text, post-editing seems to be less cognitively demanding for both groups, although professional translators process the target text differently, with students requiring fewer fixations when translating from scratch compared to

post-editing and professional translators requiring a comparable number of fixations for both methods. Further analysis of the actual text production and final translations is needed to get a better idea of what is really happening, and whether or not what is happening is successful. This knowledge can then be used to better train students or provide feedback to professionals. Perhaps the professional translators treat post-editing more as a regular translation task, or they know how to move through a text more efficiently than students, considering they have more experience with spotting and solving translation issues.

There was no significant difference for the overall time spent in external resources by students or professionals, during human translation or post-editing, although the impact on overall speed and quality for both was different. Students relied significantly more on dictionaries than professionals, in line with findings by A. Jensen (1999) that the usage of dictionaries decreases with experience. This goes to show that external resources are crucial for students and professionals alike, independent of translation task. Integrating the most often consulted resources (dictionaries, concordancer, *Google Search*) into a translation tool could perhaps save translators some time. Seeing how professional translators depend less on dictionaries than students, perhaps translator training needs to focus more on translation as a higher-order task and not just the lexical aspects of translation. In addition, if the need to consult a dictionary arises from lack of confidence in one's own lexical abilities, perhaps attention could be given to vocabulary development as well.

Supporting the findings by Jääskeläinen and Tirkkonen-Condit (1991); Kiraly (1995), and Jääskeläinen (1996), we found that the more experienced translators are not necessarily the more successful translators, with students producing products of comparable overall quality, at least in this particular experiment. There seems to be no statistically significant difference in quality between human translation and post-editing either, which confirms that post-editing can produce texts that are as least as good as human translations (Garcia, 2011). The detailed analysis did provide some additional information. Students seem to struggle with meaning shifts and logical problems. This is in line with findings by Séguinot (1991), who characterized structure, cohesion and register as advanced translation issues, and might in part be explained by the fact that students treat translation as a linguistic task (Tirkkonen-Condit, 1990). We further found that professional translators specialised in the translation of general texts - the most common text type in the students' training - outperformed translators who do not specialise in general text translation (Jääskeläinen, 2010). We expect to see more significant differences between students and professionals for other types of specialisation: students encounter some specialist texts in their classes, but presumably not enough to outperform professionals with a few years of specialised experience.

Given that our student participants are students with experience in translation, but not in post-editing, it can be assumed that they have developed successful resource

consulting strategies when translating throughout their studies, whereas post-editing is a new type of translation, giving rise to different problems, questions, and strategies, which are not always as successful as when translating. It is striking that we did not find the same effect for post-editing for the professional translators, as most of the professionals in our study did not have much experience with post-editing either. We speculate that a possible explanation for these findings can be found in the machine translation (MT) quality. On the one hand, students might be too trusting of MT quality (as evidenced by the fact that a lower number of external resources is consulted when post-editing), on the other hand, they encounter very different problems when post-editing than when translating from scratch, making it hard to find the exact cause of a problem, and - in extension - to decide on the most appropriate external resources to consult. Perhaps the machine translation output primes certain - misguided - search strategies, leading to the students being unable to solve certain problems even when consulting external resources. The difference with professional translators would then be explained by the fact that professional translators in general are more mistrusting of machine translation output and the importance of confidence during translation (House, 2000; Kiraly, 1995), of which professional translators presumably have more than students (Fraser, 2000). Other factors might also provide different insights into the translation and post-editing process, such as translation styles (Carl, Dragsted, & Jakobsen, 2011) or translation patterns (Asadi & Séguinot, 2005) rather than experience.

From the surveys, we can tentatively conclude that student and professional translators hold similar opinions, and that preferences seem to be caused by individual differences rather than between group differences (in line with Guerberof (2013)). Both groups seem to prefer human translation, although they do not necessarily mind post-editing, and while they are not always convinced of post-editing quality, they mostly agree that post-editing is faster than human translation, especially after participating in the experiment. It might be a good idea to make translators more aware of the final quality of post-edited texts, seeing as they are not convinced of its quality, yet we found no significant difference with human translation quality. We can only spot a more obvious difference between students and professionals when considering the least tiring translation method. Professional translators experienced no obvious difference, whereas students seemed to consider post-editing the least tiring method of translation. This might be explained in part by the findings by Tirkkonen-Condit (1990) that non-professional participants treat translation as a linguistic task, and they rely mostly on dictionaries to solve problems. In a post-editing condition, lexical information is already provided by the MT output, which might reduce the need to look for additional information, and thus make the students experience the process as less tiring than regular human translation.

5.6 Conclusion

Our findings imply that post-editing is a viable alternative for human translation, even for general text types: it is faster without leading to lower quality results, and it is cognitively less demanding. The fact that the professionals did not obviously outperform students might mean that the current translation curriculum prepares students well for the translation of general texts, although students could benefit from more effective search strategies during post-editing, especially regarding adequacy issues. Looking at the benefits of post-editing and the fact that most participants weren't opposed to post-editing after participating, perhaps specific post-editing training could be added to the translation curriculum to make for an even better future generation of translators.

Chapter 6 Impact of MT quality on PE effort indicators[1]

Even though we found post-editing to be faster than human translation (Plitt & Masselot, 2010), and we confirmed that post-editing output is sometimes assessed more favourably than regular human translation (Garcia, 2010), machine translation systems are nowhere near perfect yet. With statistical machine translation systems being so unpredictable, Offersgaard, Povlsen, Almsten, and Maegaard (2008) argued that a post-editor does not only need the qualities of a good translator, but in addition needs to be able to quickly decide whether or not machine translation output can be used, or whether it would be faster to simply translate from scratch. Letting the post-editors make this decision, however, costs time and effort. Therefore, it would be much more cost-efficient if this step could be performed automatically, with a system capable of pre-assessing MT suitability (Schütz, 2008), a sentiment echoed by Denkowski and Lavie (2012):

> To avoid wasting translators' time, systems should be able to predict when MT output is sufficiently good to serve as a starting point for post-editing or sufficiently bad to require total re-translation and recommend accordingly.

However, it is largely unclear how post-editing effort should be defined and measured, which will be discussed in the following sections.

[1] Part of the work described here has been published as Daems, Vandepitte, Hartsuiker, and Macken (2015).

6.1 Related research

6.1.1 Assessing PE effort via product analysis

Often used automatic metrics like *BLEU* (Papineni et al., 2002) or *METEOR* (Banerjee & Lavie, 2005) compare MT output to reference translations to evaluate a machine translation system's performance. Whereas the values given by such metrics can be used to benchmark and improve MT systems, they are created on the basis of translations made independently from the MT output and thus do not necessarily provide post-editors with valid information about the effort that would be involved in post-editing the output. In addition, a score given by such metrics is an uninformative value that says nothing about the complexity of specific errors that need to be fixed during post-editing.

More recently, research into machine translation quality estimation (QE, sometimes also called confidence estimation, CE) has moved away from independent reference translations to reference post-edited sentences. Many QE systems have been trained on human-targeted translation error rate (HTER) data (Snover, Dorr, Schwartz, Micciulla, & Makhoul, 2006). HTER measures the edit distance between the MT output and a post-edited version of that MT output. The benefit of using HTER is the fact that it is relatively easy to apply, as it only requires MT output and a post-edited text, but HTER has some important limitations. The underlying assumptions when using HTER are that HTER is an indication of actual post-editing effort (Specia & Farzindar, 2010), which implies that all edits made to MT output by a human are expected to require a comparable amount of effort. However, HTER focuses on the final product, without taking the actual process into account, so its relationship to post-editing effort is questionable. In that sense, HTER is a little bit like a car's GPS without intelligent feedback: it provides you with the shortest way to get from point A to point B, but it does not take any obstructions or personal factors into account. For example, a post-editor can return to the same phrase multiple times during the post-editing process, changing that particular phrase each time, but settling on one specific solution in the end. HTER will indicate how different this final solution is from the original MT output, but it does not take into account all edits made during the process. In addition, the number of edits required to change an agreement issue could be comparable to or greater than the number of edits required to deal with a logical problem, although the first issue has a straightforward mechanical solution (which is assumed to be cognitively less demanding), while the second requires more effort to solve.

Taking these aspects into account, we hypothesise that process factors are a more accurate representation of actual post-editing effort than product factors such as HTER.

6.1.2 Assessing PE effort via process analysis

According to Krings (2001), there are three main types of post-editing effort. Of these three, the easiest to define and measure is temporal effort: how much time does a post-editor need to turn machine translation output into a high quality translation? The second type of post-editing effort is somewhat harder to measure, namely technical effort. Technical effort includes all physical actions required to post-edit a text, such as deletions and insertions. It can, for example, be measured by pauses and production units. Production units are editing events, i.e., continuous typing activity, separated from one another by pauses of at least one second. The final type of effort is cognitive effort. It refers to the mental processes and cognitive load in a translator's mind during post-editing, and can be measured by means of fixation data.

Yet, the distinction between temporal, technical, and cognitive effort is not a strict one, and post-editing effort indicators have been assigned to different categories, with cognitive effort being the overarching one. Koponen et al. (2012), for example, used a cognitively motivated MT error classification created by Temnikova (2010) and found evidence that post-editing time can also be an indication of cognitive effort. They further looked at a technical effort indicator - keystrokes - and its relationship to cognitive load, but here they found that keystrokes were influenced more by individual differences between participants than by cognitive load. We therefore decided to look for other technical effort indicators than keystrokes. Related to keystrokes are production units (sequences of coherent typing activity) and pauses (any interruption between keystrokes lasting longer than one second). O'Brien (2006) suggested pause ratio (total time in pauses divided by the total editing time) as a possible indicator of cognitive effort, but she did not find conclusive evidence for a relationship between both. Later, Lacruz, Shreve, and Angelone (2012) introduced the average pause ratio (the average time per pause in the segment divided by the average time per word in the segment) as an answer to O'Brien's pause ratio (2006), arguing that pause ratio is not sensitive enough as a measure for cognitive activity, as it does not take average pause length into account. We decided to include both pause measures in our study, to establish whether or not they can both be used, and whether or not they are indicators for different causes of effort. In addition, we will look at the number of production units as a possible effort predictor as well, as Lacruz et al. (2012) found a relationship between average pause ratio and the number of production units.

Effort indicators that seem to be exclusively related to cognitive post-editing effort are the average fixation duration and the number of fixations. Jakobsen and Jensen (2008) found longer average fixation durations and a higher number of fixations as the complexity of the task increased from reading to translation. Doherty and O'Brien (2009), however, found a higher number of fixations for bad MT output than for good MT output, but they did not find a significant difference between the average fixation

durations for both types. We will include both average fixation duration and number of fixations as potential cognitive post-editing effort indicators.

6.1.3 Impact of MT quality

Denkowski and Lavie (2012) make a clear distinction between analysing MT as a final product and MT fit for post-editing, saying that evaluation methods for the first may not necessarily be appropriate for the latter. It is the latter that the present section will be concerned with: how can we analyse MT quality in order to predict post-editing effort? More specifically: Which kinds of MT errors have the highest impact on PE effort?

The problem has been approached in different ways, using error typologies (Koponen et al., 2012; Stymne et al., 2012) and human ratings ranging from 'good' to 'bad' (Doherty & O'Brien, 2009; Koponen, 2012; Popovic et al., 2014). Koponen et al. (2012) used the classification proposed by Temnikova (2010), which contains various MT output errors ranked according to cognitive demand, and Stymne et al. (2012) used the more limited classification proposed by Vilar et al. (2006). This difference in classification makes it hard to compare both studies, although they both found word order errors and incorrect words to impact post-editing effort the most (Koponen et al. studied their relationship with post-editing time, and Stymne et al. studied their relationship with fixation duration). Popovic et al. (2014), Doherty and O'Brien (2009) and Koponen (2012) used human-assigned sentence ratings with four or five levels, with the highest score indicating that no post-editing was needed and the lowest score indicating that it would be better to translate from scratch. Whereas Popovic et al. (2014) used all levels in their analysis, Doherty and O'Brien (2009) and Koponen (2012) limited theirs to the MT segments with highest and lowest quality. It is therefore, again, hard to directly compare both studies. For the lower quality sentences, Popovic et al. (2014) and Koponen (2012) found an increase in the number of word order edits and Doherty and O'Brien (2009) found a higher number of fixations. For future work, researchers suggest using more fine-grained error typologies (Koponen et al., 2012; Stymne et al., 2012) and different languages (Koponen et al., 2012; Popovic et al., 2014; Stymne et al., 2012).

6.1.4 Impact of experience

Inexperienced translators have been shown to treat the translation task as a mainly lexical task, whereas professional translators pay more attention to higher-order concerns such as coherence and style (Séguinot, 1991; Tirkkonen-Condit, 1990). Students have also been shown to require more time (Tirkkonen-Condit, 1990), more

fixations and more pauses than professional translators while translating (Dragsted, 2010).

A comparable trend is found in revision research. Sommers (1980), for example, found that experts adopt a non-linear strategy, focusing more on meaning and composition, whereas student revisers work on a sentence level and rarely ever reorder or add information. Hayes et al. (1987) too reported that expert revisers first attend to the global structure of a text - the so-called higher-order concerns - whereas novice revisers attend to the surface level of a text - lower-order concerns. Revision is seen as a complex process that puts a great demand on working memory. Broekkamp and van den Bergh (1996) found that students were heavily influenced by textual cues during the revision process. For example, if a text contained many grammatical errors, the reviser's focus switched to solving grammatical issues, and higher-order concerns were ignored.

We expect to be able to extrapolate these findings to post-editing, although research linking experience to the post-editing process is hard to find. The study by de Almeida and O'Brien (2010) is, to the best of our knowledge, one of the only ones. They found that more experienced translators were faster post-editors and made more essential changes as well as preferential changes. They further suggest that differences in keyboard and mouse usage can be attributed to individual preferences rather than differences in experience.

6.2 MT error analysis

We used *Google Translate* to provide the Dutch machine translation for the texts (output obtained January 24, 2014, see Appendix 3). To be able to identify the relationship between specific machine translation problems and post-editing effort, the same two annotators from the pretests and main experiment annotated the MT output of all texts for quality on the basis of the translation quality assessment approach developed during the pretests and used in the main analysis. After the annotation process, the annotators discussed discrepancies in their annotations, and only the annotations that both annotators agreed on were kept for the final analysis. As with the pretests and final products discussed in the main analysis section, annotations were made with the brat rapid annotation tool (Stenetorp et al., 2012).

Out of 63 segments, 60 segments contained at least one error. There were more acceptability issues (201 instances) than adequacy issues (86 instances). The error categorisation described in section 3.2.1 contained 35 types of acceptability issues and 17 types of adequacy issues, but not all issues were found in the machine translation

output. To be able to perform statistical analyses, some of the error categories were grouped together, so that each error type occurred at least 10 times. The final classification can be seen in Figure 43. The category 'adequacy' contains all forms of adequacy issues. Other meaning shifts and word sense issues occurred frequently enough to be considered as separate categories, the other subcategories (additions, deletions, misplaced words, function words, part of speech, inconsistent terminology) were grouped together into 'adequacy other'. Within the grammar and syntax category (the most common error category for MT output), word order issues, structural issues and incorrect verb forms occurred more than ten times each. The different types of agreement issues (noun-adjective, article-noun, subject-verb, and reference) were grouped into a new 'agreement' category, and the other grammatical issues are contained in the 'grammar other' category (superfluous or missing elements). For coherence issues, the category 'logical problem' occurred more than ten times, but the other categories together (conjunction, missing info, paragraph, and inconsistency) did not occur more than ten times, so all coherence categories were grouped together. For lexicon, the subcategory 'wrong collocation' appeared often enough to stand alone, the other subcategories (wrong preposition, named entity, and word non-existent) have been grouped into 'lexicon other'. All subcategories for style and spelling have been grouped together into the main categories, since there were very few instances of these subcategories.

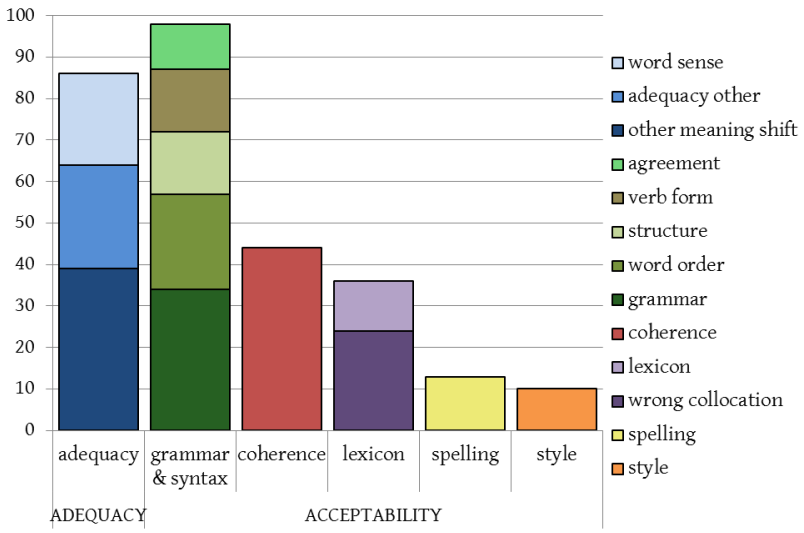


Figure 43 Overview of regrouping and number of occurrences of each error type in the MT output.

The most common errors are grammatical errors (grammar & syntax bar), followed closely by adequacy issues. Spelling and style issues barely occur.

6.3 Hypotheses

Building on the abovementioned research, we expect different types of post-editing effort indicators (e.g., fixation duration, number of fixations, (average) pause ratio, post-editing time) to be impacted by different types of machine translation errors, and we expect that the process post-editing effort indicators are impacted by different machine translation error types than the product post-editing effort indicator HTER. More specifically, we expect the process effort indicators to be influenced most by machine translation errors that require higher cognitive effort to correct, whereas we do not expect this to necessarily hold true for HTER as well. Post-editing time would, for example, be influenced most by mistranslations and word order issues (Koponen et al., 2012) and lexical or semantic issues (Popovic et al., 2014). Following Alves and Gonçalves (2013), we expect the number of production units to be influenced most by morpho-syntactic elements and word order. Regarding fixations, we expect to see more and longer fixations for word order errors and mistranslations (Stymne et al., 2012).

Building on the notion that inexperienced revisers focus more on grammatical errors when there is an abundance of grammatical errors (Broekkamp & van den Bergh, 1996), and the fact that student translators treat translation as a lexical task (Tirkkonen-Condit, 1990), we expect students to focus mostly on the grammatical and lexical issues, whereas professional translators are expected to pay more attention to coherence, meaning, structural, and style issues (Sommers, 1980).

6.4 Analysis

The final dataset comprised 721 post-edited segments, concatenated from all post-editing sessions from the main experiment and enriched with MT quality information, i.e., the average error weight per word in that segment for each error type. The statistical software package *R* (R Core Team, 2014) was used to analyse the data. As for our main analysis, we used the *lme4* package (Bates et al., 2014) and the *lmerTest* package (Kuznetsova et al., 2014) to perform linear mixed effects analyses. In this particular study, we used the various post-editing effort indicators as dependent variables, machine translation quality and experience as independent variables, and participants and sentence codes (a unique id for each source text sentence) as random effects, because we expected there to be individual differences across participants as well as sentence-inherent effects.

For each of the models discussed below, we start by building a null model with the post-editing effort indicator as dependent variable and participant and sentence code as random effects. This model is then tested against the model containing two predictors (the continuous variable machine translation quality and the categorical variable experience, which was either 'student' or 'professional') plus interaction effect. As a measure of model fit, we used Akaike's Information Criterion (AIC) values (Akaike, 1974). The model with the lower AIC value, has a better fit compared to the model with the higher AIC value, especially when the difference between both models is greater than 4 (Burnham & Anderson, 2004), but an AIC value in itself has no absolute meaning. We looked at the impact of machine translation quality on different post-editing effort indicators in three different analyses, increasing the granularity of the machine translation quality assessment with each analysis, as we expect a more fine-grained analysis to highlight the differences between the different post-editing effort indicators more clearly.

The first analysis was the most coarse-grained, in which we used the average total machine translation error weight and experience as predictor variables. For each of the post-editing effort indicators, we built a null model and a model with both predictor variables, which we then compared.

The second analysis was a little more fine-grained. Rather than using the total machine translation error weight, we looked at the total acceptability error weight and the total adequacy error weight as possible predictor variables. For each of the post-editing effort indicators, we built a null model and a model with acceptability and adequacy error weight as predictors. We tested whether acceptability or adequacy errors - or both - added anything to the model. On the basis of this analysis, we built the final models. For example, if only acceptability was found to be relevant, the final model contained acceptability and experience plus interaction effect. If both acceptability and adequacy were found to be relevant, we built two models: one with acceptability and experience plus interaction effect, one with adequacy and experience plus interaction effect.

For the third and final analysis, we went even more fine-grained, as suggested in previous research (Koponen et al., 2012; Stymne et al., 2012). We again built separate models for each of the post-editing effort indicators. This time, we added all the machine translation error subcategories as possible predictors. We then tested which of the subcategories were retained by the step-function from the *lmerTest* package. These subcategories were used as predictors in the final model, to which we again added experience and interaction effect.

6.5 Results

6.5.1 Analysis 1: coarse-grained

A summary of the models with average total machine translation error weight and experience plus interaction as possible predictor variables can be found in Table 23. The table contains the dependent variable, i.e., the post-editing effort indicator under scrutiny, the AIC values for the null model (i.e., model without predictor), as well as the model with predictor, an overview of the significant predictors (i.e., MT error weight and/or experience, with or without interaction effect), the effect size for each significant predictor, and the p-value to indicate significance. The effect column shows the actual effect an independent variable has on a dependent variable[2].

For most post-editing effort indicators, the machine translation quality has an impact, but experience does not, with the exception of both pause measures. The effect of machine translation quality on the effort indicators is in line with expectations, with a decrease in quality leading to an increase in time needed, an increase in the number of fixations and the number of production units, an increase in HTER score, and a decrease in average pause ratio. The only post-editing effort indicator the machine translation quality seems to have no statistically significant impact on, is pause ratio. This effort indicator is impacted by experience rather than machine translation quality, whereas the average pause ratio is impacted by experience as well as machine translation quality. Students have a significantly higher pause ratio than professionals, presumably requiring more time to think before changing anything, but also a higher average pause ratio, which is somewhat more surprising. It must be noted, however, that this latter effect is not quite significant (p=0.05) and so it must be interpreted with caution. We further found no significant effect of either machine translation error weight or experience on the average fixation duration.

[2] For example, if the MT error weight increases with a value of one, the average duration per word will increase with 3.5 seconds, with a standard error of 1.2 seconds, meaning that the expected increase ranges from 2.3 seconds to 4.7 seconds.

Table 23 Summary of mixed models with average total MT error weight and experience plus interaction effect as fixed effects.

Dependent variable	AIC without predictor	AIC with predictor	significant predictors	effect	p
Avg duration per word (in ms)	13729	13724	MT error weight	3526 (± 1224)	0.005
Avg fixation duration (in ms)	6946	6948	n/a	n/a	n/a
Avg number of fixations	5201	5194	MT error weight	10 (± 2.7)	< 0.001
Avg number of production units	-123	143	MT error weight	0.3 (± 0.06)	< 0.001
Pause Ratio	-373	-383	Experience	0.1 (± 0.03)	0.002
Avg Pause Ratio	3400	3383	MT error weight	-2.33 (± 0.75)	0.002
			Experience	1.06 (± 0.53)	0.05
HTER	-95	-120	MT error weight	0.44 (± 0.07)	< 0.001

6.5.2 Analysis 2: finer granularity

Table 24 contains a summary of the second level analysis, in which we first established whether acceptability, adequacy or both significantly influenced the various types of post-editing effort indicators. The first part of the analysis shows that acceptability error weight is a better predictor of post-editing effort than adequacy error weight, except for the average number of fixations, which is influenced by both (second column). In a second step, we made separate models for the significant MT error type and added 'experience' with interaction as possible predictor variables. The significant predictors are shown in the fifth column. In case the model with adequacy and acceptability did not outperform the null model (as indicated by a higher AIC value), we did not build the additional models and the tables contain 'n/a' (not applicable).

The effects are comparable to the effects found for the level 1 analysis, with all effort indicators - with the exception of average pause ratio - increasing with a decrease in MT quality, and average fixation duration not being predicted by either MT quality or experience. Again, experience only has an impact for (average) pause ratio. It is the only significant predictor for pause ratio, and one of the predictors for average pause ratio, with error weight plus interaction effect adding to the model in this case. In contrast with the findings from the level 1 analysis, the effect of experience as well as the interaction effect is statistically significant. Although students have a higher average

pause ratio than professional translators, the average pause ratio decreases more rapidly than it does for the professionals with a decrease in MT quality.

Table 24 Summary of mixed models with average total adequacy and acceptability error weight as potential fixed effects, and experience plus interaction effect as added fixed effect.

Dependent variable	ACC and /or AD?	AIC without predictor	AIC with predictor	significant predictors	effect	p
Avg duration per word (in ms)	ACC	13729	13726	MT error weight	3605 (±1394)	0.01
Avg fixation duration (in ms)	ACC	6947	6948	n/a	n/a	n/a
Avg # fixations	AD	5202	5197	MT error weight	13.4 (± 4.02)	0.001
	ACC	5202	5198	MT error weight	9.72 (± 3.15)	0.003
Avg # production units	ACC	-123	-138	MT error weight	0.32 (± 0.07)	<0.001
Pause Ratio	ACC	-374	-382	Experience	0.11 (± 0.03)	0.001
Average Pause Ratio	ACC	3401	3384	MT error weight	-2.43 (± 0.85)	0.005
				Experience	1.14 (± 0.52)	0.035
				MT EW:Exp	-1.77 (± 0.85)	0.039
HTER	ACC	-95	-111	MT error weight	0.42 (± 0.09)	<0.001

6.5.3 Analysis 3: finest granularity

We get a more nuanced picture when looking at the results for the most fine-grained level analysis in Table 25. Again, we first built a model including all possible machine translation error types as possible predictors for the different post-editing effort indicators. The significant error types are listed in column two, showing that different types of machine translation errors impact different post-editing effort indicators. The more technical effort indicators (number of production units, pause ratio, and average pause ratio, HTER) are mostly impacted by grammatical errors (grammar, structure, word order), whereas the more cognitive effort indicators (fixations and time) are influenced most by coherence and other meaning shifts. We then built a model for each of the significant MT error types separately and added experience plus interaction effect as possible predictors. The significant ($p<0.05$) and almost significant ($p<0.06$) effects for these models can be seen in column five, with the actual effect in column six. For non-significant effects, the table contains 'n/a'.

It is interesting to see how experience and/or experience with interaction effect now become relevant for the models with average duration, average fixation duration, and HTER as dependent variables. In the case of average duration, students seem to be impacted less by an increase in coherence issues than professional translators, although the significance levels are not convincing (the experience effect only just reaches significance, the interaction effect almost reaches significance). In the case of average fixation duration, only students seem to be impacted by an increase in other meaning shifts, whereas the average fixation duration for professional translators remains comparable. In the case of HTER, students seem to make fewer edits than professional translators with an increase in adequacy issues. These consisted mainly of deletions and part of speech errors. In the models for average number of production units and average pause ratio, a few different machine translation error types seem to influence the post-editing effort indicators, but once split into a separate model containing experience with interaction effect as additional predictors, the predictors are no longer significant. The model with word order as predictor variable of the average number of production units approaches significance, but that is the only one.

Table 25 Summary of mixed models with average MT error weight for the subcategories retained by step function as fixed effects, and experience plus interaction effect as potential additional fixed effects.

Dependent variable	predictor retained	AIC without predictor	AIC with predictor	significant predictors	effect	p
Average duration per word (in ms)	coherence	13729	13719			
				MT error weight	9866 (± 2642)	0.003
				Experience	1338 (± 622)	0.041
				MT EW:Exp	-3799 (± 1963)	0.054
Average fixation duration (in ms)	other meaning shift	6947	6938			
				MT EW:Exp	51.64 (± 16.63)	0.002
Average number of fixations	other meaning shift	5202	5197			
				MT error weight	17.85 (± 5.56)	0.002
	coherence	5202	5194			
				MT error weight	23 (± 6)	0.002
Average number of production units	other meaning shift	-123	-127			
				n/a	n/a	n/a
	grammar	-123	-122			
				n/a	n/a	n/a
	structure	-123	-121			
				n/a	n/a	n/a
	word order	-123	-122			
				MT error weight	0.43(± 0.22)	0.052
Pause Ratio	grammar	-374	-384			
				Experience	0.08	0.002
Average Pause Ratio	coherence	3401	3397			
				n/a	n/a	n/a
	structure	3401	3401			
				n/a	n/a	n/a
HTER	adequacy other	-95	-101			
				MT error weight	0.77 (± 0.26)	0.004
				MT EW:Exp	-0.47 (± 0.18)	0.01
	agreement	-95	-95			
				MT error weight	0.65 (± 0.27)	0.02

6.6 Discussion

In order to improve Translation Environment Tools, we need to find objective ways to assess post-editing effort before presenting machine translation output to the translator. We expected machine translation quality to be a possible objective predictor of post-editing effort, but post-editing effort can be measured in various ways. In order to determine which types of post-editing effort could be predicted by which types of

machine translation errors, we studied the impact of machine translation quality on seven types of post-editing effort indicators (six process-based effort indicators and one product-based effort indicator), in three stages of increasing MT quality assessment granularity. We added experience as a predictor because we expected professional translators and student translators to react differently to different types of machine translation errors, as students have been shown to focus on different elements during translation (Séguinot, 1991; Tirkkonen-Condit, 1990), students and professionals exhibit differences in translation styles (Dragsted, 2010), and students in particular focus more on grammatical issues when there is an abundance of grammatical issues (Broekkamp & van den Bergh, 1996) - which is the case in the present study.

From the two more coarse-grained analyses, we learned that post-editing effort indicators - including the expected deviant HTER - can indeed be predicted by machine translation quality, with the exception of average fixation duration and pause ratio. Previous studies have shown that average fixation duration does not differ significantly for good and bad MT quality (Doherty & O'Brien, 2009; Doherty et al., 2010), and so perhaps the average fixation duration is not a good measure of post-editing effort. For pause ratio, it seems that students require significantly more time in pauses than professional translators, and this effect outweighs the impact of machine translation quality. Perhaps students need more time to think about the correct course of action, whereas this process has been become more automatic for professionals. This confirms previous findings on students' pause behaviour (Dragsted, 2010). Our findings also offer some confirmation that average pause ratio indeed measures something else than pause ratio (Lacruz et al., 2012). Experience does seem to influence average pause ratio as well, especially for the more fine-grained analysis, but the p-values are rather high and so these results should be interpreted with caution. Interesting here as well is the fact that the average number of fixations is impacted by acceptability as well as adequacy issues. Adequacy issues are intuitively more cognitively demanding to solve, and the average number of fixations is one of the only effort indicators that is exclusively related to cognitive effort.

The most fine-grained analysis shows that there is indeed a difference between the different types of post-editing effort, as different types of machine translation errors predict different types of post-editing effort. Still, there are a few machine translation error types that occur with more than one post-editing effort indicator, such as coherence issues, other meaning shifts, grammar and structural issues. The only post-editing effort indicator that, in line with our expectations, has nothing in common with the other types of post-editing effort indicators, is HTER, which is influenced by agreement issues and certain types of adequacy issues. Upon closer inspection, we found these adequacy issues to consist mostly of deletions, which would of course automatically show up in a HTER score. What is striking in this analysis, is the lack of word order issues as an influential category. We only found word order issues to be

relevant for the average number of production units, but not for other post-editing effort indicators, even though we expected them to influence post-editing time (Koponen et al., 2012), the average fixation duration, and number of fixations (Stymne et al., 2012) as well. A possible explanation for these findings is the fact that our error classification includes error types such as coherence, which was not included in Temnikova's (2010) classification, and those error types outweigh the effect of previously used error types such as word order issues. Further research on a larger dataset and different languages is needed to support or refute these claims. Some of our findings are in line with expectations as well. For example, the number of production units was indeed influenced most by grammatical issues (Alves & Gonçalves, 2013), and fixations were influenced most by other meaning shifts (comparable to mistranslations), as suggested by Stymne et al. (2012). With respect to both pause measures, we again find support for the claim made by Lacruz et al. (2012) that average pause ratio is a better measure of cognitive effort than the pause ratio suggested by O'Brien (2006), who did not find conclusive evidence to support her claim. Pause ratio seems more technically motivated, with grammatical issues having the largest impact, but the experience effect still outweighs the MT quality effect. The average pause ratio, on the other hand, is influenced by structural issues - which are presumably technically as well as cognitively demanding to solve - in addition to coherence issues, which can be assumed to require more cognitive processing as well.

Regarding experience effects, we expected students to be more heavily influenced by lower-order concerns and professional translators to be more heavily influenced by higher-order concerns (Hayes et al., 1987; Séguinot, 1991; Sommers, 1980; Tirkkonen-Condit, 1990). This almost holds true for the average duration per word, where an increase in coherence issues leads to an increase in the time needed and this effect is stronger for professional translators than for students. It must be noted, however, that this interaction effect is not quite significant ($p=0.054$). For the average fixation duration, which is influenced most heavily by other meaning shifts, only the interaction effect between the increase in other meaning shifts and experience is significant. While the average fixation duration for professional translators remains more or less constant as the number of other meaning shifts increases, the average fixation duration for students goes up under the same circumstances. This can be an indication that other meaning shifts are cognitively more demanding for students, seeing how it is not an issue they usually focus on during translation, whereas spotting and solving these issues is probably more routinised for professional translators, causing their average fixation duration to remain constant (Séguinot, 1991; Tirkkonen-Condit, 1990). The opposite is true for the effect of adequacy issues and experience on HTER. Here, it is the professional translators who make more edits than students with an increase in adequacy issues. In order to determine what causes this discrepancy, we need to look

into the final translations. Perhaps students did not spot these errors or professional translators solved them in more creative ways.

Our analysis, and especially the most fine-grained level of analysis, provided evidence for our main hypotheses: that different types of post-editing effort indicators are influenced by different types of machine translation errors, and that HTER in itself is not a sufficient measure of post-editing effort, as it is influenced by a small subset of MT errors, a subset which, furthermore, does not seem to significantly impact any of the other post-editing effort indicators. This shows that the question of what influences post-editing effort depends on which type of effort is meant, with coherence issues, grammatical issues and other meaning shifts being good candidates for effort prediction on the basis of MT quality.

Although we included data from students as well as professional translators, the total dataset remains relatively small. While our most fine-grained analysis showed some promising results, it must be noted that with the more fine-grained analyses, fewer observations are taken into account, and so these results need to be interpreted with caution. Further experiments with the same categorisation on more data and different language pairs should be carried out in order to further develop the claims made in this chapter.

6.7 Conclusion

We confirmed that certain machine translation error types impact certain types of post-editing effort indicators, and that this effect is - in some cases - influenced by experience as well. We found that, while most machine translation error types occur with more than one post-editing effort indicator, HTER does not share machine translation error types with any of the other effort indicators, providing support for our hypothesis that product effort measures do not measure actual post-editing effort the way process effort measures do.

Once the most important error types and their impact have been determined, this information can - in future work - be used to improve translation tools, by only providing MT output to a translator when the effort to post-edit a sentence is expected to be lower than the effort to translate the sentence from scratch, and by taking into account that post-editor's level of experience. Additionally, translator training that incorporates post-editing can be adapted to make future translators more aware of effortful machine translation errors. By learning how to spot and solve these types of issues, the post-editing process can, in turn, become less strenuous as well.

Chapter 7 Impact of MT quality on students' processing of multi-word units[1]

"A multiword is a lexical unit formed by two or more words to yield a new concept, different from the composition of the meaning of its elements" (Portela, Mamede, & Baptista, 2011, p. 1). As such, they offer quite a challenge to machine translation systems, as literal translations often do not suffice (Wehrli & Nerima, 2013). This would not be so problematic if multi-word units were not as ubiquitous as they are (Mendoza Rivera, Mitkov, & Corpas Pastor, 2013). Even though phrase-based statistical machine translation (SMT) systems are generally better suited to translate multi-word units (MWUs) than rule-based machine translation (RBMT) systems (as SMT systems can extract multi-word units from corpora and therefore have a better coverage of multi-word units than RBMT systems), SMT systems often still lack the semantico-syntactic knowledge required to properly translate MWUs (Monti, Barreiro, Elia, Marano, & Napoli, 2011).

In order to improve machine translation systems' processing of multi-word units, most research has gone into identifying and classifying MWUs in machine translation output (Monti, Mitkov, Corpas Pastor, & Seretan, 2013). What most of these researchers have in common, is that they have looked at multi-word units solely from the 'how is it processed by the machine translation system' point of view. This information is then used to improve the processing of multi-word units by machine translation systems via linguistic pre-processing, POS-pattern definition, syntactic and semantic processing (Mendoza Rivera et al., 2013). Although improving the output of machine translation systems is definitely an important goal in itself, most machine translation output is subsequently post-edited in order to obtain high quality products. Still, research into the effects of MWUs on subsequent post-editing is limited. If post-editing is used in multi-word unit research, it is often as a measure of the machine translation quality, in

[1] Part of the work described here will be published as Daems, Vandepitte, Hartsuiker, and Macken (2016a).

which a higher edit rate is equal to lower machine translation quality (Barreiro, Monti, Orliac, & Batista, 2013; Seretan, 2015). In this case, the assumption is made that human post-editors detect and correct all machine translation errors, which is, unfortunately, not always true.

In a post-editing scenario, it makes sense to ask 'which types of multi-word units are problematic for a post-editor' in addition to 'which types of multi-word units are problematic for the machine translation system', and 'how can we improve the machine translation quality'. Seretan (2015) already took a step in this direction by looking at the effort involved in post-editing MWUs. She found that 63.2% of multi-word units in user-generated content were translated correctly by the MT system (correctly meaning that the post-editors either changed nothing, or the changes were 'minor', such as agreement or number changes), and that the effort involved in post-editing MWUs accounted for 20% of all post-editing effort (as measured by a variety of edit rates between MT output and the post-edited product). When it comes to quality evaluation, however, Seretan (2015) as well considers the post-edited product to be an indication of machine translation quality, not a translation product that needs to be evaluated in its own right.

When evaluating multi-word units in a post-editing context, however, we believe that it is important to understand how different types of multi-word units are processed by post-editors, and how this affects the final quality. As discussed in section 6.2, we used the same quality assessment approach for the raw MT output as for the post-edited texts in order to better compare both products. Our two-step translation quality assessment approach takes into account some of the main concerns Monti et al. (2011) had regarding quality evaluation by means of comparison with reference translations: "All these metrics only partially give reliable results concerning machine translation quality, since the judgement is based not on whether a machine translation system translates accurately the meaning and the message of an original text, but only how well it scores against references" (p. 17). The researchers continued by stating that the output should be evaluated from two perspectives: as a text derived from a source text, and as an autonomous text in the target language and culture. These are precisely the two aspects we take into account in our translation quality assessment approach. In addition, we suggest taking the external resources consulted during the post-editing process into account, as they can give a better idea of post-editors' problem-solving strategies (Göpferich, 2010) and are as such an indication of their uncertainty while post-editing.

In order to shift the focus to the impact of machine-translated multi-word units on subsequent post-editing, we tried to keep our multi-word unit classification itself as simple as possible, but we added the factor 'contrast with the target language' to the classification. This distinction is of course language-dependent, but we believe it to be a necessary addition for this type of analysis. If, for example, an idiom is not contained in

the corpus a statistical machine translation system was trained on, the system will resort to word-by-word translation. Depending on the target language, that idiom might actually still make sense. While the English idiom 'it's raining cats and dogs' cannot be translated into Dutch as *het regent katten en honden*, the English 'the apple does not fall far from the tree' can literally be translated into Dutch as *de appel valt niet ver van de boom* and retain its idiomatic meaning. The first would be highly confusing for the post-editor and would not make sense in Dutch at all, the second would not even require post-editing. Even though these examples belong to the same category (idioms), they have a different effect on subsequent post-editing because of their degree of contrast with the target language. We believe that the contrast with the target language and the subsequent post-editing process can provide interesting new insights in the processing of multi-word units in a machine translation scenario. These insights can, in turn, be used to improve the translation interface to better aid post-editors with their work by, for example, highlighting specific types of multi-word units that post-editors often find problematic.

In this chapter, we provide examples for our simplified categorisation, followed by a summary of the quality of machine translated multi-word units in our corpus, after which we analyse how these multi-word units are subsequently processed by student post-editors. We discuss the final quality and the usage of external resources.

7.1 Classification of multi-word units

On the basis of the abovementioned assumptions, we decided to classify multi-word units for this study in two ways: by category (compound, collocation, multi-word verb) and by contrastiveness: if a direct translation of the English multi-word unit would be correct in Dutch, the unit was classified as 'non-contrastive', whereas multi-word units that could not be translated literally into Dutch were classified as 'contrastive'. An explanation plus example of each type can be found below.

Compound: lexical units with more than one base functioning grammatically and semantically as a single unit

(44) a. high-rise
hoogstijger (=literal translation)
'hoogbouw' (=correct translation) = CONTRASTIVE
b. climate warming
klimaatopwarming (=literal translation)
'klimaatopwarming' (=correct translation) = NON-CONTRASTIVE

Collocation: a semantically autonomous base with an additional element which makes its own semantic contribution to the whole, but the selection of this element is dependent on the base

(45) a. for centuries
voor eeuwen (=literal translation)
'eeuwenlang' (=correct translation) = CONTRASTIVE
b. steep mountains
steile bergen (=literal translation)
'steile bergen' (=correct translation) = NON-CONTRASTIVE

Multi-word verb: a multi-word unit functioning like a single verb. This term encompasses phrasal verbs, prepositional verbs, phrasal-prepositional verbs and other multi-word verb constructions

(46) a. listen up
luister op (=literal translation)
'luister' (=correct translation) = CONTRASTIVE
b. count on
rekenen op (=literal translation)
'rekenen op' (=correct translation) = NON-CONTRASTIVE

In total, we found 99 multi-word units in the texts, 52 of which were collocations, 37 could be classified as compounds and 10 were multi-word verbs. Though we assumed it would be interesting to compare these three categories due to their varying degrees of semantic autonomy, we are mainly interested in the translation of multi-word units. Since translation is language-dependent, we classified our multi-word units further as contrastive or non-contrastive with Dutch (can the multi-word unit be translated literally or not?). The collocations are fairly evenly divided across categories, with 24 collocations being contrastive and 28 being non-contrastive. Of the compounds, 13 belong to the 'contrastive' category, the other 24 to the 'non-contrastive' category. Most of the multi-word verbs (7) are contrastive and only 3 of them are non-contrastive.

7.2 MT quality of multi-word units

In order to analyse the effect of machine translation quality on post-editing, we first need to establish how problematic each type of multi-word unit is for machine translation. Rather than comparing the MT output with a reference translation, we look at the MT output as a text in the target language (acceptability), and we compare the MT output to the source text, to assess its adequacy, as suggested by Monti et al. (2011).

As explained in section 6.2, we annotated the raw MT output for quality on the basis of our two-step translation quality assessment approach. As such, we get a detailed impression of the problems in the machine translated text, regarding both the number of problems and the type of problems. Figure 44 shows how common MT errors are for each type of multi-word unit. A value of zero means that there were no errors in the MT version of the multi-word unit, a value of one means that there was one error, and this up to three errors.

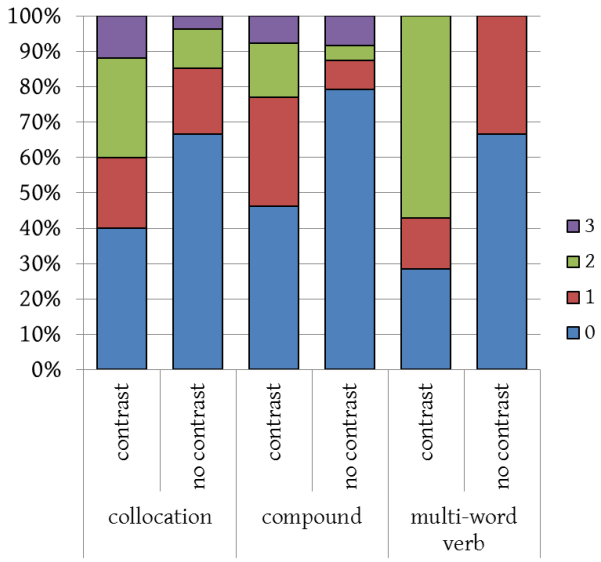


Figure 44 Frequency of zero, one, two or three errors occurring in multi-word units processed by MT, for each type of multi-word unit.

In line with our expectations, MT performs much better with non-contrastive MWUs than with contrastive MWUs, with 60 to 80% (depending on the MWU category) of non-contrastive MWUs being unproblematic for MT. In the contrastive conditions, only 30 to 45% of MWUs is processed correctly by the MT system. We also see a higher percentage of MWUs containing two or three errors after MT in the contrastive condition. Figure 45 gives an overview of the most common error types found in the MT output.

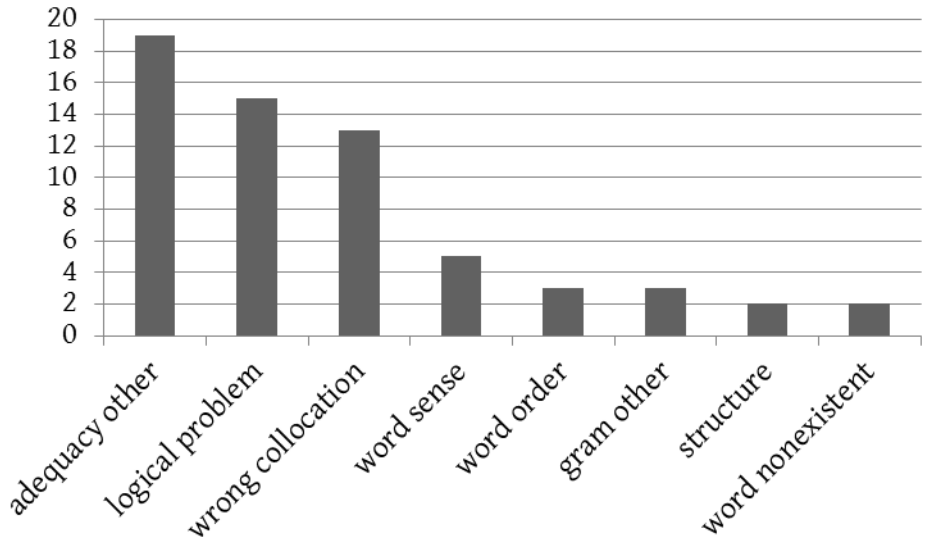


Figure 45 Error types occurring at least twice in machine translated multi-word units.

As can be seen in Figure 45, unclassified adequacy errors (adequacy other) are most common in the MT output, with logical problems and wrong collocations occurring frequently as well. The adequacy issues (word sense or others) often cause logical problems.

(47) ... the volunteers keep us company
... de vrijwilligers houden ons gezelschap (=correct translation)
... 'de vrijwilligers houden ons bedrijf' (MT, word sense error and logical problem: 'company' is translated in the sense of a business organisation)

In example (47), the word 'company' is mistranslated as *bedrijf* by the MT system. From an adequacy perspective, this is a word sense error ('company' can be translated as *bedrijf*, just not in this context), but it is also a logical problem from an acceptability perspective (*bedrijf* makes no sense in this sentence). As expected, grammatical errors also occur in the machine translation output (word order, structure, and other types of grammatical problems).

7.3 Post-edited multi-word units

After establishing how problematic the various multi-word units are for a statistical machine translation system, we take a closer look at the effectiveness of post-editing (PE). We anticipated four scenarios, which we briefly discuss with examples below.

No problem: the MWU was not problematic after MT, and none of the student post-editors introduced errors during the post-editing process.

(48) In exchange for...
'In ruil voor' ... (MT, correct translation)
'In ruil voor' ... (PE, correct translation)

Not solved: the MWU was problematic in MT, and is still problematic after post-editing for at least one student post-editor.

(49) Families are holding tight to their cash...
Gezinnen zijn strak vast te houden aan hun geld... (MT, incorrect too literal translation)
Gezinnen houden zich vast aan hun geld... (PE, incorrect too literal translation)

Solved: the MWU was problematic in MT, and is no longer problematic in any of the post-edited versions.

(50) self-described
'zelf-beschreven' (MT, incorrect too literal translation)
'zoals ze zichzelf noemen' (PE, correct translation)

Problem introduced: the MWU was not problematic in MT, but errors were introduced by at least one student post-editor during the post-editing process.

(51) finger-painting
'vingerverven' (MT, correct translation)
'vingerverfschilderijen' (PE, incorrect translation, verb as noun)

Figure 46 gives an overview of the outcome after post-editing for multi-word units in each category. Quite a few errors found in the machine translated MWUs are not solved after post-editing, especially in the contrastive condition. Example (52) shows a contrastive collocation that was mistranslated by MT and was still problematic for most post-editors. Post-editors three (P3) and ten (P10) provide correct translations, whereas post-editor one (P1) does correct the mistranslation of 'pulled', but not the mistranslation of 'together'. Post-editors five (P5) and eight (P8) correctly interpret the 'together' as belonging with 'pulled', but failed to choose a Dutch verb that completely covers the meaning of 'pull together'.

(52) Researchers have pulled together data...
'Onderzoekers hebben samen gegevens getrokken'... (MT)
'Onderzoekers hebben samen gegevens geanalyseerd' (P1)
'Onderzoekers hebben gegevens verzameld' (P5, P8)
'Onderzoekers hebben gegevens ... samengebracht' (P3)
'Onderzoekers hebben gegevens...samengebracht en onderzocht' (P10)

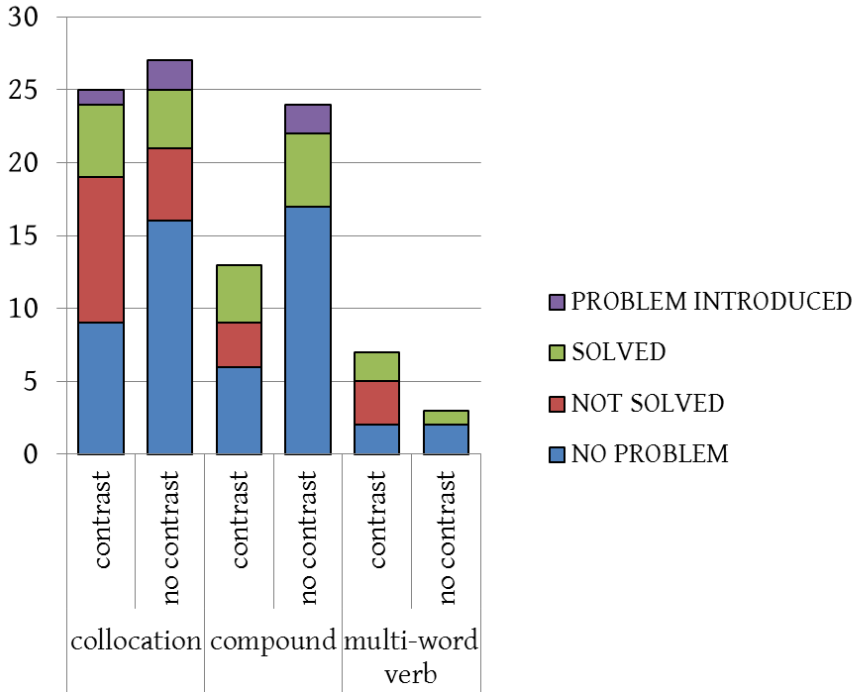


Figure 46 Occurrence of the four possible scenarios after post-editing (problem introduced, solved, not solved, no problem) for each type of multi-word unit.

In addition, collocations seem to be more problematic than compounds, which might be explained by the fact that in collocations, each part is semantically more autonomous than in a compound, so it might be harder for post-editors to spot errors there. Example (53) shows a case where a collocation is translated literally by the MT system. Yet, in context, the 'trip' is an indication of distance and not an actual 'trip', so the literal translation does not work in Dutch. This error goes unnoticed by one post-editor. An error in a compound noun, as shown in example (54), is much more obvious and is solved by all post-editors.

(53) ...the museum is just a short trip for
...'het museum is slechts een korte reis voor' (MT)
... 'het museum is slechts een korte reis voor' (P9)

(54) museum exhibit
'museumstuk' (MT)
'tentoonstelling' (PE)

The post-editors only introduced errors that were not there in the MT output in a few cases, most notably in the non-contrastive condition.

(55) urban lifestyles
'stedelijke levensstijl' (MT)
'straatcultuur' (PE)

An explanation would be that student post-editors do not trust the MT output and feel like they should change something, even when it is correct. Another explanation might

be that student post-editors correct certain MT errors, but introduce errors of their own. To verify this assumption, we take a closer look at the error types found in the 'not solved' condition, so those multi-word units that were problematic in machine translation and still problematic for at least one student post-editor (Figure 47).

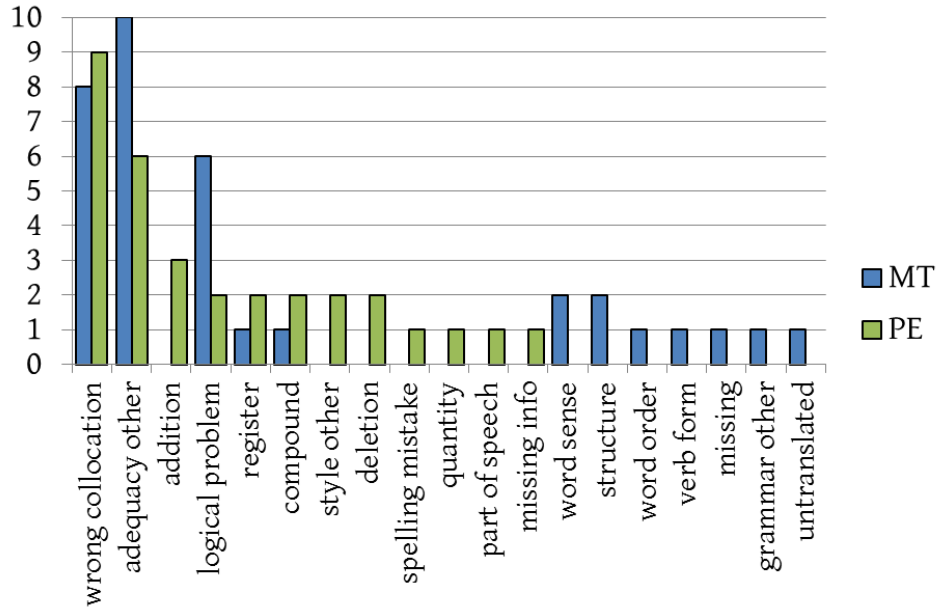


Figure 47 Error types in the 'not solved' condition for MT and PE.

From Figure 47, we can derive that 'wrong collocation' errors are abundant, both in the machine translation output and after post-editing. Unclassified adequacy errors (*adequacy other*) are more common in the MT output than in the final post-edited version, as are logical problems. Some (mostly grammatical) error types can only be found in the MT output, whereas adequacy and style issues only appear after post-editing. It can be assumed that grammatical errors are easily spotted and solved by a student post-editor, whereas some adequacy errors (i.e., contrasts between source and target text) are easily overlooked, perhaps because the text itself is fluent. In line with expectations, student post-editors treat the text more freely than the MT system, as evidenced by the number of additions, deletions and stylistic problems.

7.4 Usage of external resources

Though looking at quality tells us something about how difficult the processing of MWUs is for MT systems and subsequent post-editing, it does not tell the whole story.

We therefore take a closer look at the external resources consulted during the post-editing process.

The first question is 'for which types of MWUs do student post-editors consult external resources?' Figure 48 gives an overview of the multi-word units and whether or not resources were consulted when post-editing that particular type of multi-word unit. 'Yes' means that at least one post-editor consulted external resources during the translation of a multi-word unit, 'no' means that no post-editors looked up external resources for a multi-word unit.

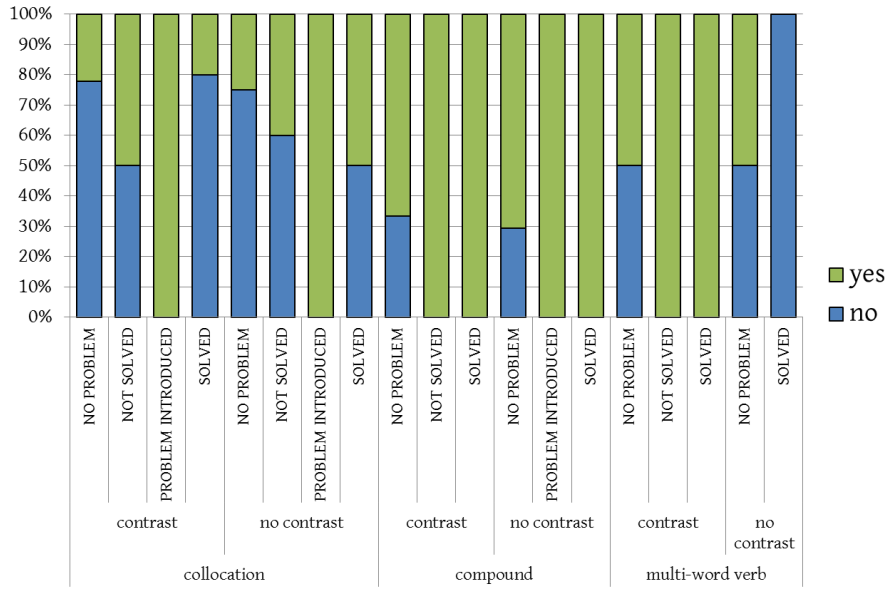


Figure 48 Proportion of multi-word units per category for which resources have been consulted by at least one post-editor.

As can be seen in Figure 48, student post-editors consult external resources for each type of multi-word unit. Remarkably, it is more common for post-editors to consult external resources when translating compounds than when translating collocations, even if the collocations are incorrect in the MT output and have been solved by post-editing. This might, in part, be due to the type of errors. As shown in Figure 48, many MT errors consist of grammatical errors, which presumably can be solved without consulting external resources. We further expect grammatical errors to occur more frequently within collocations than within compounds. Then again, there is still an abundance of collocations that were not solved by all post-editors, and it seems that post-editors consulted external resources for less than half of those cases. A comparison of contrastive and non-contrastive multi-word units shows comparable results for compounds, but when looking at collocations it seems that, overall, resources are consulted more frequently in the non-contrastive condition, which again is a little odd. To better try and understand these findings, we take a closer look at the time spent in the various types of external resources, and a few examples of search strategies.

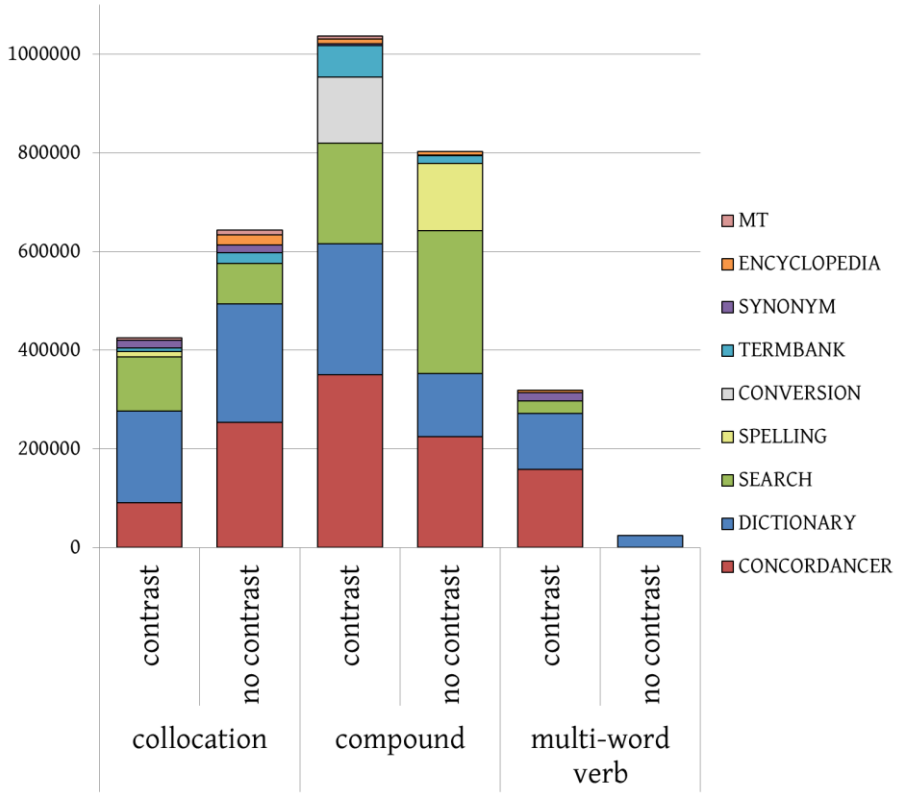


Figure 49 Total time spent in external resources for each type of MWU.

Figure 49 gives a general overview of the total time spent in each type of external resource for each type of multi-word unit. Even though there are more collocations than compounds in the data, a lot more time is spent looking up external resources when post-editing compounds than when post-editing collocations. It is also remarkable that, for collocations, more time is spent in external resources for non-contrastive collocations than for contrastive collocations. Perhaps the latter can be considered 'false friends', where the student post-editor believes that nothing should be changed, when actually the MT output is incorrect. This might also explain the abundance of 'not solved' MT problems in the contrastive collocations condition. Perhaps the frequency of multi-word units also plays a part. If a multi-word unit is not frequently used in the target language, the post-editor might want to verify that it is a correct translation, even for the non-contrastive multi-word units. We did not look at frequency in the present study, as it is rather hard to get accurate frequency information for collocations, definitely when they are split up in the sentence.

Overall, we see that most time is spent in dictionaries, concordancers and search engines, the latter being used more for compounds than for collocations. Perhaps student post-editors use a search query to verify that a particular compound exists, and if the search engine returns enough results, it is no longer necessary to consult other resources. The common choice of dictionaries for collocations is counterintuitive, as the words in collocations are more independent than compounds, and so we would not expect those to appear in dictionaries. A possible explanation is that student post-

editors do not realise that a word is part of a collocation and they try to solve the problems by looking up parts of the collocation rather than the whole. Closer inspection of the categories 'conversion' and 'spelling' reveal that both were used for only one multi-word unit each ('conversion' for '183-square-foot' and 'spelling' for '21st-century'), so these should not be considered as typical.

Table 26 Search strategy for multi-word unit 'low interest payments'.

Source descriptor	Time	Dur	Keystrokes	Type
Van Dale	758460	42500	interst[..]est m[.]pauy[..]yment	dictionary
Nieuw tabblad - Google Chrome	815038	4235	linguee	navigation
Linguee | Nederlands-Engels woordenboek	819273	5749	interest payments	concordancer
interest payments - Nederlandse vertaling – Linguee woordenboek	825022	22485		concordancer
Nieuw tabblad - Google Chrome	847507	4672	iate.europa.eu	navigation
IATE - De veeltalige databank van de EU	852179	18703	"interest paymne[..]ents"	termbank
IATE - Zoekresultaat	870882	30422	"interes[.]t payment"	termbank

An example of a search strategy can be seen in Table 26. During the twelfth minute of post-editing, the post-editor navigates towards the Dutch dictionary *Van Dale* and types in the words 'interest payment'. The post-editor remains on this page for forty-two seconds before opening a new tab and looking up 'interest payments' on *Linguee* (a concordancer website). This page remains in focus for twenty-two seconds, after which the post-editor navigates to the European multilingual term base *IATE* to look up the plural 'interest payments' as well as the singular 'interest payment'.

An additional question here is: how effective is the time spent in external resources? Does spending a lot of time in external resources equal better quality, and how much time is spent on passages that were not problematic to begin with? Figure 50 shows the average time spent in external resources for MWUs that were incorrectly translated by the MT system and were correctly translated by all student post-editors. Calculations were made by dividing the total time spent in each external resource by the number of participants that consulted that type of external resource and the number of multi-word units in a particular category.

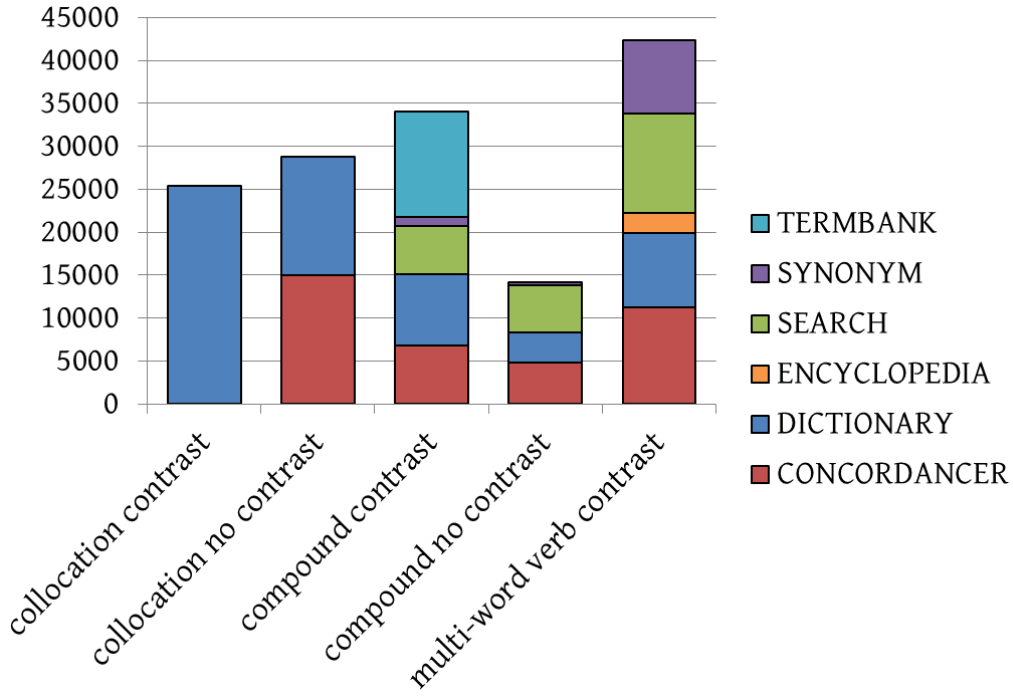


Figure 50 Average time (in ms) spent in external resources per multi-word unit incorrectly translated by the MT system and correctly translated by all student post-editors, for each category.

It seems that, on average, between twenty-five and forty seconds are needed to look up enough information to correctly post-edit multi-word units that were incorrectly translated by the machine translation system. In the contrastive condition, much more time on average is spent in external resources when post-editing compounds or multi-word verbs than when post-editing collocations. In addition, there is less variety in the types of resources consulted when solving collocations than when solving compounds and multi-word units. When solving contrastive compounds, a lot of time seems to be spent in term banks as well, though closer inspections reveals this search to be conducted by one post-editor only for one specific multi-word unit (interest payments), the one shown in Table 26.

Figure 51 gives an overview of the sources consulted while post-editing those multi-word units that were incorrectly translated by the machine translation system, and incorrectly post-edited by at least one post-editor. We compare the sources used by the post-editors that did not correctly post-edit the multi-word unit ('not solved' in the graph) with the sources used by post-editors that did manage to correctly post-edit the multi-word unit ('solved' in the graph). Calculations were made by dividing the total time in each external resource within each category of MWU by the number of MWUs in that category and the number of participants that consulted external resources for that category.

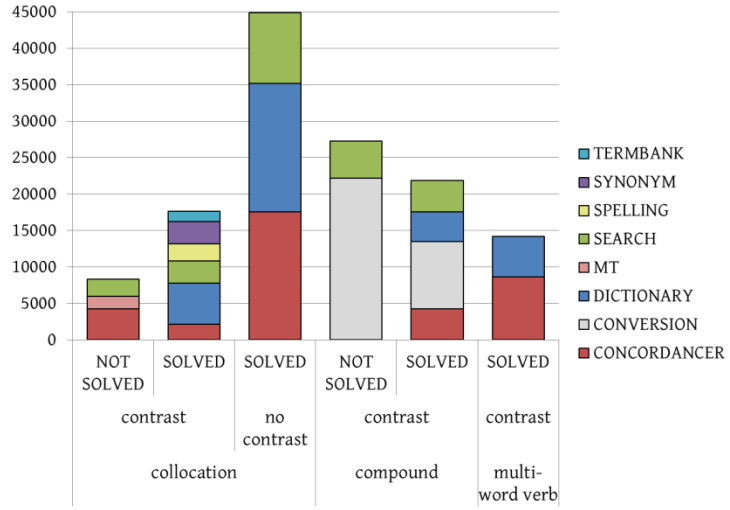


Figure 51 average time (in ms) spent in external resources for MWUs that were incorrectly translated by MT and not corrected by at least one student post-editor.

It can be derived from this graph that less time is spent in external resources by the post-editors that failed to solve the problems compared to those post-editors who corrected the machine translation errors. For compounds, this is not so obvious from the graph, but, as mentioned before, conversion is not a typical category, and in this particular case, there was only one participant who consulted resources to convert '183-square-foot' to square meters, leading to the high bar for contrastive compounds that were not solved. Taking this exception into account, Figure 51 seems to indicate that consulting external resources can help student post-editors solve problems related to multi-word units. There were only four cases where post-editors who consulted external resources when trying to correct an incorrectly translated multi-word unit failed to provide a good translation. Two of these were by the same post-editor, and in one of the cases, the error that remained after post-editing was a spelling error rather than the original logical problem found in the MT output. In contrast with the multi-word units that were correctly translated by all post-editors, we see a larger variety of sources consulted and more total time needed when translating collocations than when translating multi-word verbs, with the translation of contrastive compounds demanding the most time in external resources. There is also less variety in the types of resources consulted when translating multi-word verbs. In addition, term banks are only used when translating contrastive collocations, and not contrastive compounds, as was the case for multi-word units that were correctly translated by all post-editors. Table 27 shows the search strategy for the contrastive collocation 'fail their polygraph tests'. The collocation was translated correctly by all post-editors with the exception of post-editor nine (P9).

Table 27 Search strategies of five different student post-editors for multi-word unit 'fail their polygraph tests'.

Participant	Source descriptor	Time	Dur	Keystrokes	Type
P2	Van Dale	368148	9484	polydrap[....]graph	dictionary
P4	Van Dale	395901	6859	polygraaf	dictionary
P4	Van Dale	430381	3563		dictionary
P4	Nieuw tabblad - Google Chrome	505189	4531	polygraaf	navigation
P4	polygraaftest - Google zoeken	510236	4391	[....]	search
P6	- Google zoeken	574400	8672	polygraaftest	search
P6	polygraaftest - Google zoeken	583072	1281		search
P6	polygraaftest - Google zoeken	591737	4422		search
P6	polygraaftest - Google zoeken	919262	2235		search
P6	Nieuw tabblad - Google Chrome	921497	3203	]groene boekje	navigation
P6	groene boekje - Google zoeken	924700	1719		search
P6	Woordenlijst Nederlandse Taal - Officiële Spelling	926419	5765	test	spelling
P6	Nieuw tabblad - Google Chrome	932184	6360	testente[..] tests	navigation
P6	tests testen taaladvies - Google zoeken	938544	3046		search
P6	Testen / tests	941590	5688		spelling
P7	Van Dale - Google Chrome	320232	7938	polygraph	dictionary
P9	polygraph tests - Nederlandse vertaling – Linguee woordenboek	300124	9453		concordancer
P9	Nieuw tabblad - Google Chrome	309577	2828	polygrra[..]aaf	navigation
P9	Google - Google Chrome	312405	1360		navigation
P9	polygraaf - Google zoeken	313765	5406		search

Participants number two and seven simply look up the word 'polygraph' in a dictionary (*Van Dale*). Participant number four has a slightly more elaborate search strategy, looking up the Dutch word 'polygraaf' in the same dictionary and then navigating to Google to search for the word 'polygraaftest'. Participant number six has the most elaborate search strategy: he looks up 'polygraaftest' in *Google Search*, and switches back to consult the results of the search query throughout his translation process. The first time is around nine minutes and a half, with a few checks quickly following the first, then there's another check at around fifteen minutes. The post-editor then switches to *Groene Boekje*, which is the official word list of the Dutch language, to look up the correct spelling. The same query is given to the site *Taaladvies*, which is another website for checking Dutch spelling. Judging by the keystrokes, the post-editor wanted to know whether the Dutch plural of *test* is *tests* or *testen*. The last post-editor, also the only post-editor that made a mistake in the final translation of this multi-word unit, is the only person to use a concordancer (*Linguee*) to look up 'polygraph tests', after which she also consults *Google Search* to look up *polygraaf*. What's remarkable in this example is that the main issue in the machine translation output was the translation of 'fail' rather than the translation of 'polygraph test', yet all post-editors focus on 'polygraph test' in their searches. Though most post-editors correctly translate 'fail' as well, post-editor nine does not. It might be possible that in cases like this, where a compound (polygraph test) is part of a collocation (fail a test), post-editors focus on the compound rather than on the collocation as a whole.

The above findings indicate that looking up external resources can help student post-editors correct errors made by the MT system. Nevertheless, the success of looking up external resources is also determined by knowing when to look things up. A key post-editing skill is knowing when the machine translation is correct, and when it is not. We therefore look at all external resources consulted by participants for multi-word units that were correctly translated by the MT system and were either correctly translated by all post-editors (no problem), or where at least one post-editor introduced an error of their own (problem introduced). To study the usage of external resources in more detail, we looked at participants within the category 'problem introduced' and grouped them together according to their personal end result: 'no problem', or, if they had indeed introduced an error of their own, 'problem introduced'. Averages were once again obtained by dividing the total time spent in each external resource type by the number of MWUs in a particular category and the number of participants that consulted that particular type of external resource within that category.

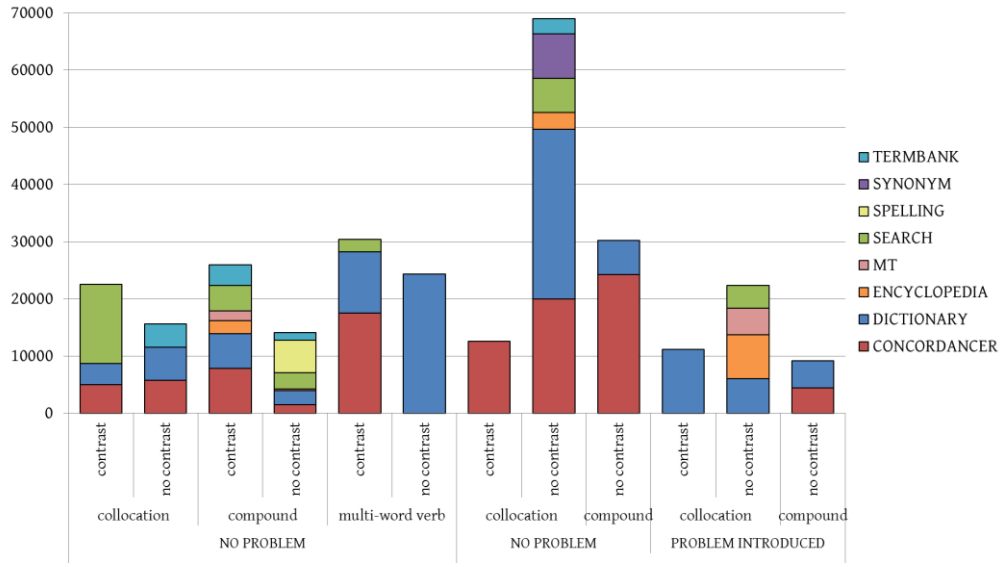


Figure 52 Average time (in ms) spent in external resources for MWUs that were correctly translated by MT.

From Figure 52, we can derive that students spend a lot of time looking up external resources when post-editing multi-word units, even when the multi-word units have been translated correctly by the machine translation system. The time spent on non-contrastive collocations in particular is striking, as is the wide variety of resources used to look up compounds. This finding is a little counterintuitive, as we would expect post-editors to not think twice about correctly translated compounds, and we expect contrastive collocations to require more time in external resources than non-contrastive collocations. Collocations in general are expected to require additional searches for verification due to their freer compositional nature than compounds. This could perhaps be a sign that post-editors do not consider contrastive collocations to be a whole, whereas they are more accustomed to compounds. Or, as mentioned before, this might be due to the low frequency of the compounds.

Table 28 Search strategy for multi-word unit 'high-rise'.

Source descriptor	Time	Dur	Keystrokes	Type
- Nederlandse vertaling - bab.la Engels-Nederlands woordenboek	1029750	6469	hih[.]gh-rise	dictionary
high rise - Nederlandse vertaling - bab.la Engels-Nederlands woordenboek	1036219	30687		dictionary
dagelijkse levensstijl - Google zoeken	1066906	3281	high-rise	search
high-rise - Google zoeken	1070187	4063		search
main document	1074250	2687		
high-rise - Google zoeken	1076937	3688	hoogbouw	search
hoogbouw - Google zoeken	1080625	6969	een	search
een hoogbouw - Google zoeken	1087594	3172		search
main document	1090766	48526		
...	...	...	...	...
een hoogbouw - Google zoeken	1330352	1297		search

In Table 28, we see an example of a student post-editor looking up 'high-rise', a contrastive compound that was correctly translated by the machine translation system as 'hoogbouw'. The post-editor first looks up 'high-rise' in the English-Dutch dictionary *bab.la* and via *Google Search*. He then returns to the main document for two seconds, and continues to use *Google Search*, this time to look up the Dutch word 'hoogbouw'. He adds the article 'een' to the search query and returns to the main document for almost an entire minute. The post-editor proceeds with the rest of the text for a while (omitted from example) and checks the search results again four minutes later. In total, the post-editor spent almost two minutes verifying a correct translation.

7.5 Discussion

In this chapter, we discussed the machine translation quality of various types of multi-word units, the subsequent post-editing process by students and the final quality of the product after post-editing. We suggest adding 'contrast with the target language' as a new factor in the evaluation and analysis of multi-word units in machine translation. Contrastive multi-word units were found to be more difficult than non-contrastive multi-word units for *Google Translate* to process as well as for the post-editors to correct. We further found collocations to be harder to post-edit than compounds. Fine-grained error analysis shows that grammatical errors and logical problems are usually corrected by the post-editors, whereas wrong collocation errors and adequacy issues remain after post-editing.

A closer look at the resources consulted during the post-editing of multi-word units showed that students consult resources more frequently and spend a lot more time looking up external resources when post-editing compounds than when post-editing collocations, which might indicate they need to be made more aware of collocations occurring in the text. We found that, if sources are consulted, the machine translation errors are usually corrected by the post-editor. More time is used to successfully process the contrastive multi-word units than the non-contrastive multi-word units, with the exception of collocations. The limited time spent in external resources when post-editing contrastive collocations might be the reason that a fair number of contrastive collocations remain problematic after post-editing. In addition, post-editors spend quite some time looking up multi-word units that were correctly translated by the machine translation system.

7.6 Conclusion

We can conclude that the difference between contrastive and non-contrastive multi-word units is a useful new way of classifying multi-word units regarding machine translation and subsequent post-editing. While post-editors' search strategies seem to be successful, they need to be made aware of contrastive collocations, and they could further benefit from some sort of MT quality estimation to prevent them from spending a lot of time looking up resources for correctly translated multi-word units.

Conclusion

This dissertation set out to gain a better understanding of the differences between human translation and post-editing from English into Dutch for general texts, and the impact of translators' experience on these differences.

As translators can no longer keep up with the increased need for translation, post-editing machine translation has become a new key skill for modern translators. While the benefits of post-editing for technical texts with customised systems have been established, especially regarding post-editing speed, findings for general text types are more mixed. As with any form of technology, the automated process (in this case, post-editing) needs to be compared with the manual process (human translation) in order to better understand how both processes work, what makes them work, and under which circumstances they work. In addition, translators' attitude and experience need to be taken into account, as the uptake of new technology depends on its perceived usefulness, and less experienced translators have been shown to translate and post-edit differently. We studied the different aspects of both methods of translation (human translation and post-editing) for two levels of experience (students and professionals) in order to answer the following main research questions:

1) What are the differences in process between human translation and post-editing (and is there a difference between students and professional translators)?

2) What are the differences in product between human translation and post-editing (and is there a difference between students and professional translators)?

3) What is the impact of machine translation quality on post-editing (and is there a difference between students and professional translators)?

4) What are the differences in attitude towards human translation and post-editing (and is there a difference between students and professional translators)?

In order to gain detailed and holistic information, we used a combination of keystroke logging tools, an eye tracker, and surveys before and after the experiment. We

developed a translation quality assessment approach suitable for a fine-grained and comparative analysis of human translations, post-edited texts, and machine-translated texts. We controlled for as many variables as possible in our main experiment through rigorous text selection and by using a balanced design, suitable for advanced statistical analysis.

In the following sections, we first discuss the empirical findings of our work and their theoretical implications, followed by more practical implications. We then highlight some important methodological choices that can be relevant to other researchers. The final sections address the limitations of the studies presented in this dissertation and contain suggestions for future research.

Empirical findings and theoretical implications

Process

What are the process differences between human translation and post-editing?

1) Is post-editing faster than human translation?

Our findings from the pretests as well as the main experiment indicate that post-editing is significantly faster than human translation, with human translators requiring approximately a second (main experiment) or two seconds (pretests) more to translate a word than post-editors. This supports industry findings (Groves & Schmidtke, 2009; Plitt & Masselot, 2010; Zhechev, 2014), and strengthens the (although statistically non-significant) findings from other studies with more general text types (Carl, Dragsted, Elming, et al., 2011; Garcia, 2011). The main implication is that post-editing can be faster than human translation from English into Dutch, even for general text types when using a statistical machine translation system that has not been customised.

In contrast with expectations based on de Almeida and O'Brien (2010), we did not find significant differences in time between students and professional translators. Kiraly (1995) and Jääskeläinen (1996) have suggested translator confidence as a more important factor than experience, and our findings support the idea that experience itself is not a sufficient predictor of speed.

2) Is post-editing cognitively more demanding than human translation?

Average fixation duration was significantly shorter during post-editing compared to human translation, indicating that post-editing is cognitively less demanding than

human translation. This corresponds to findings by O'Brien (2007), and suggests that the presence of machine translation output facilitates cognitive processing during translation, despite the fact that machine translation output is an additional resource that could take up working memory space (Krings, 2001). We tentatively attribute this to the presence of lexical information and the fact that, in the case of multiple translation options, the translator can simply take the option offered in the MT output without having to decide on an option themselves.

In contrast with our expectations, however, this presence of lexical information did not have a greater impact on students' cognitive processing than on that of professional translators (Sweller, 1988): we did not find a significant effect of experience on average fixation duration. Perhaps cognitive effort can be attributed more to individual differences than to differences in experience, but further research is needed to verify this suggestion.

3) Is the fixation behaviour different for post-editing and human translation?

The general trend for both methods of translation is longer average fixation duration on the target text compared to the source text, in line with Carl, Dragsted, Elming, et al. (2011) and Nitzke and Oster (2016).

When considering fixations on the source text, there were fewer fixations when post-editing compared to when translating from scratch. In addition, the average fixation duration on the source text was higher for students when translating from scratch, but not for professionals. The picture is somewhat different when considering fixations on the target text. Here, the number of fixations was higher when post-editing, but only for the students, and the average fixation duration was lower when post-editing.

This confirms our expectation that translators rely less on the source text when post-editing (Carl, Dragsted, Elming, et al., 2011; Carl et al., 2015). While professional translators' fixation behaviour remained constant when considering the average fixation duration on the source text or the number of fixations on the target text, students exhibited more diverse behaviour for both translation methods. Their heavier reliance on the source text during translation could be explained by a higher level of insecurity (Laukkanen, 1993). Their higher number of fixations on the target text during post-editing, however, is somewhat harder to interpret. Jakobsen and Jensen (2008) established that an increase in the number of fixation corresponded to an increase in cognitive processing, although this interpretation would be in contrast with our finding that the average fixation duration on the target text is shorter (and its processing thus cognitively less demanding) during post-editing than during human translation. Another possible explanation is that the number of fixations does not necessarily correlate with cognitive effort, but more with the way a text is processed, such as linear versus non-linear, although more in-depth analysis is needed to verify these assumptions.

4) Are more (or other) external resources consulted in human translation compared to post-editing?

From the pretest, we learned that there was a significant difference in the number of bilingual dictionaries consulted between human translation and post-editing, and almost a significant difference in the total number of external resources consulted. However, as the usage of external resources during the pretest was self-reported, these results need to be interpreted with caution.

Turning to the main experiment, we could not find a significant difference between human translation and post-editing when comparing overall time spent in external resources. Confirming the trend spotted in the pretest, there was a significant difference in the total number of external resources consulted, with fewer resources being consulted when post-editing. This both contradicts and confirms our expectations that the MT output already provides lexical information, reducing the need to consult additional external resources: while the number of resources consulted is indeed lower when post-editing, the time spent in external resources is comparable.

When looking at the types of resources in more detail, we found that only the time spent in dictionaries differed significantly between both participant groups, with students spending more time in dictionaries than professionals (A. Jensen, 1999; Prassl, 2010). On the basis of our data, we cannot establish whether professional translators indeed preferred monolingual sources (Jääskeläinen, 1990) or not (Kiraly, 1995), as *Van Dale* was the most frequently used dictionary, and the information registered by *Inputlog* for this website was not sufficiently detailed to discriminate between monolingual and bilingual consultations.

Product

What are the product differences between human translation and post-editing?

1) Is there a difference in overall quality between the product of human translation and the product of post-edited machine translation output?

As we had anticipated, we did not find statistically significant differences in overall quality between human translation and post-editing, or between students and professionals (Carl, Dragsted, Elming, et al., 2011; Kiraly, 1995). This shows that post-editing can lead to products of comparable quality to human translation for general text types, while being faster than human translation. We did not, however, find evidence for the trend that post-editing is evaluated as being better than human translation (Carl, Dragsted, Elming, et al., 2011; Garcia, 2011). As the method of evaluation in each study is different, it is hard to directly compare these findings.

The lack of impact of experience on overall quality confirms that more experienced translators are not necessarily the better translators (Jääskeläinen, 1996). We did find a significant correlation between the number of errors and the level of specialisation (proportion of translation work spent on general text types) for professional translators. Perhaps an increased level of specialisation leads to greater confidence and thus quality (House, 2000; Kiraly, 1995), or level of specialisation is yet another factor that could be taken into account for future research.

2) Is there a difference in the most common error types in human translations and post-edited texts?

On a high level, we can distinguish between acceptability and adequacy as error types. From the preliminary exploratory tests on the pretest data, we derived that human translation outperformed post-editing for acceptability (Guerberof, 2009), but post-editing outperformed human translation for adequacy (Lee & Liao, 2011). In our main experiment, however, we could not find statistically significant differences between human translation and post-editing for either category. We did find students overall to make more adequacy errors than professional translators.

On a more fine-grained level, we found that the most frequent error types are comparable across post-editing and human translation, with meaning shifts, logical problems, wrong collocations, word sense issues and deletions making up at least 5% of all errors made for at least one participant group. Meaning shifts make up a larger portion of human translation errors than of post-editing errors, and misplaced words are a more serious issue for post-editing than for human translation. For students, the proportion of logical problems and the proportion of word sense issues are greater when post-editing than when translating from scratch, and the proportion of deletions is lower when post-editing than when translating from scratch. Our category 'other meaning shifts' corresponds roughly to 'mistranslations' as used in Guerberof (2009) and confirms her finding that human translation scores worse than post-editing for mistranslations. The greater presence of logical problems and word sense issues in students' post-edited text could offer some support to the fact that students treat translation as a linguistic task (Séguinot, 1991; Tirkkonen-Condit, 1990), causing them to overlook other errors. The fact that we found deletions to be less of a problem for students when post-editing is in line with Lee and Liao (2011), but not with Guerberof (2009), who found that accuracy, which contained deletions, was a greater issue for post-editing than for human translation. However, as the category 'accuracy' used by Guerberof encompassed more than just deletions, and as Guerberof worked with professional translators, we cannot directly compare these findings.

Regarding the differences between students and professionals, it is interesting to note that the difference in occurrence of error types between human translation and

post-editing is greater for students than for professional translators. Considering the fact that students' fixation behaviour on source and target text was also more diverse than that of professionals when comparing human translation and post-editing, we can tentatively assume that professional translators treat both processes in a comparable way, whereas students respond to both processes differently, leading to these more diverse results. This is a fascinating finding given our participants and their lack of experience with post-editing, otherwise the findings could simply be attributed to post-editing experience.

What is the impact of machine translation quality on post-editing?

1) What is the impact of overall machine translation quality on post-editing effort?

A decrease in MT quality led to a significant increase in post-editing effort, as measured by the following post-editing effort indicators: duration per word, number of fixations, number of production units, average pause ratio, and HTER. These effects were the same for students and professional translators. We did not find a significant effect of MT quality on the average fixation duration, even though average fixation duration has been shown to correlate with increased effort (Jakobsen & Jensen, 2008) and MT errors (Stymne et al., 2012). Our findings support those of Doherty and O'Brien (2009), who found a higher number of fixations for lower-quality MT output, but no significant differences in fixation duration.

2) What is the impact of specific machine translation errors on post-editing effort?

We found that different types of machine translation errors impact different types of post-editing effort indicators, although coherence issues, other meaning shifts, grammatical and structural issues impact more than one post-editing effort indicator. In line with our expectations, HTER (which is currently frequently used as a measure of post-editing effort) is the only effort indicator that does not have any MT error types in common with other effort indicators. Interestingly, although overall quality had no impact on the average fixation duration, the more specific category 'other meaning shift' was found to have a significant effect on the average fixation duration. Surprisingly, word order never showed up as a significant predictor, even though plenty of studies have suggested word order issues have an effect on total time (Koponen et al., 2012), the number of production units (Alves & Gonçalves, 2013), and even fixations (Stymne et al., 2012). We can only assume that in our study, the effects of other error types outweighed the word order effects.

For this more fine-grained level, the factor 'experience' seems to have greater influence than when looking at the overall MT quality. For students, the average fixation duration increased more than for professional translators with an increase of meaning shifts in the MT output. Meaning shifts were the most common problem for

both students and professional translators for both methods of translation, and were equally common for both participant groups, but this interaction effect shows that students do cognitively process these types of issues differently than professionals. The pause ratio was higher for students than that of professionals regardless of MT output quality, and HTER increased less for students than it did for professional translators with an increase of the number of adequacy other issues. As most 'adequacy' other issues consisted of deletions, and students were found to have more deletions than professionals in their final post-edited output, it is not surprising to see this reflected in lower HTER scores (if a deletion is not edited, it does not lead to a higher edit rate).

3) How does the machine translation output for multi-word units affect post-editing quality?

We introduced 'contrast with the target language' as an additional way of classifying multi-word units in a post-editing context. This addition seems to be a crucial one, as contrastive multi-word units (multi-word units that cannot be translated literally into the target language) contained more errors in the MT output and were subsequently harder to post-edit. The most common issues after post-editing were wrong collocations and adequacy issues, whereas the category 'logical problem' was abundant in the MT output, but much less so after post-editing.

4) How does the machine translation output for multi-word units affect the consultation of external resources during post-editing?

Students' research strategies during post-editing of multi-word units are not as efficient as they could be. A lot of time was spent in external resources even when the machine translation output was correct, and students spent more time in external resources for compounds than for collocations, whereas the latter were found to be more problematic. While not efficient, the research process did seem to be effective, as participants who spent time in external resources mostly managed to correct the machine translation errors.

Attitude

What are the differences in attitude towards human translation and post-editing?

1) How rewarding is post-editing compared to human translation?

In line with expectations, most participants found human translation more rewarding than post-editing (Fulford, 2002), although half of the participants indicated that they did not mind post-editing. Students' feelings seem somewhat more positive towards post-editing than professional translators' (Carl et al., 2015; Kliffer, 2005).

2) How useful is MT output according to translators?

Despite the preference for human translation, all participants in the pretests and in the main experiments indicated that they found MT output 'sometimes' or 'often' useful, in contrast with findings by (Koehn, 2009).

3) Which translation method is perceived as being faster?

In the pretests, half of the students thought post-editing to be faster than human translation, the others' opinions were equally divided between 'human translation is faster' and 'they are equally fast'.

Before participating in the experiment, most of the professional translators thought human translation was faster or just as fast as post-editing (Gaspari et al., 2014), whereas half of the students believed post-editing to be faster than human translation.

4) How is the quality of both methods of translation perceived?

Only three participants in the main experiment believed that post-edited texts would be of higher quality than human translations. In the pretest, twice as many students thought human translation would be better than students who believed both methods could lead to equally high-quality products. This number shifted a little in the main experiment, with an equal number of students indicating they thought human translation and post-editing would be of comparable quality as the number of students indicating that human translation would lead to better quality. More than half of the professional translators, however, assumed both translation methods would lead to products of comparable quality. This is an interesting difference, as we expected students to be somewhat more positive towards post-editing (Moorkens & O'Brien, 2015). A possible explanation could be the concept of 'translator confidence': professional translators are presumably confident they are capable of delivering a high-quality translation, regardless of translation method (Fraser, 2000; House, 2000; Kiraly, 1995).

5) Which translation method is the most preferred translation method?

In the pretests, most students either showed a clear preference for human translation, or a preference for human translation while not minding post-editing.

Comparable to the pretest findings, almost all students participating in the main experiment indicated that they preferred human translation, despite them experiencing post-editing as less or equally tiring to human translation, and despite them being more positive towards post-editing when asked about the most rewarding translation method. The professional translators' opinions were a little more mixed: half of them still preferred human translation, but four participants indicated that they preferred post-editing, whereas only two indicated that they had no preference. Their attitudes

were equally mixed when asked about the least tiring translation method: all three options (human translation; equally tiring; post-editing) were chosen by a comparable number of professional translators. As such, we could not confirm that professional translators perceive post-editing as being more effortful, as suggested by Dragsted (2006) and Guerberof (2013). In contrast with our expectations (Depraetere, 2010; Moorkens & O'Brien, 2015), professional translators seem to prefer post-editing more than students do (Gaspari et al., 2014).

6) Is there a difference in perception before and after the experiment?

As we tentatively assumed on the basis of Garcia (2010), if there was a change in attitude, it was usually in favour of post-editing. In the pretests, more students felt more positive about post-editing after participating than more negative about it, although most students indicated that they felt the same after participating as they had before participating. In the main experiment, there was a change in perceived speed after participating in experiment, mostly in favour of post-editing, although some participants also changed their mind in favour of human translation.

Practical implications

Building on these empirical findings, we can make a few practical suggestions regarding development of translation tools, improving machine translation output to better suit post-editors' needs, and translator training.

Translation tools

As most of a translators' attention goes to processing the target text, it would make sense to add visual clues to the target text in a translation tool, where they are more certain to be noticed. Perhaps the target text itself could be made visually more prominent by making it larger than the source text. Especially in the case of novice translators, who had a higher number of fixations on the target text during post-editing than professional translators, this could lead to a more efficient processing. The ideal translation tool would be tailored to a specific translator's needs, taking into account their personal preferences and experience. Students, for example, seem to rely on dictionaries more often. Integrating lexical information or dictionary searches into a translation tool could help them save time. Knowing that students make more adequacy errors, however, they would have to be made more aware of possible adequacy issues,

for example, by highlighting possible polysemous words in either the source text or the machine translation output. This would especially be necessary when post-editing, considering the higher number of word sense issues students failed to spot, the fact that meaning shifts in MT output were cognitively more demanding to process for students than for professional translators, and the fact that students' increased time in external resources did not reduce the number of adequacy errors, indicating that their current research strategies are not effective. They could further benefit from integrated word alignment information so that deletions in the machine translation output can be highlighted, as students often did not spot these issues during post-editing.

Machine translation quality and post-editing effort

Machine translation developers interested in reducing post-editing effort should not rely solely on HTER scores as a measure of effort. We found that the error types that best predict HTER are different from the ones that predict other effort indicators, such as time and average fixation duration. In addition, HTER was impacted greatly by the presence of deletions, but less so if the post-editors did not solve the deletions, as was often the case for students. HTER is a measure of how many changes a post-editor has made, but does not necessarily correspond to an increase in quality, or an increase in cognitive effort. As coherence issues, meaning shifts, and grammatical and structural issues had the greatest impact on the widest variety of post-editing effort indicators, developers would do best to invest in ways of either detecting these types of effort or reducing the number of times these errors occur. If it is possible to detect them, post-editors could be warned and receive additional visual cues or other types of support when there is a high chance of these errors occurring in a particular MT segment.

Translator training

As post-editing was found to be faster than human translation, while leading to comparable quality, one might wonder whether specific post-editing training is really necessary. However, there still seems to be room for improvement. The most straightforward reason for teaching post-editing is making people aware of its existence and limitations. We found that if participants changed their minds after the experiment, it was usually in favour of post-editing, indicating that understanding indeed leads to acceptance. In addition, we believe the process can be made more efficient, especially for students. When post-editing, their research strategies were not sufficient to reduce the number of adequacy errors, and the significantly higher number of fixations on the target text during post-editing compared to professional translators could also indicate that they do not know what to look for exactly. Post-editor training should focus on

helping students spot typical machine translation errors that currently often go unnoticed (such as meaning shifts, wrong collocations, logical problems, and word sense issues), and ways to solve them. But it is not just students who benefit from hands-on experience with post-editing. The comments professional translators provided after the experiment indicate that they were mostly surprised about *Google Translate*'s quality, and that they enjoyed the post-editing more than they had anticipated.

Methodological suggestions

In addition to practical suggestions, our used methodology could also be of use to other researchers, and we would like to highlight a few important suggestions and concerns.

Translation Quality Assessment

We believe the acceptability-adequacy distinction is an important one to maintain. It is a relatively well-established distinction in translation evaluation (although it sometimes goes by different names) and machine translation evaluation alike. Both error types have a different impact on cognitive effort and, in the case of machine translation, subsequent post-editing. The concept of acceptability-adequacy has been integrated in other metrics as well, although annotators usually have to mark both error types at the same time. We have tested this in our first TQA attempts, but came to comparable conclusions as Stymne and Ahrenberg (2012), that deciding between acceptability and adequacy is really hard when judging both at the same time. By dividing the process into two separate steps, and allowing problematic sections to contain errors of different types, we effectively took away the previous doubts. If the error can be spotted just by looking at the target text, it is an acceptability issue, if the error can be spotted by comparing source and target text, it is an adequacy issue. We further confirmed the need of a consolidation phase with multiple evaluators, as was also suggested by Stymne and Ahrenberg (2012).

Regarding the more practical side of making annotations, we greatly enjoyed using the brat tool (Stenetorp et al., 2012). It is easy configurable with different types of annotation schemes, and it has the option to add subcategories, relations, and notes. For the purpose of our translation quality assessment, it was more than sufficient. An important remark to make, is the fact that brat currently handles plain text only. This was not a problem for our work, as we did not look at layout and other forms of textual

markup, but this is something to take into account when layout is important to the evaluation of the task at hand.

An ideal future version of the brat tool would work seamlessly with word alignment tools, such as the *YAWAT* tool used by the *TPR-DB* (Germann, 2008). As such, it would become easier to map errors back to source text words and segments, and include this quality information in the different data files for further analysis. In our study, a lot of work still had to be done either manually or with self-written scripts, but a more permanent solution would of course be better.

Logging external resources

Where other researchers have worked with screen capture software that was manually annotated for the usage of external resources, we decided to use *Inputlog* (Leijten & Waes, 2008) to automatically register what goes on outside of the main translation interface. Processing this data still required some recoding (labelling different websites with the correct type of external resource), but we do believe the work was easier and more accurate than working with screen recordings. Ours is the first study to integrate *CASMACAT* with *Inputlog*, offering the possibility of an even more holistic translation process analysis than hitherto possible.

Although easier in many ways, the usage of *Inputlog* without screen capture had a few drawbacks we had not anticipated. The tool registers the name of a screen or tool whenever it is opened. It happened that a translator opened a webpage to look up a certain word, and then returned to that same webpage at a later point, either to look for another word, or to open a different type of resource on the same webpage. In these cases, *Inputlog* first registers the old page name, and then the new page name. Simply counting the number of sources without looking at the time spent in these sources is therefore not always accurate. Another issue is the issue we encountered with the *Van Dale* dictionary. *Inputlog* registers the name of the webpage as is, and while most dictionaries include the search query and language combination at the top, the descriptor of the *Van Dale* website simply says 'Home - Van Dale'. *Inputlog* does register keystrokes, so this helped us identify the search words, but we could not determine which type of dictionary translators used. *Van Dale* is a bilingual as well as a monolingual dictionary, and the participants' selection did not show up in the page descriptor. We therefore suggest running some pretests when using the tool, to see how different types of external resources show up. In addition, if this type of information is important to the research, a screen recording could still be added as a backup for verification purposes rather than as the main source of information.

Limitations and future work

We worked on translation and post-editing for general text types from English into Dutch. Research on other language combinations and text types is needed to verify whether our findings can be extrapolated to other text types and languages. The same goes for the choice of machine translation system. We worked with *Google Translate* as this system produces good results for general translations from English into Dutch. Other machine translation systems would of course lead to other results. While not all of our more practical suggestions will necessarily carry over to different contexts, we do believe our proposed methodological suggestions can be of use to other researchers.

Even though we tried to take as many factors into account in our analyses, we could not control for everything, nor analyse all the data we gathered. While our design was balanced, ideally, we would have had sixteen students and sixteen professional translators, to truly cover all possible variations of the experiment. We could not find sufficient participants during the period the sessions were planned, and so we had to work with the number we had. As with most research, the more data, the more robust the findings, and it would be interesting to see our findings confirmed in a larger-scale study. Future studies should also look at different types of participants. Even though we found some processing differences between professional translators and students, there was no significant difference in final quality or time needed. As we did find a significant correlation between the percentage of professional translators' work consisting of general text translation and final quality, we believe that specialisation and confidence are intriguing factors to take into account for future work.

The richness of the data we gathered had the advantage of allowing us to observe human translation and post-editing in detail from many different angles, using data from different sources. The disadvantage of that same richness, however, is the fact that it is impossible to analyse all available data and take all possible angles into account. The *TPR-DB* files themselves already contain an abundance of additional information that could be looked at in future work. Regarding the translation and post-editing process, we did not currently look at different stages (such as drafting stage and revision stage) or strategies (reading source text first or starting to work on the target text immediately). Keystrokes could also provide interesting additional information. We found post-editing to be faster than human translation, but the time spent in external resources was not significantly shorter, raising the question where exactly this time gain comes from. A decrease in typing activity could be a possible answer, but we need to look at typing activity (and take into account typing proficiency and speed) to verify this assumption. When studying problem-solving during translation and post-editing, it would be interesting to take pause and fixation data into account as well, to see how problems are processed: is there a longer pause, a regression in the text, or a shift to the

source text before participants turn to external resources? We also wish to link our findings back to the data from the retrospective session. Participants highlighted passages they found problematic, and it would be interesting to verify whether their actual behaviour when processing these passages was indicative of this perceived difficulty. We further wish to see whether different participants selected the same passages as being difficult or not.

A translation robot for each translator?

The title question has been called 'a real teaser', and it is, purposely so. It is a deliberately open question that can be interpreted in many ways, in part to keep the reader aware of a possible greater picture, in part to keep the reader critical of this same greater picture: is this really where we are going? Is this where we want to go?

In the section below, we discuss just a few aspects of the title question and the first steps the research presented in this dissertation has taken to answering it. As we only looked at certain aspects of this broader question, some of the answers inevitably contain speculation.

What would a translation robot look like?

While mechanical robots such as the ones depicted on the cover would certainly be fun to have around, a translation robot would be more like a bot: a software agent. A translation robot would be different from current translation environment tools in that it would be able to make autonomous decisions about the user interface and translation process, using a variety of inputs.

It would be able to automatically assess the provided source text and adapt the translation environment to the translator's needs. For example, the metadata of a certain text could let the system know that a particular lexicon should be integrated.

The translation robot would be aware of potential difficulties in a source text or MT output and would provide the translator with appropriate feedback. It could, for example, as suggested in a previous section, highlight words and offer dictionary suggestions for polysemous words.

What would really set the translation robot apart, however, would be its ability to interpret and respond to user activity data. An increase in production units, more fixations, and more pauses were all found to correlate with a decrease in MT quality, and are an indication of increased effort. Upon noticing these increases, a translation

robot could provide the translator with additional information to help solve common problems such as grammatical or coherence issues, for example, by highlighting linguistic markers of coherence in the text. Alternatively, when the effort reaches a certain threshold, the robot could decide to remove the suggested MT output so that the translator could work from scratch.

Can all types of translators benefit from using a translation robot?

We have shown that post-editing is faster and cognitively less demanding than human translation for English-Dutch translation of general text types, without reducing the translation's quality, for professional translators and students alike. This already shows the benefits of integrating MT into a translation tool for different types of translators, but a translation robot would of course do more. In the more specific analysis on multi-word units, we found that students' research strategies are not efficient. By picking up on different behavioural signals, a translation robot could offer a translator either the right tools to solve the problem, or an assurance of a translation's quality in the form of, for example, a quality estimation score.

As we only took experience into account as translator type, it remains to be seen whether, for example, using translators with different types of specialization would lead to different results.

Do translators want a translation robot?

Most participants in our study preferred human translation over post-editing, but all participants found machine translation output useful. Finding technology useful is a first step towards the acceptance of technology (Dillon, 2001), and so we assume translators to at least accept machine translation in their work. Other steps include experience, training, and implementation, for which the research presented in this dissertation offered insights and suggestions.

Machine translation would of course only be a small aspect of a translation robot. The experimental evidence and translators' feedback show that the addition of other aspects could be beneficial as well. The sometimes inefficient search strategies, increase in effort for low-quality MT output and errors in the final product indicate that different types of process-monitoring and assistance could be helpful. In addition, participants mentioned the need for spell-checkers, integration of their own preferred dictionaries, and automatic completion as ways to improve the translation environment. The desire for customisation was expressed by multiple participants, and a translation robot capable of automatically adapting to a given situation or person's preferences would therefore be the ideal solution.

How feasible is the creation of a translation robot?

In essence, a translation robot is a more advanced, automated version of a translation environment tool. While a lot more research is needed to either create or improve the envisioned aspects necessary for such a robot to be useful to translators, it is not impossible to imagine a translation robot as the future of the translation environment tool.

Recent translation environment tools such as Lilt already make use of automated and adaptive processes such as automatic completion during typing.

Some translation environment tools use keystroke logging to generate process reports after a translation, but this functionality could also be used during the translation process itself to alert the system of translation difficulties and help it select the appropriate support.

Eye trackers are not currently a part of translation environment tools, but as they are becoming ever more affordable and compact, it is not hard to imagine they might be in the future. This is especially true if we consider how a translation robot could benefit from the addition of fixation information: it could detect regressions and help solve coherence issues or it could detect longer than average fixations and provide additional information or translation suggestions as needed.

Concluding remarks

While there will always be more to discover and investigate, our work brings us closer to a better understanding of the benefits and limitations of post-editing. An understanding which will, ultimately, help us develop technologies that work for the people using them, making translation, as Kay (1980, reprint 1997) envisioned it, "more rewarding, more exciting, more human" (p. 1).

Bibliography

Akaike, H. (1974). A new Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723.

Alabau, V., Bonk, R., Buck, C., Carl, M., Casacuberta, F., García-Martínez, M., . . . Tsoukala, C. (2013). CASMACAT: An Open Source Workbench for Advanced Computer Aided Translation. *The Prague Bulletin of Mathematical Linguistics, 100*, 101-112.

Alves, F. (Ed.) (2003). *Triangulating translation: perspectives in process oriented research* (Vol. 45). Amsterdam/Philadelphia: John Benjamins.

Alves, F., & Campos, T. L. (2009). Translation technology in time: investigating the impact of translation memory systems and time pressure on types of internal and external support. In S. Göpferich, A. L. Jakobsen, & I. M. Mees (Eds.), *Behind the mind: Methods, models and results in translation* (pp. 191-218). Frederiksberg C: Samfundslitteratur.

Alves, F., & Gonçalves, J. L. V. R. (2013). Investigating the conceptual-procedural distinction in the translation process. *Target, 25*(1), 107-124.

Angelone, E., & Shreve, G. M. (2011). Uncertainty management, metacognitive bundling in problem-solving and translation quality. In S. O'Brien (Ed.), *Cognitive Explorations of Translation* (pp. 108-130). London & New York: Continuum.

Aranberri, N., Labaka, G., Diaz de Ilarraza, A., & Sarasola, K. (2014). *Comparison of post-editing productivity between professional translators and lay users*. Paper presented at the AMTA 2014 3rd Workshop on Post-editing Technology and Practice (WPTP-3), Vancouver, Canada.

Asadi, P., & Séguinot, C. (2005). Shortcuts, Strategies and General Patterns in a Process Study of Nine Professionals. *Meta, 50*(2), 522-547.

ATA (Producer). (2009). Flowchart for error point decisions. Retrieved from http://www.atanet.org/certification/aboutexams_flowchart.pdf

Austermühl, F. (2001). *Electronic Tools for Translators*. Manchester: St. Jerome.

Aziz, W., De Sousa, S., & Specia, L. (2012). *PET: a tool for post-editing and assessing machine translation*. Paper presented at the 8th International Conference on Language Resources and Evaluation (LREC'12).

Babych, B., Elliott, D., & Hartley, T. (n.d.). *A Comparative Evaluation of Two Machine Translation Systems*. Retrieved from http://www.translution.com/TranslutionDownloads/Softwares/CaseStudies/Is%20MT%20%20fit%20for%20purpose-Aug04.pdf

Banerjee, S., & Lavie, A. (2005). *METEOR: An automatic metric for MT evaluation with improved correlation with human judgments*. Paper presented at the ACL 05 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, Ann Arbor, Michigan.

Barreiro, A., Monti, J., Orliac, B., & Batista, F. (2013). *When multiwords go bad in machine translation*. Paper presented at the MT Summit XIV Workshop on Multi-word Units in Machine Translation and Translation Technology (MUMTTT 2013), Nice, France.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7. Retrieved from http://CRAN.R-project.org/package=lme4

Belam, J. (2003). *'Buying up to falling down': a deductive approach to teaching post-editing*. Paper presented at the MT Summit IX workshop on Teaching Translation Technologies and Tools (T4), New Orleans.

Bentivogli, L., Bertoldi, N., Cettolo, M., Federico, M., Negri, M., & Turchi, M. (2016). On the evaluation of adaptive machine translation for human post-editing. *IEEE/ACMtransactions on audio, speech, and language processing, 24*(2), 388-399.

Berka, J., Bojar, O., Fishel, M., Popovic, M., & Zeman, D. (2012). *Automatic MT error analysis: Hjerson helping Addicter*. Paper presented at the 8th International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey.

Bernardini, S. (2001). Think-aloud protocols in translation research. *Target, 13*(2), 241-263.

Bojar, O., Buck, C., Callison-Burch, C., Federman, C., Haddow, B., Koehn, P., . . . Specia, L. (2013). *Findings of the 2013 Workshop on Statistical Machine Translation*. Paper presented at the ACL 2013 Eight Workshop on Statistical Machine Translation (WMT13), Sofia, Bulgaria.

Bowker, L., & Buitrago Ciro, J. (2015). Investigating the usefulness of machine translation for newcomers at the public library. *Translation and Interpreting Studies, 10*(2), 165-186.

Broekkamp, H., & van den Bergh, H. (1996). Attention strategies in revising a foreign language text. In G. Rijlaarsdam, H. van den Bergh, & M. Couzijn (Eds.), *Theories, models and methodology in writing research* (pp. 170-181). Amsterdam: Amsterdam University Press.

Burnham, K., & Anderson, D. (2004). Multimodel inference: understanding AIC and BIC in Model Selection. *Sociological Methods & Research, 33*, 261-304.

Carl, M., Dragsted, B., Elming, J., Hardt, D., & Jakobsen, A. L. (2011). The process of post-editing: A pilot study. In B. Sharp, M. Zock, M. Carl, & A. L. Jakobsen (Eds.), *Proceedings of the 8th International NLPCS Workshop* (pp. 131-142). Frederiksberg C: Samfundsliterattur.

Carl, M., Dragsted, B., & Jakobsen, A. L. (2011). A Taxonomy of Human Translation Styles. *Translation Journal, 16*(2).

Carl, M., Gutermuth, S., & Hansen-Schirra, S. (2015). Post-editing machine translation: Efficiency, strategies, and revision processes in professional translation settings. In A. Ferreira & J. W. Schwieter (Eds.), *Psycholinguistic and Cognitive Inquiries into Translation and Interpreting* (pp. 145-174). Amsterdam: John Benjamins.

Carl, M., Schaeffer, M., & Bangalore, S. (2016). The CRITT Translation Process Research Database. In M. Carl, M. Schaeffer, & S. Bangalore (Eds.), *New Directions in Emprical Translation Process Research. Exploring the CRITT TPR-DB* (pp. 13-54). Switzerland: Springer International Publishing.

Chodkiewicz, M. (2015). Undergraduate Students' Use of External Sources in Revising and Justifying Their Translation Decisions Based on Instructor Feedback. *Lublin Studies in Modern Languages and Literature, 39*(2), 124-141.

Colina, S. (2009). Further evidence for a functionalist approach to translation quality evaluation. *Target, 21*(2), 235-264. doi:10.1075/target.21.2.02col

Daems, J., Carl, M., Vandepitte, S., Hartsuiker, R., & Macken, L. (2016). The effectiveness of consulting external resources during translation and postediting of general text

types. In M. Carl, S. Bangalore, & M. Schaeffer (Eds.), *New directions in empirical translation process research : exploring the CRITT TPR-DB* (pp. 111-133): Springer.

Daems, J., Macken, L., & Vandepitte, S. (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*. Paper presented at the MT Summit XIV 2nd Workshop on Post-editing Technology and Practice (WPTP-2), Nice, France.

Daems, J., Macken, L., & Vandepitte, S. (2014). *On the origin of errors: a fine-grained analysis of MT and PE errors and their relationship*. Paper presented at the 9th International Conference on Language Resources and Evanluation (LREC'14).

Daems, J., Vandepitte, S., Hartsuiker, R., & Macken, L. (2015). *The impact of machine translation error types on post-editing effort indicators*. Paper presented at the MT SUMMIT XV 4th Workshop on Post-Editing Technology and Practice (WPTP-4), Miami, USA.

Daems, J., Vandepitte, S., Hartsuiker, R., & Macken, L. (2016a). How do students cope with machine translation output of Multi-Word Units? An exploratory study. In R. Mitkov, J. Monti, G. Corpas Pastor, & V. Seretan (Eds.), *Multi-word Units in Machine Translation and Translation Technology*. Amsterdam: John Benjamins.

Daems, J., Vandepitte, S., Hartsuiker, R., & Macken, L. (2016b). Translation Methods and Experience: A Comparative Analysis of Human Translation and Post-editing with Students and Professional Translators. *Meta*.

Dancette, J. (2007). Mapping meaning and comprehension processes in translation. In J. Danks, G. M. Shreve, S. Fountain, & M. McBeath (Eds.), *Cognitive Processes in Translation and Interpreting*. Kent: Sage.

de Almeida, G. (2013). *Translating the post-editor: an investigation of post-editing changes and correlations with professional experience across two Romance languages* (Doctoral dissertation), Dublin City University, Dublin.

de Almeida, G., & O'Brien, S. (2010). *Analysing Post-Editing Performance: Correlations with Years of Translation Experience*. Paper presented at the 14th annual conference of the European Association for Machine Translation (EAMT 2010), St. Raphaël, France.

de Souza, J., Turchi, M., & Negri, M. (2014). *Machine translation quality estimation across domains*. Paper presented at the 25th International Conference on Computational Linguistics: Technical Papers (COLING 2014), Dublin, Ireland.

Demiris, G., Rantz, M., Aud, M., Marek, K., Tyrer, H., Skubic, M., & Hussam, A. (2004). Older adults' attitudes towards and perceptions of 'smart home' technologies: a pilot study. *Medical Informatics & The Internet in Medicine, 29*(2), 87-94.

Denkowski, M., & Lavie, A. (2012). *Challenges in Predicting Machine Translation Utility for Human Post-Editors.* Paper presented at the 10th Conference of the Association for Machine Translation in the Americas (AMTA 2012).

Depraetere, I. (2010). *What counts as useful advice in a university post-editing training context? Report on a case study*. Paper presented at the 14th annual conference of the European Association for Machine Translation (EAMT 2010), St Raphael, France.

Dillinger, M. (2014). Introduction. In S. O'Brien, L. Winther-Balling, M. Carl, M. Simard, & L. Specia (Eds.), *Post-editing of Machine Translation: Processes and Applications* (pp. ix-xv). Newcastle Upon Tyne: Cambridge Scholars Publishing.

Dillon, A. (2001). User acceptance of information technology. In W. Karwowski (Ed.), *Encyclopedia of Human Factors and Ergonomics*. London: Taylor and Francis.

Doherty, S. (2016). The impact of translation technologies on the process and product of translation. *International Journal of Communication, 10*, 647-969.

Doherty, S., & Kenny, D. (2014). The design and evaluation of a Statistical Machine Translation syllabus for translation students. *The Interpreter and Translator Trainer, 8*(2), 295-315.

Doherty, S., & O'Brien, S. (2009). *Can MT Output be Evaluated through Eye Tracking?* Paper presented at the 12th Machine Translation Summit (MT Summit XII), Ottawa, Canada.

Doherty, S., O'Brien, S., & Carl, M. (2010). Eye tracking as an MT evaluation technique. *Machine Translation, 24*, 1-13.

Dragsted, B. (2006). Computer-aided translation as a distributed cognitive task. *Pragmatics & Cognition, 14*(2), 443-464.

Dragsted, B. (2010). Coordination of reading and writing processes in translation: An eye on uncharted territory. In G. M. Shreve & E. Angelone (Eds.), *Translation and Cognition*. Amsterdam/Philadelphia: John Benjamins.

Drugan, J. (2013). *Quality in Professional Translation: Assessment and Improvement*. London: Bloomsbury.

Ehrensberger-Dow, M., & Künzli, A. (2010). Methods of accessing metalinguistic awareness: A question of quality? In S. Göpferich, F. Alves, & I. M. Mees (Eds.), *New approaches in translation process research* (pp. 113-132). Frederiksberg C: Samfundslitteratur.

Eyckmans, J., Anckaert, P., & Segers, W. (2009). The perks of norm-referenced translation evaluation. In C. Angelelli & H. E. Jacobson (Eds.), *Testing and Assessment in Translation and Interpreting* (pp. 73-93). Amsterdam: John Benjamins.

Farrús, M., Costa-jussà, M., & Mariño, J. (2011). Overcoming statistical machine translation limitations: error analysis and proposed solutions for the Catalan-Spanish language pair. *Language Resources & Evaluation, 45*(2), 181-208.

Federico, M., Bertoldi, N., Cettolo, M., Negri, M., Turchi, M., Trombetti, M., . . . Germann, U. (2014). *The MateCat Tool*. Paper presented at the 25th International Conference on Computational Linguistics (COLING 2014), Dublin, Ireland.

Fiederer, R., & O'Brien, S. (2009). Quality and Machine Translation: A realistic objective? *The Journal of Specialised Translation, 11*, 52-74.

Fomicheva, A. (2015). *Google translates översättningsteknik från svenska till ryska: en analys av fyra översatta texter med olika bakgrunder.* (Bachelor's thesis), Tartu Ülikool, Tartu.

Fraser, J. (2000). What do real translators do? Developing the use of TAPs from professional translators. In S. Tirkkonen-Condit & R. Jääskeläinen (Eds.), *Tapping and mapping the processes of translation and interpreting* (Vol. 37, pp. 111-122). Amsterdam: John Benjamins.

Fulford, H. (2002). *Freelance Translators and Machine Translation: an Investigation of Perceptions, Uptake, Experience and Training Needs*. Paper presented at the 6th annual conference of the European Association for Machine Translation Workshop on Teaching Machine Translation, Manchester, United Kingdom.

Fulford, H., & Granell-Zafra, J. (2004). Translation and Technology: a Study of UK Freelance Translators. *JoSTrans*(4).

Gambier, Y. (2014). Changing landscape in translation. *International Journal of Society, Culture & Language, 2*(2), 1-12.

Gambier, Y. (2016). Rapid and radical changes in translation and translation studies. *International Journal of Communication, 10*, 887-906.

Garcia, I. (2010). Is machine translation ready yet? *Target, 22*(1), 7-21.

Garcia, I. (2011). Translating by post-editing: is it the way forward? *Machine Translation, 25*, 217-237.

Gaspari, F., Toral, A., Naskar, S. K., Groves, D., & Way, A. (2014). *Perception vs Reality: Measuring Machine Translation Post-Editing Productivity* Paper presented at the AMTA 2014 3rd Workshop on Post-editing Technology and Practice (WPTP-3).

Gerloff, P. (1988). *From French to English: A Look at the Translation Process in Students, Bilinguals, and Professional Translators*. doctoral dissertation. Harvard University, University Microfilms International

Germann, U. (2008). *Yawat: Yet Another Word Alignment Tool*. Paper presented at the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies Demo Session (Companion Volume) (ACL-08: HLT), Columbus.

Gibb, D. K. (1985). *Computer Linguistics and Translation*. Paper presented at the the Savonlinna School of Translation Studies.

Göpferich, S. (2009). Towards a model of translation competence and its acquisition: the longitudinal study *TransComp*. In S. Göpferich, A. L. Jakobsen, & I. M. Mees (Eds.), *Behind the mind: Methods, models and results in translation process researcg* (pp. 11-38). Frederiksberg C: Samfundslitteratur.

Göpferich, S. (2010). The translation of instructive texts from a cognitive perspective. In F. Alves, S. Göpferich, & I. M. Mees (Eds.), *New approaches in Translation Process Research* (pp. 5-65). Frederiksberg C: Samfundslitteratur.

Göpferich, S., Jakobsen, A. L., & Mees, I. M. (Eds.). (2009). *Behind the mind: Methods, models and results in translation process research*. Frederiksberg C: Samfundslitteratur.

Gouadec, D. (1981). Paramètres de l'évaluation des traductions. *Meta, 26*(2), 99-116.

Green, S., Chuang, J., Heer, J., & Manning, C. D. (2014). *Predictive Translation Memory: A mixed-initiative system for human language translation*. Paper presented at the 27th annual ACM symposium on User interface software and technology, Honolulu, USA.

Green, S., Heer, J., & Manning, C. D. (2013). *The Efficacy of Human Post-Editing for Language Translation*. Paper presented at the ACM Human Factors in Computing Systems (CHI), Paris, France.

Groves, D., & Schmidtke, D. (2009). *Identification and analysis of post-editing patterns for MT*. Paper presented at the 12th Machine Translation Summit (MT Summit XII), Ottawa, ON.

Guerberof, A. (2009). *Productivity and quality in MT post-editing*. Paper presented at the MT Summit XII Beyond Translation Memories Workshop (WS3), Ottawa, Ontario, Canada.

Guerberof, A. (2012). *Productivity and quality in the post-editing of outputs from translation memories and machine translation.* Universitat Rovira i Virgili, Spain, Tarragona.

Guerberof, A. (2013). What do professional translators think about postediting? *JoSTrans, 19*, 75-95.

Guerra Martínez, L. (2003). *Human Translation versus Machine Translation and Full Postediting of Raw Machine Translation Output.* (Master's thesis), Dublin City University, Dublin.

Hailin, W., Hanhui, L., & Zhurnei, S. (2010). *Fatigue Driving Detection System Design Based on Driving Behavior* Paper presented at the International Conference on Optoelectronics and Image Processing (ICOIP 2010), Haiko, Hainan, China.

Hansen, G. (2010a). Integrative Description in Translation process Research. In G. M. Shreve & E. Angelone (Eds.), *Translation and Cognition* (pp. 189-211). Amsterdam / Philadelphia: John Benjamins.

Hansen, G. (2010b). Translation 'errors'. In Y. Gambier & L. V. Doorslaer (Eds.), *Handbook of Translation Studies* (Vol. 1, pp. 385-388).

Hayes, J., Flower, L., Scriver, K., Stratman, J., & Carey, L. (1987). Cognitive Processing in Revision. In S. Rosenberg (Ed.), *Advances in Applied Psycholinguistics: Vol. 2. Reading, writing, and language processes* (pp. 176-240). New York: Cambridge University Press.

House, J. (2000). Consciousness and the Strategic Use of Aids in Translation. In S. Tirkkonen-Condit & R. Jääskeläinen (Eds.), *Tapping and Mapping the Processes of Translation and Interpreting: Outlooks on Empirical Research* (Vol. 37). Amsterdam: John Benjamins.

Hunziker Heeb, A. (2012). The problem-solving processes of experienced and nonexperienced translators. In S. Kersten, C. Ludwig, D. Meer, & B. Rüschoff (Eds.), *Language learning and language use - applied linguistics approaches: Papers Selected from the Junior Research Meeting Essen 2011* (pp. 177-186). Duisburg: UVVR Universitätsverlag Rhein-Ruhr.

ITR. (2002). ITR announces breakthrough in translation evaluation technology. Retrieved from http://www.itr.co.uk/bulletins/itr-announces-breakthrough-translation-evaluation-technology

Jääskeläinen, R. (1990). *Features of Successful Translation Processes: A Think-aloud Protocol Study.* (Licentiate thesis), University of Joensuu, Savonlinna School of Translation Studies.

Jääskeläinen, R. (1996). Hard Work Will Bear Beautiful Fruit. A Comparison of Two Think-Aloud Protocol Studies. *Meta, 41*(1), 60-74.

Jääskeläinen, R. (2002). Think-aloud protocol studies into translation: An annotated bibliography. *Target, 14*(1), 107-136.

Jääskeläinen, R. (2010). Are All Professionals Experts? Definitions of Expertise and Reinterpretation of Research Evidence in Process Studies. In G. M. Shreve & E. Angelone (Eds.), *Translation and Cognition* (pp. 213-262). Amsterdam / Philadelphia: John Benjamins.

Jääskeläinen, R., & Tirkkonen-Condit, S. (1991). Automatised Processes in Professional vs. Non-professional translation: A think-aloud protocol study. In S. Tirkkonen-Condit (Ed.), *Empirical Research in Translation and Intercultural Studies* (pp. 89-109). Tübingen: Gunter Narr.

Jakobsen, A. L. (2003). Effects of think aloud on translation speed, revision, and segmentation. In F. Alves (Ed.), *Triangulating Translation: Perspectives in Process Oriented Research* (pp. 69-96). Amsterdam: John Benjamins.

Jakobsen, A. L. (2006). Research methods in translation: Translog. In K. P. H. Sullivan & E. Lindgren (Eds.), *Computer keystroke logging and writing: methods and applications* (Vol. 18, pp. 95-105). Amsterdam: Elsevier.

Jakobsen, A. L., & Jensen, K. T. H. (2008). Eye movement behaviour across four different types of reading task. In S. Göpferich, A. L. Jakobsen, & I. M. Mees (Eds.), *Looking at Eyes: EyeTracking Studies of Reading and Translation Processing* (pp. 103-124). Frederiksberg C: Samfundslitteratur.

Jensen, A. (1999). Time pressure in translation. In G. Hansen (Ed.), *Probing the process in translation: methods and results* (Vol. 24, pp. 103-119). Frederiksberg C: Samfundslitteratur.

Jensen, K. T. H. (2009). Indicators of text complexity. In S. Göpferich, A. L. Jakobsen, & I. M. Mees (Eds.), *Behind the mind: Methods, models and results in translation process research* (pp. 61-80). Frederiksberg C: Samfundslitteratur.

Jimenez-Crespo, M. A. (2011). A corpus-based error typology: towards a more objective approach to measuring quality in localization. *Perspectives: Studies in Translatology, 19*(4), 315-338. doi:10.1080/0907676X.2011.615409

Just, M., & Carpenter, P. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*(4), 329-354.

Kay, M. (1980, reprint 1997). The proper place of men and machines in language translation. *Machine Translation, 12*, 3-23.

Kelly, N. (2014, 06/19/2014). Why so many translators hate translation technology. *The Huffington Post*. Retrieved from http://www.huffingtonpost.com/nataly-kelly/why-so-many-translators-h_b_5506533.html

Kiraly, D. (1995). *Pathways to Translation: Pedagogy and Process*. Kent, Ohio: Kent State University Press.

Kliffer, M. (2005). *An Experiment in MT Post-Editing by a Class of Intermediate/Advanced French Majors.* Paper presented at the 10th annual conference of the European Association for Machine Translation (EAMT 2005), Budapest.

Koehn, P. (2009). A Process Study of Computed Aided Translation. *Machine Translation, 23*(4), 241-263.

Koglin, A. (2015). An empirical investigation of cognitive effort required to post-edit machine translated metaphors compared to the translation of metaphors. *The International Journal for Translation & Interpreting Research, 7*(1), 126-141.

Koponen, M. (2012). *Comparing human perceptions of post-editing effort with post-editing operations.* Paper presented at the 7th workshop on statistical machine translation (NAACL 2012), Montréal, Canada.

Koponen, M. (2016a). Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. *JoSTrans*(25), 131-148.

Koponen, M. (2016b). *Machine Translation Post-editing and Effort Empirical Studies on the Post-editing Process.* (Doctoral dissertation), University of Helsinki.

Koponen, M., Aziz, W., Ramos, L., & Specia, L. (2012). *Post-editing Time as a Measure of Cognitive Effort.* Paper presented at the AMTA 2012 Workshop on Post-editing Technology and Practice (WPTP), San Diego, California.

Krings, H. (2001). *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes* (G. Koby, G. M. Shreve, K. Mischerikov, & S. Litzer, Trans. G. Koby Ed.). Kent, Ohio, & London: The Kent State University Press.

Kußmaul, P., & Tirkkonen-Condit, S. (1995). Think-Aloud Protocol Analysis in Translation Studies. *TTR : traduction, terminologie, rédaction, 8*(1), 177-199.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2014). lmerTest: Tests in Linear Mixed Effects Models. R package version 2.0-20. Retrieved from http://CRAN.R-project.org/package=lmerTest

Lacruz, I., Shreve, G. M., & Angelone, E. (2012). *Average Pause Ratio as an Indicator of Cognitive Effort in Post-Editing: A Case Study.* Paper presented at the AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP), Stroudsburg, PA.

Larose, R. (1998). Méthodologie de l'évaluation des traductions. *Meta, 43*(2), 163-186.

Laukkanen, J. (1993). *Routine vs. Non-routine Processes in Translation: A Think-aloud Protocol Study*. pro gradu thesis. University of Joensuu, Savonlinna School of Translation Studies.

Lee, J., & Liao, P. (2011). A Comparative Study of Human Translation and Machine Translation with Post-editing. *Compilation and Translation Review, 4*(2), 105-149.

Leijten, M., & Van Waes, L. (2013). Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes. *Written Communication, 30*(3), 325-343.

Leijten, M., Van Waes, L., Schriver, K., & Hayes, J. (2014). Writing in the workplace: Constructing documents using multiple digital sources. *Journal of Writing Research, 5*(3), 285-336.

Leijten, M., & Waes, L. V. (2008). Inputlog: new perspectives on the logging of on-line writing processes in a Windows environment. In K. P. H. Sullivan & E. Lindgren (Eds.), *Computer keystroke logging and writing: methods and applications* (pp. 73-94): Emerald.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior research methods, 44*, 325-343.

Letchner, J., Krumm, J., & Horvitz, E. (2006). Trip Router with Individualized Preferences (TRIP): Incorporating Personalization into Route Planning. *American Association for Artificial Intelligence*, 1795-1800.

Luukkainen, T. (1996). *Comparisons of translations made with and without reference material: a think-aloud protocol study.* (Master's thesis), University of Joensuu, Finland.

Macizo, P., & Bajo, M. T. (2006). Reading for repetition and reading for translation: do they involve the same processes? *Cognition, 99*, 1-34.

Macken, L., De Clercq, O., & Paulussen, H. (2011). Dutch Parallel Corpus: a balanced copyright-cleared parallel corpus. *Meta, 56*(2), 374-390.

Macklovitch, E. (2006). *TransType2: The Last Word* Paper presented at the 5th International Conference on Languages Resources and Evaluation (LREC'06), Genoa, Italy.

Macklovitch, E., Lapalme, G., & Gotti, F. (2008). *TransSearch: What are translators looking for?* Paper presented at the 8th Conference of the Association for Machine Translation in the Americas (AMTA 2008), Waikiki, Hawai, USA.

Mareček, D., Rosa, R., Galuščáková, P., & Bojar, O. (2011). *Two-step translation with grammatical post-processing.* Paper presented at the EMNLP 2011 6th Workshop on Statistical Machine Translation, Edinburgh, Scotland, UK.

Martínez, L. G. (2003). *Human translation versus machine translation and full post-editing of raw machine translation output.* (Master's thesis), Dublin City University, Dublin, Ireland.

Mateo, R. M. (2014). A deeper look into metrics for translation quality assessment (TQA): a case study. *miscelánea: a journal of english and american studies, 49*(73-93).

MeLLANGE. (2006). MeLLANGE WP4 Tanslation Error Typology. Retrieved from http://corpus.leeds.ac.uk/mellange/images/mellange_error_typology_en.jpg

Mendoza Rivera, O., Mitkov, R., & Corpas Pastor, G. (2013). *A flexible framework for collocation retrieval and translation from parallel and comparable corpora.* Paper presented at the MT Summit XIV Workshop on Multi-word Units in Machine Translation and Translation Technology (MUMTTT 2013), Nice, France.

Merat, N., & Jamson, H. (2009). *How do drivers behave in a highly automated car?* Paper presented at the 5th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, Iowa City, IA: The University of Iowa.

Meschtscherjakov, A., Wilfinger, D., Scherndl, T., & Tscheligi, M. (2009). *Acceptance of Future Persuasive In-Car Interfaces Towards a More Economic Driving Behaviour*. Paper presented at the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications.

Miller, G. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*(2), 81-97.

Monti, J., Barreiro, A., Elia, A., Marano, F., & Napoli, A. (2011). *Taking on new challenges in multi-word unit processing for machine translation.* Paper presented at the 2nd International Workshop on Free/Open-Source Rule-Based Machine Translation, Barcelona, Spain.

Monti, J., Mitkov, R., Corpas Pastor, G., & Seretan, V. (2013). *Proceedings of the workshop on multi-word units in machine translation and translation technologies*, Nice, France.

Moorkens, J., & O'Brien, S. (2015). *Post-Editing Evaluations: Trade-offs between Novice and Professional Participants* Paper presented at the 18th Conference of the European Association for Machine Translation (EAMT 2015), Antalya, Turkey.

Naert, S. (2013). *Uitbreiden van de annotatietool brat.* (Master), HoGent.

Naskar, S. K., Toral, A., Gaspari, F., & Way, A. (2011). A Framework for Diagnostic Evaluation of MT Based on Linguistic Checkpoints. *13th Machine Translation Summit (MT Summit XIII)*.

Nitzke, J., & Oster, K. (2016). Comparing Translation and Post-editing: An Annotation Schema for Activity Units. In M. Carl, S. Bangalore, & M. Schaeffer (Eds.), *New Directions in Emprical Translation Process Research. Exploring the CRITT TPR-DB* (pp. 293-308). Switzerland: Springer International Publishing.

O'Brien, S. (2002). *Teaching Post-Editing: A Proposal for Course Content*. Paper presented at the 6th annual conference of the European Association for Machine Translation Workshop Teaching Machine Translation, Manchester.

O'Brien, S. (2004). *Machine Translatability and Post-Editing Effort: How do they Relate?* Paper presented at the Translating and the Computer Conference, London.

O'Brien, S. (2006). Pauses as Indicators of Cognitive Effort in Post-Editing Machine Translation Output. *Across languages and cultures, 7*(1), 1-21.

O'Brien, S. (2007). Eye-tracking and Translation Memory Matches. *Perspectives: Studies in Translatology, 14*(3), 185-205.

O'Brien, S. (2012). Towards a Dynamic Quality Evaluation Model for Translation. *The Journal of Specialised Translation*(17), 55-77.

O'Curran, E. (2014). *Translation quality in post-edited versus human-translated segments: a case study*. Paper presented at the AMTA 2014 3rd Workshop on Post-editing Technology and Practice (WPTP-3), Vancouver, Canada.

Och, F., & Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics, 29*(1), 19-51.

Offersgaard, L., Povlsen, C., Almsten, L., & Maegaard, B. (2008). *Domain specific MT in use*. Paper presented at the 12th annual conference of the European Association for Machine Translation (EAMT 2008), HITEC e.V, Vogt-Kölln Strasse 30, Hamburg, Germany.

Optimale. (2012). *Optimale employer survey and consultation*. Retrieved from http://www.ressources.univ-rennes2.fr/service-relations-internationales/optimale/attachments/article/52/WP4_Synthesis_report.pdf

PACTE. (2003). Building a Translation Competence Model. In F. Alves (Ed.), *Triangulating Translation: Perspectives in Process Oriented Research* (pp. 43-66). Amsterdam: John Benjamins.

PACTE. (2005). Investigating Translation Competence: Conceptual and Methodological Issues. *Meta, 50*(2), 609-619.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). *BLEU: a Method for Automatic Evaluation of Machine Translation*. Paper presented at the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia.

Pavlović, N. (2007). *Directionality in collaborative translation processes: a study of novice translators.* (Doctoral dissertation), Universitat Rovira i Virgili, Spain.

Pavlović, N., & Jensen, K. T. H. (2009). Eye tracking translation directionality. In A. Pym & A. Perekrestenko (Eds.), *Translation Research Projects 2* (pp. 101-119). Tarragona, Spain: Intercultural Studies Group.

Pilipovic, M., Spasojevic, D., Velikic, I., & Teslic, N. (2014). *Toward Intelligent Driver-Assist Technologies and Piloted Driving: Overview, Motivation and Challenges* Paper presented at the X International Symposium on Industrial Electronics INDEL, Banja Luka.

Plitt, M., & Masselot, F. (2010). A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. *The Prague Bulletin of Mathematical Linguistics, 93*, 7-16.

Popovic, M., Lommel, A. R., Burchardt, A., Avramidis, E., & Uszkoreit, H. (2014). *Relations between different types of post-editing operations, cognitive effort and temporal effort*. Paper presented at the 17th annual conference of the European Association for Machine Translation (EAMT 2014), Dubrovnik, Croatia.

Portela, R., Mamede, N., & Baptista, J. (2011). *Multiword Identification*. Paper presented at the Terceiro Simpósio de Informáctica (INFORUM 2011), Departamento de Engenharia Informática da Universidade de Coimbra.

Prassl, F. (2010). Translators' decision-making processes in research and knowledge integration. In S. Göpferich, F. Alves, & I. M. Mees (Eds.), *New approaches in translation process research* (pp. 57-81). Frederiksberg C: Samfundsliteratur.

Pym, A. (2013). Translation Skill-Sets in a Machine-Translation Age. *Meta, 58*(3), 487-503.

QTLaunchPad. (2013, 2013 September 3). Multidimensional Quality Metric Quality Issue Types. version 2.5.5. Retrieved from https://docs.google.com/document/d/1E8IR1-8bR_M7VouHQhogUPpP2htpprwMMx1Mk9KPBTI/pub

R Core Team. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org

Raído, V. E. (2014). *Translation and Web Searching*. New York: Routledge.

Reinke, U. (2013). State of the art in translation memory technology. *Translation: Computation, Corpora, Cognition, 3*(1), 27-48.

Rinsche, A., & Portera-Zanotti, N. (2009). *The size of the language industry in the EU*. Retrieved from http://ec.europa.eu/dgs/translation/publications/studies/index_en.htm

Sager, J. (1994). *Language engineering and translation: Consequences of automation*. Amsterdam: John Benjamins.

Schäfer, F. (2003). *MT post-editing: How to shed light on the" unknown task" Experices made at SAP*. Paper presented at the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop (EAMT/CLAW 2003), Dublin, Ireland.

Schütz, J. (2008). *Artificial Cognitive MT Post-Editing Intelligence*. Paper presented at the 8th Conference of the Association for Machine Translation in the Americas (AMTA 2008), Waikiki, Hawai'i, USA.

Screen, B. (2016). What does Translation Memory do to translation? The effect of Translation Memory output on specific aspects of the translation process. *The International Journal for Translation & Interpreting Research, 8*(1), 1-18.

Secara, A. (2005). *Translation evaluation - a state of the art survey.* Paper presented at the eCoLoRe/MeLLANGE Workshop, Leeds, UK.

Séguinot, C. (1989). Understanding Why Translators Make Mistakes. *TTR : traduction, terminologie, rédaction, 2*(2).

Séguinot, C. (1991). A Study of Student Translation Strategies. In S. Tirkkonen-Condit (Ed.), *Empirical Research in Translation and Intercultural Studies* (pp. 79-88). Tübingen: Gunter Narr.

Seretan, V. (2015). *Multi-word expressions in user-generated content: How many and how well translated? Evidence from a post-editing experiment*. Paper presented at the EUROPHRAS2015 2nd Workshop on Multi-word Units in Machine Translation and Translation Technology (MUMTTT2015), Malaga, Spain.

Shiyab, S. (2010). Chapter One: Globalization and its impact on Translation. In S. Shiyab, M. G. Rose, J. House, & J. Duval (Eds.), *Globalization and Aspects of Translation*. Newcastle upon Tyne, UK: Cambridge Scholars Publishing.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). *A Study of Translation Edit Rate with Targeted Human Annotation*. Paper presented at the 7th Conference of the Association for Machine Translation in the Americas (AMTA 2006), Cambridge.

Somers, H. (2003a). Translation memory systems. In H. Somers (Ed.), *Computers and Translation: A translator's guide* (Vol. 35, pp. 31-48). Amsterdam: John Benjamins.

Somers, H. (2003b). The translator's workstation. In H. Somers (Ed.), *Computers and Translation: A translator's guide* (Vol. 35, pp. 13-30). Amsterdam: John Benjamins.

Sommers, N. (1980). Revision Strategies of Student Writers and Experienced Adult Writers. *College Composition and Communication, 31*(4), 378-388.

Specia, L., & Farzindar, A. (2010). *Estimating Machine Translation Post-Editing Effort with HTER.* Paper presented at the 2nd Joint EM+CNGL Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry", Denver, CO.

Stenetorp, P., Pyysalo, S., Topic, G., Ohta, T., Anaiadou, S., & Tsujii, J. i. (2012). *brat: a Web-based Tool for NLP-Assisted Text Annotation*. Paper presented at the Demonstrations Session at EACL 2012.

Stymne, S., & Ahrenberg, L. (2012). *On the practice of error analysis for machine translation evaluation.* Paper presented at the 8th International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey.

Stymne, S., Danielsson, H., Bremin, S., Hu, H., Karlsson, J., Lillkull, A. P., & Wester, M. (2012). *Eye Tracking as a Tool for Machine Translation Error Analysis*. Paper presented at the 8th International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey.

Sun, S. (2015). Measuring translation difficulty: theoretical and methodological considerations. *Across languages and cultures, 16*(1), 29-54.

Sweller, J. (1988). Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science, 12*, 257-285.

Tatsumi, M. (2010). *Post-Editing Machine Translated Text in a Commercial Setting: Observation and Statistical Analysis.* (Doctoral dissertation), Dublin City University.

TAUS. (2013a). *Quality Evaluation using Adequacy and/or Fluency Approaches*. Retrieved from https://www.taus.net/downloads/finish/57-articles/428-adequacy-fluency-guidelines

TAUS. (2013b). *Quality Evaluation Using an Error Typology Approach*. Retrieved from https://evaluation.taus.net/component/jdownloads/finish/5-articles/10-error-typology-guidelines

Taylor, M., Woolley, J., & Zito, R. (2000). Integration of the global positioning system and geographical information systems for traffic congestion studies. *Transportation research part C, 8*, 257-285.

Temnikova, I. (2010). *Cognitive Evaluation Approach for a Controlled Language Post-Editing Experiment.* Paper presented at the 7th International Conference on Language Resources and Evaluation (LREC'10), Valetta, Malta.

Tirkkonen-Condit, S. (1990). Professional vs. Non-professional Translation: A Think-Aloud Protocol Study. In M. A. K. Halliday, J. Gibbons, & H. Nicholas (Eds.), *Learning, Keeping and Using Language: Selected Papers from the Eighth World Congress of Applied Linguistics* (pp. 381-394). Amsterdam: John Benjamins.

Tondeleir, L. (2013). *Machine translation: Friend or Foe?* (Master's thesis), HoGent, Ghent.

Toury, G. (1995). The Nature and Role of Norms in Translation. In G. Toury (Ed.), *Descriptive Translation Studies and Beyond* (pp. 53-69). Amsterdam-Philadelphia: John-Benjamins.

Valotkaite, J., & Asadullah, M. (2012). *Error Detection for Post-editing Rule-based Machine Translation.* Paper presented at the AMTA 2012 Workshop on Post-editing Technology and Practice (WPTP), San Diego, USA.

Vilar, D., Xu, J., D'Haro, L. F., & Ney, H. (2006). *Error analysis of machine translation output*. Paper presented at the 5th International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy.

Waddington, C. (2001). Different Methods of Evaluating Student Translations: the Question of Validity. *Meta, 46*(2), 312-325.

Websoft. CourseLab 2.7 User Manual.

Wehrli, E., & Nerima, L. (2013). *Anaphora Resolution, Collocations and Translation.* Paper presented at the MT Summit XIV Workshop on Multi-word Units in Machine Translation and Translation Technology (MUMTTT 2013), Nice, France.

Weiss, S., & Ahrenberg, L. (2012). *Error profilingfor evaluation of machine-translated text: a Polish-English case study.* Paper presented at the 8th International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey.

White, J., O'Connell, T., & O'Mara, F. (1994). *The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Further Approaches.* Paper presented at the Conference of the Association for Machine Translation in the Americas (AMTA 1994).

Williams, M. (2009). Translation quality assessment. *Mutatis Mutandis, 2*(1), 3-23.

Wisniewski, G., Singh, A. K., & Yvon, F. (2013). Quality estimation for machine translation: some lessons learned. *Machine Translation, 27*(3-4), 213-238.

Yamada, M. (2015). Can college students be post-editors? An investigation into employing language learners in machine translation plus post-editing settings. *Machine Translation, 29*, 49-67.

Young, M., & Stanton, N. (2007). What's skill got to do with it? Vehicle automation and driver mental workload. *Ergonomics, 50*(8), 1324-1339.

Zhechev, V. (2014). Analysing the Post-Editing of Machine Translation at Autodesk. In S. O'Brien, L. W. Balling, M. Carl, M. Simard, & L. Specia (Eds.), *Post-editing of Machine Translation: Processes and Applications* (pp. 2-23). Newcastle upon Tyne, UK: Cambridge Scholars Publishing.

Appendix

Appendix 1: Texts and MT output pretests

Pretest 1: newspaper articles

Source text 1: New diseases arise as environments destroyed, says UN

Changes to the environment that are sweeping the planet are bringing about a rise in infectious diseases, the United Nations Environment Programme (Unep) has warned. Loss of forests; the building of roads and dams; urban growth; the clearing of natural habitats for agriculture; mining; and pollution of coastal waters are promoting conditions under which new and old pathogens can thrive, according to research published today in Unep's Global Environment Outlook Year Book for 2004/2005. Ailments previously unknown in human beings are appearing, such as the Nipah virus, which until recently was found normally in Asian fruit bats, according to the report. Nipah's emergence in the late 1990s as an often fatal disease in humans has been linked to a combination of forest fires in Sumatra and the clearance of natural forests in Malaysia for palm plantations. In searching for fruit, bats were forced into closer contact with domestic pigs, giving the virus its chance to spread to humans. Climate change in particular may aggravate the threats of infectious diseases in three ways, the report suggests. First, by increasing the temperatures under which many diseases and their carriers flourish. Second, by further stressing and altering habitats. Third, climate change may increase the number of environmental refugees who are forced to migrate to other communities, or even countries. This in turn will also favour the spread of diseases from one location to another. Overall, it seems that intact habitats and landscapes tend to keep infectious agents in check.

The issue of environmental degradation and a rise of many new and old infectious diseases is a complex, sometimes subtle one that is causing increasing concern among scientists and disease specialists.

MT output text 1: Nieuwe ziekten ontstaan als omgeving vernietigd, zegt VN

Veranderingen in de omgeving die het vegen van de planeet tot stand brengen van een stijging van de besmettelijke ziekten, heeft het United Nations Environment Programme (UNEP) gewaarschuwd. Verlies van bossen, de aanleg van wegen en dammen, stedelijke groei, de clearing van de natuurlijke habitats voor de landbouw, mijnbouw, en

vervuiling van de kustwateren zijn het bevorderen van de voorwaarden waaronder nieuwe en oude ziekteverwekkers kunnen gedijen, blijkt uit een onderzoek dat vandaag wordt gepubliceerd in het Global Environment UNEP Outlook Year Book voor 2004/2005. Kwalen voorheen onbekende bij de mens verschijnen, zoals de Nipah-virus, dat tot voor kort normaal werd gevonden in de Aziatische vliegende honden, volgens het rapport. Nipah's opkomst in de late jaren 1990 als een vaak dodelijke ziekte bij de mens is gekoppeld aan een combinatie van bosbranden op Sumatra en de inklaring van natuurlijke bossen in Maleisië voor palm plantages. Bij het zoeken naar fruit, werden vleermuizen gedwongen tot nauwer contact met tamme varkens, waardoor het virus de kans om zich te verspreiden naar de mens. Klimaatverandering in het bijzonder kan verergeren de bedreigingen van infectieziekten op drie manieren, stelt het rapport. Ten eerste, door het verhogen van de temperatuur waaronder een groot aantal ziekten en hun dragers gedijen. Ten tweede, door het verder benadrukken en de wijziging van habitats. Ten derde, de klimaatverandering kan verhogen het aantal milieu-vluchtelingen die worden gedwongen om te migreren naar andere gemeenschappen, of zelfs landen. Dit zal op zijn beurt de verspreiding van ziekten ook voorstander van de ene locatie naar de andere. Over het geheel genomen lijkt het erop dat intact habitats en landschappen hebben de neiging om infectieuze agentia in toom te houden. De kwestie van aantasting van het milieu en een stijging van vele nieuwe en oude besmettelijke ziekten is een complexe, soms subtiel een die steeds meer zorgen baart tussen wetenschappers en ziekte specialisten.

Source text 2: US chat show host who sent 'coded messages' has restraining order lifted

David Letterman, the doyen of American late-night chat show hosts, has had his share of fans with unhealthy fixations, but this one probably beats them all: a New Mexico woman who claims he has been sending her secret coded messages over the airwaves so incessantly that it constitutes "mental harassment and hammering". Colleen Nestler, of Santa Fe, successfully applied for a restraining order two weeks ago forcing Mr Letterman to stay at least 100 yards from her. She alleged his subliminal messages - including, supposedly, an entreaty to marry him and become his co-host - had caused her sleep deprivation, pushed her into bankruptcy and inflicted general "mental cruelty". Since Mr Letterman lives in Connecticut, about 2,000 miles from Santa Fe, the restraining order was not exactly a crimp on his day-to-day existence. But it did offend his sense of judicial fairness, so he sent his lawyers to the New Mexico courts this week to have it lifted. The judge granted the request, noting the original restraining order was granted merely as a matter of "proper pleading" - a legal term meaning the paperwork was filled out correctly, no more and no less. She said she had begun sending Mr Letterman love messages in 1993 and that he had responded with a suggestion that she move to the East Coast. His marriage proposal supposedly came in a teaser for his

show in which he said, jokingly, "Marry me, Oprah". According to Ms Nestler, Oprah was the first of many codenames he used for her. The code later became more sophisticated and complex.

MT output text 2: Amerikaanse talkshow gastheer die verzonden 'gecodeerde berichten' heeft straatverbod opgeheven

David Letterman, de nestor van de Amerikaanse late-night talkshow hosts, heeft zijn aandeel van fans met ongezonde fixaties, maar deze waarschijnlijk verslaat ze allemaal: een New Mexico vrouw die beweert dat hij is haar geheime gecodeerde berichten versturen via de ether, zodat onophoudelijk dat het gaat het om "geestelijke intimidatie en hameren". Colleen Nestler, van Santa Fe, met succes toegepast voor een straatverbod twee weken geleden dwingen de heer Letterman ten minste 100 meter te blijven van haar. Zij beweerde zijn subliminale boodschappen - met inbegrip van, vermoedelijk, een smeekbede om met hem te trouwen en wordt zijn co-host - had veroorzaakt haar slaaptekort, duwde haar in faillissement en bracht het algemeen "geestelijke wreedheid". Aangezien de heer Letterman woont in Connecticut, ongeveer 2.000 mijl van Santa Fe, het straatverbod was niet bepaald een krimp op zijn dag-tot-dag leven. Maar het deed beledigen zijn gevoel van gerechtelijke eerlijkheid, zodat hij zijn advocaten naar het New Mexico rechtbank deze week te laten opgeheven. De rechter heeft het verzoek ingewilligd, wijzend op de oorspronkelijke straatverbod werd alleen verleend als een kwestie van 'goede pleidooi "- een juridische term betekent het papierwerk werd correct ingevuld, niet meer en niet minder. Ze zei dat ze was begonnen het verzenden van de heer Letterman liefde berichten in 1993 en dat hij antwoordde met een suggestie dat ze te verplaatsen naar de Oostkust. Zijn huwelijk voorstel vermoedelijk kwam in een teaser voor zijn show waarin hij zei, gekscherend, 'met mij te trouwen, Oprah ". Volgens mevrouw Nestler, Oprah was de eerste van vele codenamen die hij gebruikte voor haar. De code werd later meer geavanceerde en complexe.

Source text 3: 'Miracle' cures shown to work

Doctors have found statistical evidence that alternative treatments such as special diets, herbal potions and faith healing can cure apparently terminal illness, but they remain unsure about the reasons. A study of patients with incurable lung cancer who were given weeks to live and received only low-dose radiotherapy to make their final weeks more comfortable found a small number recovered completely. Researchers who followed 2,337 patients whose disease was too advanced for curative treatment found that 25 had survived five years and 18 had achieved "an apparent cure". They appeared to have been cured by treatment that "would not normally be considered to have any curative potential whatsoever". The researchers, led by Michael MacManus, a

consultant radiation oncologist in Melbourne, say: "Our data indicate that a chance for prolonged survival and possibly even cure exists for approximately 1 per cent of patients with non small cell lung cancer who receive palliative radiotherapy. "It is important that the frequency of this phenomenon should be appreciated so that claims of apparent cure by novel treatment strategies or even by unconventional medicine or 'faith healing' can be seen in an appropriate context." Unorthodox cancer cures have included vitamin C, laetrile extracted from apricot stones, and the Gershon diet of raw vegetables. The discovery of a small group of patients who unexpectedly recovered could yield new insights into the disease, the researchers say. The findings are published in the online edition of Cancer, the journal of the American Cancer Society.

MT output text 3: 'Miracle' kuren blijkt te werken

Artsen hebben gevonden statistisch bewijs dat alternatieve behandelingen zoals speciale diëten, kruiden drankjes en gebedsgenezing kan blijkbaar terminale ziekte te genezen, maar ze blijven onzeker over de redenen. Een studie van patiënten met ongeneeslijke longkanker die werden in weken te leven en ontving slechts een lage dosis radiotherapie om hun laatste weken meer comfortabele vond een klein aantal herstelde volledig. Onderzoekers die volgden 2.337 patiënten bij wie de ziekte was te vooruitstrevend voor curatieve behandeling gevonden dat 25 had overleefd vijf jaar en 18 had bereikt "een schijnbaar genezen". Ze leek te zijn genezen door behandeling die "normaal gesproken niet worden beschouwd als een curatief potentieel dan ook hebben". De onderzoekers, onder leiding van Michael MacManus, een consultant radiotherapeut in Melbourne, zeggen: "Onze gegevens tonen aan dat er een kans op langdurige overleving en mogelijk zelfs te genezen bestaat voor ongeveer 1 procent van de patiënten met niet-kleincellige longkanker die ontvangen palliatieve radiotherapie.
"Het is belangrijk dat de frequentie van dit verschijnsel zal duidelijk zodat vorderingen van schijnbare genezing van nieuwe therapeutische strategieën of zelfs onconventionele medicijnen of 'gebedsgenezing' te zien in een geschikte context." Onorthodoxe kanker geneest hebben opgenomen vitamine C, laetrile gewonnen uit abrikozenpitten, en de Gershon dieet van rauwe groenten. De ontdekking van een kleine groep patiënten die onverwacht hersteld kunnen nieuwe inzichten opleveren in de ziekte, zeggen de onderzoekers. De bevindingen zijn gepubliceerd in de online editie van kanker, het tijdschrift van de American Cancer Society.

Source text 4: Gibson is accused of anti-Semitic rant after failing drink-drive test

Los Angeles police and prosecutors are examining allegations that the actor-director Mel Gibson made abusive anti-Semitic remarks when he was arrested on drink-driving

charges near his beach-side home in Malibu in the early hours of Friday morning. According to a published reproduction of the arresting officer's handwritten report, Gibson, 50, became "belligerent" after failing a blood-alcohol test and blurted out "a barrage of anti-Semitic remarks". The Los Angeles County Sheriff's Department said yesterday it was not able to confirm or deny the authenticity of the report, four pages of which appeared on a celebrity website over the weekend, because it was subject to an internal investigation and was also being scrutinised by the district attorney's office. Spokesman Steve Whitmore did confirm the identity of the arresting officer, James Mee. Saturday's Los Angeles Times, meanwhile, cited "a source close to the investigation" as confirming that the published pages were authentic. In Mr Mee's published account, Gibson yelled: "The Jews are responsible for all the wars in the world" and turned around to ask the officer, "Are you a Jew?" Gibson issued a statement blaming the incident on a "horrific relapse" into the alcoholism that has plagued his adult life. "I acted like a person completely out of control when I was arrested and said things that I do not believe to be true and which are despicable," he said. The posted pages said Gibson became angry when he realised how much trouble he was in, and how much publicity his arrest was likely to generate.

MT output text 4: Gibson wordt beschuldigd van antisemitische tirade na niet rijden onder invloed-test

Politie van Los Angeles en openbare aanklagers onderzoeken beschuldigingen dat de acteur-regisseur Mel Gibson misbruik antisemitische opmerkingen maakte toen hij werd gearresteerd op rijden onder invloed lasten de buurt van zijn strand-side huis in Malibu in de vroege uren van vrijdag ochtend. Volgens een gepubliceerde reproductie van handgeschreven de arresterende agent verslag, Gibson, 50, werd "oorlogszuchtige" na het uitblijven van een bloed-alcohol-test en flapte eruit: "een spervuur van antisemitische opmerkingen" De Los Angeles County Sheriff's Department zei gisteren dat het niet kunnen bevestigen of ontkennen van de authenticiteit van het rapport, vier pagina's van die verscheen op een beroemdheid website in het weekend, want het was onderworpen aan een intern onderzoek en werd ook onder de loep genomen door de officier van justitie het kantoor. Woordvoerder Steve Whitmore deed bevestiging van de identiteit van de arresterende agent, James Mee. Zaterdag Los Angeles Times, ondertussen, reeds "een bron dicht bij het onderzoek" als bevestiging dat de de gepubliceerde pagina's authentiek waren. In gepubliceerde verslag heer Mee's, Gibson schreeuwde: "De Joden zijn verantwoordelijk voor alle oorlogen in de wereld" en draaide zich om naar de officier vragen: "Bent u een Jood? ' Gibson een verklaring de schuld te geven van het incident op een "verschrikkelijke terugval" in het alcoholisme, dat heeft geplaagd zijn volwassen leven. "Ik handelde als een persoon volledig uit de hand toen ik gearresteerd werd en zei dingen die ik niet geloof om waar te zijn en die

verachtelijk," zei hij. De gedetacheerde pagina's zei Gibson boos werd toen hij zich realiseerde hoeveel moeite hij in, en hoeveel publiciteit zijn arrestatie was vermoedelijk zal opleveren.

Pretest 2: technical texts

Source text 1: Object Selection

In order to execute command against the object, this object has to be selected. Selected objects are outlined by markers, which are also used for object's resizing. The processing information regarding selected object is displayed in the status field within CourseLab. In case multiple objects are selected, status field contains only information pertaining to the last selected object.

Selecting objects in the workspace

- To select object, click on it using left mouse button.
- In order to select multiple objects, left click on the desired object while holding down Shift or Ctrl key.
- To undo selection selectively, left click on the object while holding down Shift or Ctrl key.
- Click anywhere within a Slide to undo selection of all objects.
- Use **Ctrl+A** combination to select all object in the Slide.

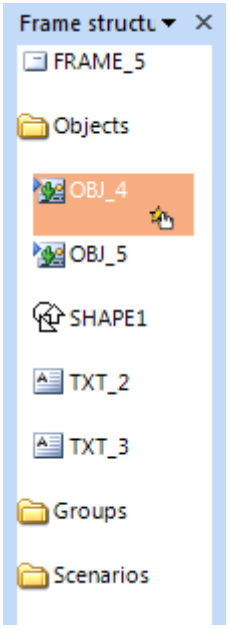
Regardless the fact that objects placed on the Master-Slide are visible on the standard Slides as well, to select such objects they need to be opened within Master-Slide.

Sequential objects Selection within workspace

When object is selected, you can also select the subsequent object by pressing Tab key.

- In order to select previous object from the sequence, use Tab key while holding down Shift key.

Object's selection in the task panel



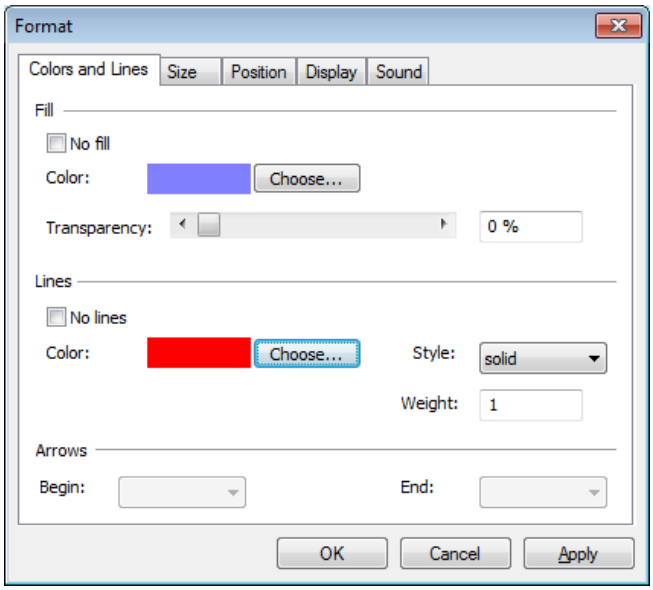
1. Use **View** -> **Tasks Pane** -> **Frame Structure** to open Frame Structure section within Tasks Panel.

2. To select the object in the Frame Structure, left click on the object's identifier. To select all the objects within the group in the Frame Structure, left click on the group's identifier.

Border and fill colors. Opacity

When you insert any Object into a Frame, it is placed into the rectangular placeholder. Use "Color" tab on the "Format" screen to specify placeholder's background color. Note that these parameters pertain only to the Object's placeholder, rather than the Object itself. If the Objects fit the entire rectangular placeholder (for example pictures), it is appropriate to specify the placeholder's background color only when the pictures contain transparent areas.

Exceptions: AutoShapes Objects, where changes to placeholder's background and border color apply to the AutoShapes directly.



Google Translate output text 1: Object Selectie

Om commando tegen het object dit object worden geselecteerd. Geselecteerde objecten beschreven door markers die ook worden gebruikt voor het schalen object. De verwerking informatie over geselecteerde object wordt weergegeven in het status veld binnen CourseLab. In het geval dat er meerdere objecten geselecteerd, status veld bevat alleen informatie met betrekking tot het laatst geselecteerde object.

Objecten selecteren in de werkruimte
• Om objecten te selecteren, klikt u erop met de linker muisknop.
• Om meerdere objecten te selecteren, klik links op de gewenste ingedrukt te houden terwijl Shift-of Ctrl-toets.
• Om selectief ongedaan maken selectie, de linker muisknop op het object met ingedrukte Shift-of Ctrl-toets.
• Klik op een willekeurige binnen een dia om de selectie van alle objecten ongedaan te maken.
• Gebruik Ctrl + Een combinatie van alle object in de dia te selecteren.

Ongeacht het feit dat de objecten op de Master-Slide zijn zichtbaar op de standaard dia's ook, om dergelijke objecten moeten ze worden geopend binnen de Master-Slide selecteren.

Sequentiële objecten Selectie binnen de werkruimte
Als object is geselecteerd, kunt u ook kiezen voor de daaropvolgende object door op Tab-toets.
• Om naar de vorige object te selecteren uit de reeks, gebruik Tab-toets terwijl u de Shift-toets te drukken.

Object selectie in het taakdeelvenster
1. Gebruik View -> Taakvenster -> Frame Structuur naar Frame Structuur sectie te openen in het deelvenster Taken.
2. Om het object in het frame structuur te selecteren, klik links op identificatie van het object. Als u alle objecten in de groep in het frame structuur te selecteren, klik links op identificatie van de groep.

Grens-en vulkleur. Dekking
Als u een object invoegt in een frame, wordt deze geplaatst in de rechthoekige tijdelijke aanduiding. Gebruik "Kleur" tab op de "Formaat" scherm om placeholder de achtergrondkleur te geven. Merk op dat deze parameters alleen betrekking hebben op

tijdelijke aanduiding van het object in plaats, dan het object zelf. Als de objecten op het volledige rechthoekige tijdelijke aanduiding (bijvoorbeeld foto's), is het wenselijk om de tijdelijke aanduiding van de achtergrond kleur op te geven wanneer de foto's transparante gebieden bevatten.
Uitzonderingen: AutoVormen Objects, waarin wijzigingen aan achtergrond tijdelijke aanduiding en randkleur rechtstreeks van toepassing op de AutoVormen.

Bing Translator output text 1: Een selectie van objecten

Dit object heeft om uit te voeren opdracht tegen het object, worden geselecteerd. Geselecteerde objecten worden beschreven bij de markeerders, die ook worden gebruikt voor het wijzigen van het formaat van het object. Het verwerken van informatie met betrekking tot geselecteerde object wordt weergegeven in het statusveld binnen de CourseLab. Indien meerdere objecten zijn geselecteerd, bevat het statusveld alleen informatie met betrekking tot de laatste object hebt geselecteerd.

Selecteren van objecten in de werkruimte

• Selecteer object, klikt u op op met behulp van de linker mouse button.

• Om meerdere objecten selecteren Klik op het gewenste object terwijl u Shift of Ctrl toets ingedrukt.

• Om selectief, selectie ongedaan te maken Klik op het object terwijl u Shift of Ctrl toets ingedrukt.

• Klik ergens in een dia om selectie van alle objecten ongedaan te maken.

• Gebruik Ctrl + A combinatie alle object te selecteren in de dia..

Ongeacht het feit dat objecten op de Master-dia geplaatst om deze objecten te selecteren zijn zichtbaar op de standaard dia's, moeten ze worden geopend binnen Master-dia.

Sequentiële objecten selectie in de werkruimte

Wanneer een object is geselecteerd, kunt u het volgende object ook selecteren door op Tab-toets te drukken.

• Om het vorige object selecteren in de reeks, Tab-toets gebruiken terwijl u de Shift-toets ingedrukt houdt.

De selectie van het object in het deelvenster Taken

1. Weergave-> taakvenster-> framestructuur gebruiken om open framestructuur sectie in het deelvenster Taken.

2. Schakel het object in de structuur van het Frame, klik links op de object-id. Schakel alle objecten in de groep in de structuur van het Frame, klik links op de groeps-id.

Kleuren van rand en een opvulling. Dekking

Wanneer u een Object in een Frame invoegt, wordt het naar de rechthoekige tijdelijke aanduiding geplaatst. Met "Kleur" tabblad op het scherm 'Format' geeft u de tijdelijke aanduiding voor de achtergrondkleur. Merk op dat deze parameters betrekking alleen op tijdelijke aanduiding van het Object, in plaats van het Object zelf hebben. Als de objecten past de hele rechthoekige tijdelijke aanduiding (bijvoorbeeld afbeeldingen), is het dienstig is te bepalen van de voorlopige afbeelding achtergrondkleur alleen wanneer de foto's transparante gebieden bevatten.

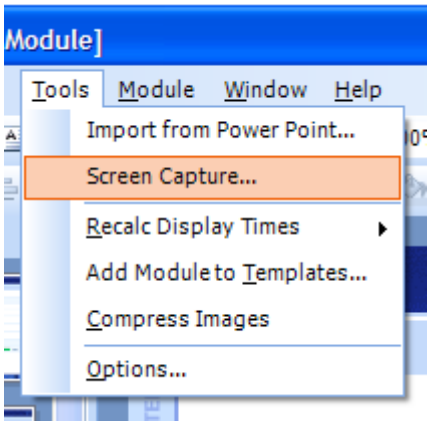
Uitzonderingen: AutoVormen objecten, waar wijzigingen in de tijdelijke aanduiding voor de achtergrond en rand kleur op de AutoVormen rechtstreeks toegepast.

Source text 2: Screen Capture

Learning courses are created for many purposes. One of the most common objectives is instructing on how to use various software. To facilitate the creation of software simulations CourseLab contains built-in screen capture mechanism, therefore no additional software needs to be installed. Simulations are recorded directly into the internal format of the editor and can be edited later as usual frames. Internet Browser's capabilities allow replaying of such animated simulations. No additional components (Flash Player, Shockwave Player, Media Player, etc.) are required.

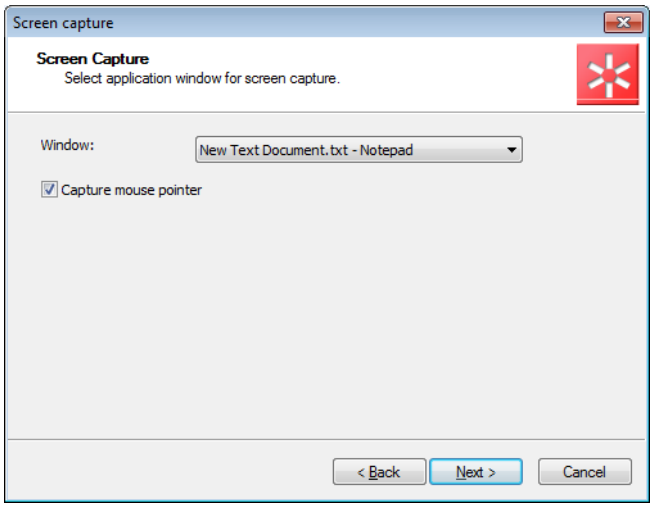
Screen Capture Wizard

While on the slide, which is to be used for recording the simulation, select "Capture Screens" item from the "Tools" menu.

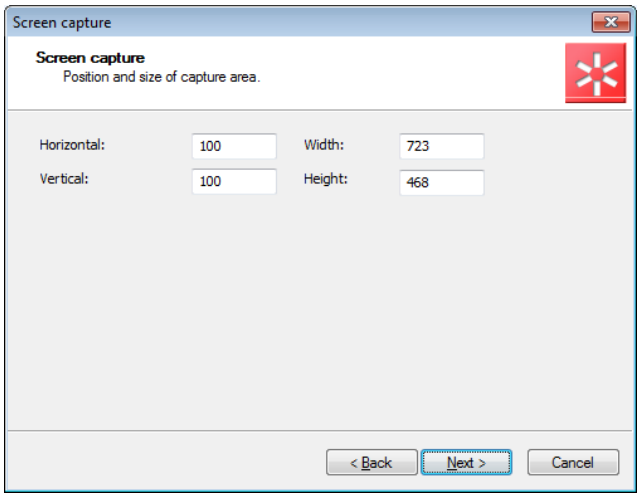


Screen Capture Wizard opens up.

From the drop-down menu select program to record the simulation from. Mark "Capture Cursor" check box if you need to record mouse's clicks and movements. Clicks and movements will be captured automatically, once the recording starts.



Specify position for the top left corner of the area for recording simulations within a frame. By default, position of top left corner of the area for recording simulations within a frame is equal to top left corner of the frame (position 0,0), however there are instances where it is not acceptable. For example, if there is a title located at the top of the frame, then top left corner of the area for recording should be placed underneath, by adding frame title's height in pixels into the "Vertical Position" field.

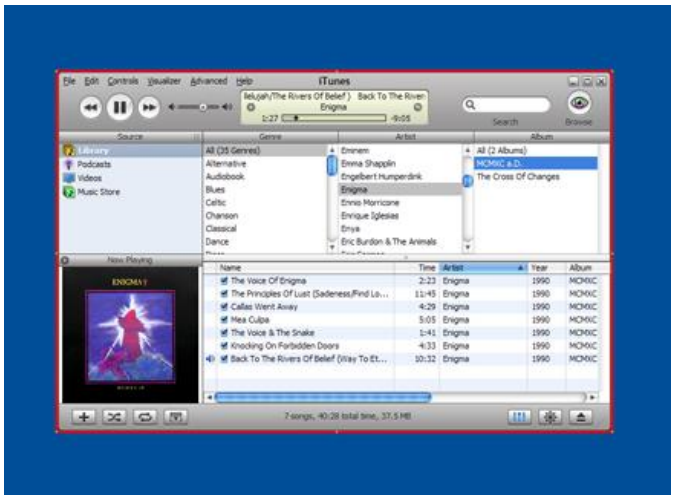


Define location and size of the area on the monitor to be captured. The editor will try to adjust captured application window to the specified size automatically, if possible.

Capture Frames

After switching to the capture mode, editor minimizes into the icon on the windows taskbar.

Red rectangular enclose appears, which limits the area or capturing. Make sure that all desired information is well fitted inside of the screen capture area.



Press **PrintScrn**, the first frame is captured. Go to the next step in target program, and press **PrintScrn** one more time. Using **PrintScrn** continue capturing all the changes that you apply to the target program, until you complete recording based on the desired scenario.

Upon capture scenario is completed, double click CourseLab icon from the taskbar. CourseLab Editor window will be restored and all captured frames will be available for editing.

Google Translate output text 2: Screen Capture

Learning cursussen zijn gemaakt voor vele doeleinden. Een van de meest voorkomende doelstellingen instrueert over diverse software. Ter vergemakkelijking van het maken van software simulaties CourseLab bevat ingebouwde screen capture-mechanisme, dus geen extra software te worden geïnstalleerd. Simulaties worden direct opgenomen in de interne indeling van de editor en kan later als gebruikelijk frames worden bewerkt. Internet Browser mogelijkheden laten afspelen van dergelijke geanimeerde simulaties. Geen extra componenten (Flash Player, Shockwave Player, Media Player, enz.) zijn verplicht.

Screen Capture Wizard

Terwijl op de dia, die moet worden gebruikt voor het opnemen van de simulatie, selecteert u "Capture Schermen" item uit de menu "Extra".

Screen Capture Wizard opent.

Vanaf de drop-down menu te selecteren programma om de simulatie opnemen. Mark "Capture Cursor" in als u nodig hebt om muis klikken en bewegingen op te nemen. Klikken en bewegingen wordt automatisch vastgelegd, zodra de opname start.

Geef positie voor de linker bovenhoek van het gebied voor het opnemen van simulaties binnen een frame. Standaard positie van linker bovenhoek van het gebied voor het opnemen van simulaties binnen een frame is gelijk aan linker bovenhoek van het frame (positie 0,0), maar er zijn gevallen waar het niet aanvaardbaar is. Bijvoorbeeld, als er een titel aan de bovenkant van het frame, dan moeten linkerbovenhoek van het gebied voor opname onder worden gebracht door toevoeging lijst titel van de lengte in pixels in de "verticale positie" veld.

Definieer locatie en de grootte van het gebied op de monitor vast te leggen. De editor zal proberen om gevangen toepassingsvenster zich automatisch aanpassen aan het opgegeven formaat, indien mogelijk.

Capture Frames

Na het inschakelen naar de opnamestand, redacteur minimaliseert in het pictogram op de Windows-taakbalk.

Rode rechthoekige omsluiten verschijnt, welke grenzen het gebied of het vastleggen. Zorg ervoor dat alle gewenste informatie en wordt aangebracht aan de binnenzijde van de screen capture gebied.

Druk op PrintScrn, het eerste frame wordt vastgelegd. Naar de volgende stap in doelprogramma en druk PrintScrn een keer. Met behulp van PrintScrn blijven vastleggen van alle wijzigingen die u op de doel-programma, totdat u de opname te voltooien op basis van het gewenste scenario.

Na capture scenario is voltooid, dubbelklikt u op CourseLab pictogram van de taakbalk. CourseLab Editor venster zal worden hersteld en alle gemaakte opnamen beschikbaar zal zijn voor het bewerken.

Bing Translator output text 2: Zeef inneming

Trainingen worden gemaakt voor vele doeleinden. Een van de meest voorkomende doelstellingen is instrueren over het gebruik van verschillende software. Bevat ingebouwde scherm vangen mechanisme ter vergemakkelijking van de oprichting van software simulaties CourseLab, dus geen extra software moet worden geïnstalleerd. Simulaties rechtstreeks naar de interne indeling van de editor zijn opgenomen en later

als gebruikelijke frames kunnen worden bewerkt. Internet Browser analysemogelijkheden herhalen van dergelijke geanimeerde simulaties. Geen extra onderdelen (Flash Player, Shockwave Player, Media Player, enz.) zijn vereist.

Screen Capture Wizard

Terwijl op de dia die moet worden gebruikt voor het opnemen van de simulatie, selecteer "Vangen schermen" item in het menu 'Extra'.

Screen Capture Wizard opent.

Selecteer vastleggen van de simulatie van het drop-down menu. Mark "Vangen Cursor" selectievakje in als u wilt opnemen van muis klikken en bewegingen. Klikken en bewegingen zal worden automatisch vastgelegd, zodra de opname begint.

Positie van de linkerbovenhoek van het gebied voor het opnemen van simulaties binnen een frame opgeven. Standaard, positie van de linkerbovenhoek van het gebied voor het opnemen van simulaties binnen een frame is gelijk aan de linker bovenhoek van het frame (positie 0,0), maar er zijn gevallen waarin het is niet aanvaardbaar. Bijvoorbeeld, als er een titel gelegen op de top van het frame, dan boven linker hoek van het gebied voor opname moet worden geplaatst onder, door toe te voegen frame titel hoogte in pixels in het veld "Verticale positie".

Definieer de locatie en grootte van het gebied op de monitor worden vastgelegd. De editor zal proberen aan te passen indien mogelijk automatisch, opgenomen toepassingsvenster tot de opgegeven grootte.

Frames opnemen

Nadat u bent overgeschakeld naar de opnamemodus, minimaliseert editor in het pictogram op de taakbalk van windows.

Red rechthoekige omsluiten wordt weergegeven, die het gebied beperkt of vastleggen. Zorg ervoor dat alle gewenste informatie is goed uitgerust binnenkant van het scherm vangen gebied.

Druk op PrintScrn, het eerste frame wordt vastgelegd. Ga naar de volgende stap in het doelprogramma, en druk op PrintScrn nog een keer. Met behulp van PrintScrn blijven alle wijzigingen die u op het doelprogramma, toepast zolang u niet hebt voltooid opname op basis van de gewenste scenario vastleggen.

Op vangst scenario is voltooid, dubbelklik CourseLab pictogram in de taakbalk. CourseLab Editor venster zal worden hersteld en alle opgenomen frames zullen beschikbaar zijn voor bewerking.

Appendix 2: Annotation Guidelines for English-Dutch Translation Quality Assessment

Introduction

Assessing translation quality is a very complex task, and depending on the goal of the assessment, a different approach is needed. Though existing translation quality assessment (TQA) metrics are suitable for a general assessment or assessment within a specific company, they lack the granularity needed to compare different methods of translation and their respective translation problems. The translation situation is not always taken into account, the assessments are often subjective and they are often limited to the sentence level. Moreover, the annotation task is often too complex, as annotators have to identify the location of errors, the main category an error belongs to and they have to give a weight to each instance of an error. In an attempt to solve these and other problems, we propose a two-step TQA approach with a fine-grained categorisation of translation problems and user-defined error weights.

Categorisation

The categories are divided into two main groups: adequacy and acceptability. Though some subcategories are suggested in this report, categories can easily be added or deleted to better suit certain assessment and language needs. Categories contain *possible* translation problems, but depending on the text type they will be considered to be errors or not.

Error weights

Rather than letting annotators judge a text as containing 'minor', 'major', or 'critical' errors, error weights are defined by the user. As such, the error weights can be adapted to the translation situation: for technical texts, for example, 'terminology' will receive a high error weight. The error weights can also be equal to zero, which can be useful for researchers or teachers interested in studying the translation characteristics, rather than errors. Hyperonymy, for example, is not always an error, but it can be interested to highlight cases of hyperonymy in order to identify differences between translators or translation methods (human translation vs. post-editing of machine translation).

Our suggested error weight set-up allows for five severity levels:

- 0: not an actual problem: the category deals with an aspect of translation that is not considered to be an error for the task at hand (for example: explicitation; terminology in general texts,...) or the category deals with an error that was caused by the translation situation (for example: the tool did not contain a spell-checker, so typos were not automatically filtered out. In a 'real' translation situation, this error would not have been made).
- 1: minor problems: the text can still be understood without effort and the information contained in the translation is equal to that of the source text, but there is a small error or the readability is affected a little (for example: capitalization errors; long or short sentences...)
- 2: medium problems: there is a slight shift in meaning between source text and target text, or the text can be understood with little effort, but it is not entirely correct regarding either grammar, lexicon, style or coherence (for example: wrong prepositions).
- 3: major problems: there is some misunderstanding of the source text, or the readability and/or understandability of the text is affected by incorrect grammatical structures or awkward expressions in the target text (for example, wrong collocations.
- 4: critical problems: errors within this category have a critical impact on the understandability and/or accuracy of a text (for example: the content doesn't make sense or there is a contradiction between source and target text), or when explicit translation instructions have been ignored (for example: terminology has not been translated according to a terminology list).

Two-step

One of the reasons error annotation is such a difficult process, is the fact that annotators have to decide whether an error belongs to 'adequacy' or 'acceptability'. Sometimes, an error can affect both. To facilitate the annotation task for the annotators, and to allow for a clear view of these two different perspectives (the source text perspective and the target text perspective), we split up the annotation process into two different steps. In a first step, annotators only receive the target text and they have to annotate the text for acceptability. By only giving them the target text, they cannot be influenced by the source text in their acceptability assessment and they can also judge the general coherence of the text.

In the second step, annotators see the source sentences next to the target text and they have to annotate the texts for adequacy.

For optimal results, we recommend informing the annotators as thoroughly as possible on the goal of the translation and the assessment, so they know which problems to highlight. Provide them with the same information the translators had during the task (Did they receive certain pictures or background information? Was there terminology involved?). If terminology is crucial, also provide the annotators with the glossary the translators were asked to adhere to, so they can judge the translation for terminological adequacy.

The technical report 2.0

This technical report contains a possible classification of translation problems for the translation of texts from English to Dutch and guidelines on how to annotate these problems. Though tuned to suit the needs of the Dutch and English language, the categorisation allows for customization to suit different language-pair needs. The categorisation was tested during two pilot studies with different text types (newspaper articles and technical texts), and has been used to assess and compare the quality of human translations, machine translations and post-edited texts.

These guidelines detail the annotation process with the brat rapid annotation tool. The tool provides an intuitive interface and it is possible to add your own categories to the tool. Recent additions to the brat tool have allowed us to improve the annotation process and facilitate the analysis afterwards. The technical report has been updated to accommodate these features.

Using the brat-tool

Hover over the word 'brat' at the top right hand corner to be able to select 'log in' and use your username and password to log in.

You are now ready to start annotating. Just double-click a word or click and drag to select smaller/larger pieces of text and the tool will give you an overview of the possible categories.

The categories are listed below 'entity type'. After selecting the appropriate category, select the correct subcategory from the drop-down menu below 'entity attributes'. Make sure to always select a subcategory, not just the main category!

Always add extra information in the 'notes' section to explain the reason of the annotation and the selected category (this facilitates the analysis afterwards and helps consolidate annotations between different annotators).

Just click 'ok' when you're done or 'cancel' when you've selected a piece of text that you didn't want to select. To change an annotation, double-click the label above the word. You can change the category, subcategory and notes, you can decide to delete the annotation or you can move the annotation. To move an annotation, first select 'move' and then select the text span where you want the annotation to move to.

! Be careful when changing the category of an annotation: sometimes the tool remembers the first chosen subcategory alongside the new subcategory (even when the first subcategory belongs to a different main category than the second). If this happens, simply delete your annotation and make a new one with the correct subcategory.

You can select a word or span more than once, so it is possible to assign different problem categories to the same word.

Annotations on sentence or word level

We advise adding a text code above each text to be annotated, as well as sentence codes at the beginning of each sentence. Sometimes, you'll have to make an annotation that concerns the entire text or an entire sentence instead of a single word or a span of text. Rather than selecting the entire sentence in this case, the text or sentence code can be selected to indicate that the annotation affects the entire sentence/text.

Linking spans

When you encounter a problem that concerns more than one word, but these words are split up by other words, you'll have to make sure that those non-relevant words are not contained in your span. You do this by selecting the two parts of the sentence that contain the problem separately (so you make two different annotations of the same category) and then you link them together with an arrow. You do this by clicking the first annotation and dragging your mouse pointer to the second annotation. You'll see an arrow appear with the words 'belongs_to'. The guidelines contain information on when you are allowed or required to insert a link between spans.

A second type of linking spans (caused_by) is also possible between all acceptability annotations and adequacy annotations. The use of this relationship shall be explained in more detail in the consolidation section.

General annotation rules

1) Annotations are made on the Dutch target text, unless otherwise specified.
2) Always select a subcategory, not just a main category.
3) Always specify why you made a certain annotation in the 'notes' section. You are allowed to use either English or Dutch for your comments.
4) If an item contains more than one problem, highlight the item as many times as there are problems, once for each problem (for example, a word that contains both a compound error and a capitalization error).
5) Only select 'other' if there is no other category that describes the problem better.

6) If you are not sure about your annotation, end your comment in the notes section with a double question mark '??' to indicate this insecurity.

7) If the same problem occurs more than once, select each occurrence separately.

8) Only problems that affect the entire sentence or problems that cannot be highlighted in the sentence itself are indicated by highlighting the sentence code. Other annotations are made within the sentence itself.

When in doubt...

If you are not sure whether or not something is an error, don't be afraid to consult external sources. You are perfectly allowed to use a dictionary or a search engine to look things up. Some useful sites that you could consult for acceptability issues are:

http://www.vandale.be/

http://taaladvies.net/

http://woordenlijst.org

http://www.vrt.be/taal/

You can refer to external sources in the 'notes' section to support your decisions. It is also allowed to look back to previous texts, to check how you annotated the same problem in a different translation.

Step 1: Annotating acceptability

The texts that you are about to annotate are the results of a translation task from English to Dutch. To be able to judge the quality of the translations, they will be marked for two important aspects: adequacy and acceptability. Adequacy is concerned with the relationship between source text and target text, whereas acceptability is concerned with the target text and language. The goal of the current assignment is to annotate translations for acceptability. Adequacy will be dealt with separately. In order to allow you to focus on the requirements of the target text and language, without being distracted by the source text, you only receive the Dutch translation, without the English reference text.

Acceptability can be described as respecting the norms of the target language and culture. A good translation should read as a native Dutch text. This includes respecting the conventions of the language (grammar, lexicon, spelling) as well as respecting the conventions of the text structure (paragraph content and coherence) and the text type (a newspaper article requires a different style than a manual, for example).

As a reviser, it is your task to make sure the translation reaches publishable and acceptable quality. You do this by marking anything that does not follow the conventions of the Dutch language or that does not respect the demands of the text structure or the text type. To facilitate the task, you can use the brat rapid annotation tool. In this tool, the different categories (grammar & syntax; lexicon; spelling, typos & punctuation; style & register; coherence) are predefined, along with their most important subcategories.

Following is an overview of all the (sub)categories for acceptability and guidelines on how to annotate these issues within the brat-tool.

Categorisation

A detailed explanation of each subcategory can be found below the overview. The information consists of the category name and colour in the brat-tool, followed by the full name, a definition, important remarks, guidelines for annotation and examples. The words that should be annotated are underlined and the information after the arrow sign is an example of a possible annotation note.

grammar and syntax

- article
- comparative/ superlative
- singular/plural
- verb form
- article-noun agreement
- noun-adjective agreement
- subject-verb agreement
- reference
- missing constituent/ preposition
- superfluous word/ constituent
- word order
- structure
- other

lexicon

- wrong preposition
- wrong collocation
- named entity
- word non-existent

spelling and typos

- capitalization
- single-word spelling mistake
- compound
- punctuation
- typo

style and register

- register
- untranslated
- repetition
- disfluent sentence/ construction
- short sentences
- long sentence
- text type
- other

coherence

- conjunction
- missing information
- logical problem
- paragraph
- inconsistency
- other

GramSyn (red) Grammar & syntax:

Definition 'This does not follow the grammatical or structural rules of the Dutch language.'

article

Definition 'There is an article that should not be there, or there is no article where there should be one.'

Be careful! If you encounter a singular noun that either needs to receive an article or needs to be replaced by the plural version in order for it to be correct, also select this category.

- e.g.: ze kunnen schijnbaar terminale ziekte genezen

Be careful! (2) If the article is incorrect for the noun, select 'article-noun agreement'.

Annotation If the article is missing, select the noun it belongs to; if the article is superfluous, select the article itself. Specify the type of error in the notes-section.

- e.g.: VN-milieuprogramma waarschuwt -> missing article, should be 'het VN-milieuprogramma waarschuwt'

comparative/superlative

Definition 'The structure or form of the comparative or superlative is incorrect.'

Annotation Select the entire comparative/superlative.

- e.g.: praktischte -> meest praktische
- e.g.: een meer belangrijk verschil is dat... -> een belangrijker verschil

singular/plural

Definition 'A word that has no plural or singular has been given this form or an incorrect plural has been given.'

Annotation Select the word.

- e.g.: Ik heb gisteren een hersen gegeten -> 'hersenen' is always plural
- e.g.: Ze luistert graag muzieken -> 'muziek' has no plural

verb form

Definition 'A grammatically incorrect tense or verb form.'

Annotation If a whole verb form is incorrect, select the entire constituent (if the verbs are split up by non-verbs, select the two parts and use an arrow to link the verbs). Specify the type of error - tense or form - in the notes-section and only select the word(s) where the issue occurs.

Be careful! Only use this category if the actual verb form is incorrect or does not exist. If the form itself is correct, but the spelling is wrong (for example 'zij', 'verhuizde') select 'spelling & typos - single-word spelling mistake). If the verb exist, but is not the one intended in this context, select 'Lexicon - wrong collocation' rather than 'verb form' (for example 'wijdt' for 'wijt' or 'opgeheven' for 'opgegeven').

- e.g.: zodat vorderingen van schijnbare genezing te zien in een geschikte context -> verb form, should be 'gezien worden'
- e.g.: Ontgind -> verb form, should be 'ontgonnen'
- e.g.: ik wordt -> verb form, should be 'ik word'

article-noun agreement

Definition 'Mismatch between article and noun.'

Annotation Select both the article and the noun and link them together with an arrow.

- e.g.: de hoofd -> wrong article

noun-adjective agreement

Definition 'Mismatch between noun and adjective.'

Annotation Select the adjective and the noun separately and link the annotations together with an arrow.

- e.g. een slimme meisje -> een slim meisje

subject-verb agreement

Definition 'The subject and verb differ in number.'

Annotation Select the subject and the verb separately and link the annotations together with an arrow.

- e.g.: onze planeet...ondergaan -> 'planeet' is singular, verb should be singular as well

reference

Definition 'Mismatch between referring expression and referent'

Annotation Select the referent and the referring expression separately and link the annotations together with an arrow. In case of a compound, select the entire compound (even if the compound is incorrectly split up).

- e.g.: Ze heeft een contact verbod aangevraagd die hem verplicht om... -> een contactverbod dat

superfluous - superfluous word or constituent

Definition 'The sentence is grammatically incorrect because there is a superfluous constituent or word. This can be a constituent that already appeared earlier in the sentence or a word that has been written twice in a row.'

Annotation Select the superfluous word or constituent. If the word appears twice and either one of them can be deleted to make the sentence grammatically correct, select the first occurrence of the word.

- e.g.: Ze heeft het verteld aan de politie verteld.
- e.g.: de de hond

missing - missing constituent or preposition

Definition 'The sentence is grammatically incorrect because a necessary part of the structure is missing, this can be a preposition or an entire constituent such as an obligatory direct object.'

Be careful! Missing articles should be highlighted as 'article', not as 'missing constituent or preposition'.

Be careful! (2) Only select 'missing constituent' if a whole constituent or preposition is missing. If part of a verb form is missing (but not the whole verb), select 'grammar & syntax - verb form', not 'missing constituent'.

Annotation If a preposition is missing, select the noun phrase following the missing preposition, even if the preposition is part of a verb with a fixed preposition. If an essential subject or object is missing, select the verb. If an entire constituent that does not belong to either a noun or a verb is missing, select the sentence code at the beginning of the sentence and specify the missing element.

- e.g.: Hij begon het verzenden van berichten -> 'hij begon met het verzenden'
- e.g.: Ze abonneerde zich het nieuwe tijdschrift ->'op het nieuwe tijdschrift'

word order

Definition 'This word order is grammatically incorrect.'

Be careful! If the word order is grammatically correct, but another word order would be better (more natural), select 'style & register - disfluent', not 'grammar & syntax - word order'.

Annotation If there are only two words or parts of the sentence that should switch places, select these parts separately and link them together with an arrow. If one word or constituent needs to move to a different place in the sentence, but more than one option is possible, select the word/constituent and specify the problem. If the word order within a part of the sentence is incorrect, and more complex than inversion, select the entire segment and specify the problem. If something more complex is wrong with the word order, affecting the entire sentence, select the sentence code at the beginning of the sentence and specify the incorrect and correct word order.

- e.g.: Ten derde klimaatverandering kan de temperatuur verhogen-> Inversion after 'ten derde'

structure - other structural problems

Definition 'There is something wrong with the grammatical structure that is not just the cause of an incorrect verb form, lack of agreement or word order.'

Annotation If the problem is contained within a part of the sentence, select the entire fragment and specify the problem. If the structure of the entire sentence is affected, select the sentence code at the beginning of the sentence and specify the problem in the 'notes' section.

- e.g.: _2_ Wouden gaan verloren; wegen en dammen worden aangelegd; de populatie neemt toe: natuurlijke habitatten worden weggedaan voor landbouw; mijnbouw; en vervuiling van zeewater bevorderen omstandigheden waaronder nieuwe en oude pathogenen kunnen bloeien -> different structures in the same sentence

gram_other - other grammatical / syntactical problems

Choose this category when you encounter a grammatical / syntactical error that does not belong to any of the abovementioned categories. Always explain why you choose this category. It is possible to link two annotations together with an arrow if they are separated by non-relevant words.

Lexic (yellow) Lexical problems

Definition 'This is a lexical problem/error when looking at the Dutch language.'

wrong preposition

Definition 'This expression requires a different preposition.'

Annotation Select the preposition.

- e.g.: de ziekte komt voor in vliegende honden -> bij vliegende honden
- e.g.: Mevrouw Nestler van Santa Fe -> uit Santa Fe
- e.g.: ze was begonnen om berichten te verzenden -> met berichten te verzenden

wrong collocation

Definition 'The word(s) exist(s) in Dutch, but is/are used in the wrong or in a strange way: uncommon combinations of words, errors in fixed expressions...'

Be careful! If the combination of words is lexically correct, but not logical in the context, select 'coherence - logical problem', not 'lexicon - wrong collocation'.

Annotation Select the word or constituent that is used in a wrong way. If there is a problem with a fixed expression, select the entire expression (prepositions included!). If an inappropriate verb or adjective is used with a noun, select the entire verb or

adjective. If the expression or collocation is split up by non-relevant words, select the two parts of the collocation separately and link them together with an arrow.

- e.g.: de vergadering gaat door om 5u -> should be: vindt plaats
- e.g.: veranderingen in het milieu zullen een stijging van besmettelijke ziektes tot stand brengen -> should be: met zich meebrengen
- e.g.: in diepe nesten zitten -> should be: diep in de nesten zitten
- e.g.: De temperaturen zorgen ervoor dat pathogenen kunnen bloeien. -> 'bloeien' is something that can't be said about 'pathogenen'

named entity

Definition 'Incorrect use of named entity (geographical location, name, company, organization, etc.) in Dutch: (partially) untranslated named entities that have an official Dutch equivalent, (partially) translated named entities that do not exist in Dutch, etc.'

Annotation Select the named entity.

- e.g.: Proceedings van de National Academy of Science -> the name is 'Proceedings of the National Academy of Science', so 'of the' should not have been translated.

word non-existent

Definition 'This word does not exist in Dutch. This category also includes untranslated words that make no sense in Dutch. '

Be careful! If the non-existent word is the result of incorrect grammar, select the appropriate category in the 'grammar & syntax' section, not 'Lexicon - word non-existent'.

- e.g.: ontgind -> non-existent verb form of 'ontgonnen', belongs to grammar & syntax - verb form

Annotation Select the word. If the word is a verb that is split up by non-relevant words, select the two parts separately and connect them with an arrow.

- e.g.: Plantaties-> should be: plantages
- e.g.: Bosverlies -> should be: ontginning van bossen

Spel_Typ (blue) Spelling, typos and punctuation

Definition 'This does not follow the rules of spelling and punctuation of the Dutch language or this is a typo.'

Be careful! If there is more than one type of spelling mistake in one word, select the word twice: once for each type. For example, when a compound is split up and spelled with a capital letter, you select the word once for 'capitalization' and once for 'compound'.

capitalisation

Definition 'This does not follow the Dutch language rules of capitalisation.'

Annotation Select the whole word.

- e.g. afrika -> Afrika

spelling_mistake - single-word spelling mistake

Definition 'General spelling mistake: wrong spelling of a single word.'

Annotation Choose this category when you encounter an error other than 'capitalization' in a word that is not part of a compound. Always explain why you choose this category. Select the entire word.

- e.g.: advokaat -> advocaat
- e.g.: financiëel -> financieel

compound

Definition 'This compound is misspelled: there is a space between two elements of a one-word compound, there is a space between the first part of the compound and the hyphen, there is a superfluous hyphen... A compound can be a noun as well as an adjective, a verb or a preposition.'

Annotation Select the entire compound, unless the second part of the compound does not follow directly after the first, in which case you only select the first part.

- e.g.: anti-semitische -> antisemitische
- e.g.: groei - en leerlijnen -> no space between first part of a compound and hyphen.

punctuation

Definition 'The punctuation of the sentence is wrong or missing.'

Be careful! If a punctuation mark within a word or compound is used in the wrong way, select 'single-word spelling mistake' or 'compound', not 'punctuation'. e.g.: de jaren '90 -> single-word spelling mistake

- e.g.: brood-rooster -> compound

Be careful! (2) However, if the punctuation mark is a quotation mark, you should select 'punctuation' and select the entire word, not just the punctuation mark.

- e.g.: 'Wonder'middeltjes -> quotation marks should be around the entire word.

Annotation Select the punctuation mark. If the punctuation mark is missing, select the sentence code at the beginning of the sentence. If there are quotation marks at the beginning or end of a sentence/constituent, but not on the other side, select the quotation mark that is there, and specify the problem. If both single and double quotation marks are used, select the first quotation mark and specify the problem. If the wrong quotation marks are used on both sides, select the opening quotation mark.

- e.g.: full stop after title, semicolon between two words rather than two main clauses,...
- e.g.: Hij zei: "Ik heb het niet gedaan. -> missing quotation mark at end of quote.
- e.g.: Ze heeft het probleem 'grondig" bestudeerd. -> Both single and double quotation marks are used (better use single quotation marks for irony)
- e.g.: Ze heeft het probleem "grondig" bestudeerd. ->use single quotation marks to express irony.

typo

Definition 'This error was probably caused by typing too fast: letters have been switched, there is a letter missing or too many, there's a superfluous or missing space,... '

Be careful! Do not interpret spelling mistakes as typos. Only select 'typo' if the error cannot be classified elsewhere, otherwise select 'spelling, typos & punctuation - single-word spelling mistake' or the appropriate category.

Annotation Select the entire word, not just the few letters that need to be changed. If a word has been typed twice, select the first time it occurs. If there is a space before a punctuation mark, select the punctuation mark.

- e.g.: dennken -> denken
- e.g.: De vis zwom in het water . -> no space before a punctuation mark

Style_reg (pink) Style and register

register

Definition 'The words have the same meaning, but the chosen word/expression is too formal/informal/... for the text or belongs to a regional variety of the language that is not entirely suitable for the target audience.'

Annotation Select the word(s) or expression from the inappropriate register. If they are split up by non-relevant words, select the two parts separately and connect them with an arrow.

- e.g.: dat wijf -> too informal / vulgar, should be 'die vrouw' in this text

untranslated

Definition 'A word / fragment of which a Dutch translation exists is left untranslated.'

Annotation Select all the untranslated words.

- e.g.: drie 183-square-foot spiegels -> use Dutch measurements

repetition

Definition 'The same or a very similar word/expression is used too often or is too close to the previous occurrence of the word/expression, it is better to use a synonym.'

Annotation Select the first occurrence of this word/expression and specify why/where it has been used too often in the notes. If the expression or constituent is split up by non-relevant words, select the two parts separately and connect them with an arrow.

disfluent - disfluent sentence/construction

Definition 'The sentence / constituent is not grammatically incorrect, but it is nonetheless very difficult to read, it could be translated in a much more idiomatic way.'

Annotation Select the sentence code at the beginning of the sentence and specify the problem.

short sentence

Definition 'This segment contains too many short sentences, which affects the readability.'

Annotation Select the sentence code at the beginning of the first short sentence.

long_sentence

Definition 'This sentence is too long, it would benefit the readability if this sentence were split up.'

Annotation Select the sentence code at the beginning of the sentence.

text type

Definition 'This is not necessarily a problem, but the text type allows or requires different constructions or deletions.'

Annotation If the issue is contained in one word or constituent, select that word/constituent. If the constituent is split up by non-relevant words, select the two parts of the constituent separately and connect them with an arrow. If the issue affects an entire sentence, select the sentence code at the beginning of the sentence where the issue occurs.

- e.g.: _4_ Tijdens een boswandeling hebben de scouts een schat gevonden -> no articles in the title of a newspaper article, 'scouts vinden schat tijdens boswandeling' is better
- e.g.: 'Meneer Letterman zei dat hij...' -> 'Letterman zei dat hij...': usually only the last name is used in a Dutch newspaper article

style_other - other style & register problems

Choose this category when you encounter a style/ register problem that does not belong to any of the abovementioned categories. Always explain why you choose this category. It is possible to link two annotations together with an arrow if they are separated by non-relevant words.

Coher (orange) Coherence

Definition 'There is something wrong with the coherence of the text: confusing relationships, lack of logical structure, undeveloped paragraphs,...'

conjunction

Definition 'The conjunction or linking word expresses a strange relationship or there seems to be a missing relationship.'

Annotation If there is a conjunction/linking word, select this word. If the conjunction or linking word is missing, select the sentence code at the beginning of the sentence and specify the problem in the notes.

missing info

Definition 'Information that is needed to easily understand the text is missing, thus reducing readability. This includes cases of implicitation.'

Annotation Select the constituent that is affected by the missing info. If the constituent is split up by non-relevant words, select the two parts of the constituent separately and link them together with an arrow. If the entire sentence is affected, select the sentence code at the beginning of the sentence and specify the problem in the notes.

logical problem

Definition 'This does not follow the logical structure of the text. The idea contradicts something that has been previously stated, or the information as such is illogical/confusing when looking at the rest of the text.'

Annotation If the logical problem is situated in one word or a part of the sentence, select this word/constituent. If the constituent is split up by non-relevant words, select the two parts of the constituent separately and link them together with an arrow. If the problem affects the entire sentence, select the sentence code at the beginning of the sentence and specify the problem in the notes.

paragraph

Definition 'The paragraph is not well-developed, contains more than one idea or the information belongs to a previous paragraph.'

Annotation Select the sentence code at the beginning of the paragraph and specify the problem(s) in the notes.

inconsistency

Definition 'Terms or notations are used inconsistently throughout the text.'

Annotation Select the first inconsistent occurrence of the word and specify the variant + amount of times they occur. If it is a problem contained in one word, use the following construct: "(word1): (number of occurrences)x, (word2): (number of occurrences)x, (comment)". If it is a different type of problem, specify the problem in the notes-section.

Be careful! If the capitalization of words is inconsistent throughout the text, select the text code at the top of the text and specify the problem in the notes-section.

- e.g.: Deze ziektes zijn dodelijk voor de mens. (...) Andere besmettelijke ziekten kunnen genezen worden door... -> ziektes: 3x, ziekten: 4x, a different plural is used throughout the text

coh_other - other coherence problems

Choose this category when you encounter a coherence problem that does not belong to any of the abovementioned categories. Always explain why you choose this category. It is possible to link two annotations together with an arrow if they are separated by non-relevant words.

Preparing the text for the next step

Once you've finished annotating the text for acceptability, you need to prepare it for the adequacy annotations. To make sure you are not distracted by the annotations you've already made, you must turn off the visibility of these annotations. You can switch off visibility by clicking on 'visible layers' at the top of your screen within the brat tool. Below 'Entities', click on all acceptability categories so that they turn blue (Grammar and syntax, Lexicon, Spelling and typos, Style and register, Coherence) and make sure 'Meaning shift 'is still orange. Click 'ok' and you should see the annotations disappear.

Next, select 'Collection' at the top of your screen and click on the 'ST' folder. Select the text with the same name as the text you just annotated and open it. At the top of your screen, select 'Data' and click 'Comparison mode'. Go to the folder which contains the texts that you've annotated for acceptability and select the text with the same name.

You should now see the source text on the left hand side of the screen, and the translation on the right hand side of the screen. Make sure you keep the source text opened in a separate browser tab as well. You are now ready to annotate the text for adequacy.

Step 2: Annotating adequacy

The goal of the current assignment is to annotate translations for adequacy. Acceptability will be dealt with separately.

Adequacy can be described as making sure the target text contains the same information as the source text. This means that all misinterpretations, contradictions, meaning shifts, additions or deletions are potential errors. Depending on the text type, a translator could be allowed to delete some details or to add a little extra information, but these are general possible translation problems. In a technical manual or medical document, for example, strict adherence to the source text information is required.

As a reviser, it is your task to make sure the translation reaches publishable quality based on information adequacy. You do this by marking all types of meaning shift, while taking the demands of the text type into account. Although the texts may contain errors against the Dutch language, this is not your concern, unless the errors also bring about a shift in meaning (in which case you are required to mark the shift). To facilitate the task, you can use the brat rapid annotation tool. In this tool, the different types of meaning shift are predefined.

Categorisation

A detailed explanation of each category can be found below the overview. The information consists of the category name in the brat-tool (the colour is always green

for adequacy annotations), followed by a definition, important remarks, guidelines for annotation and examples. The words that should be annotated are underlined and the information after the arrow sign is an example of a possible annotation note.

- contradiction
- word sense disambiguation
- part of speech
- hyponymy
- hyperonymy
- terminology
- quantity
- time
- meaning shift caused by punctuation
- meaning shift caused by incorrect translation of function word
- meaning shift caused by misplaced word
- deletion
- addition
- explicitation
- coherence
- inconsisent terminology
- other meaning shift

Contradiction

Definition 'The target text contradicts the source text.'

Annotation Select the entire constituent / section of the sentence that contains the contradiction. If the contradiction is split up by non-relevant words, select the two parts separately and link them together with an arrow.

- e.g.: EN: The clearing of natural habitats for agriculture causes climate change

NL: Het verdwijnen van landbouwgebieden veroorzaakt klimaatverandering

->In the source text, the habitats disappear so that there can be agriculture, in the target text it is the agriculture that disappears.

Word_sense - word sense disambiguation

Definition 'The Dutch word is a possible translation of the word in the ST, but not of the meaning the word has in this context.'

Annotation Select the word(s). If the translation is split up by non-relevant words, select the two parts separately and link them together with an arrow.

Be careful! If the word is a function word, select 'function word', not word sense disambiguation.

- e.g.: EN: The frequency of this phenomenon should be appreciated so that claims of apparent cure by novel treatment strategies can be seen in an appropriate context.

  NL: ... zodat vorderingen van schijnbare genezing in de juiste context gezien kunnen worden

  ->'claim' can mean 'vordering' in a legal context, but here 'bewering' is meant.

- e.g.: EN: Climate change aggravates the threats of infectious diseases by further stressing habitats

  NL: Klimaatverandering verhoogt het risico op besmettelijke ziektes door meer de nadruk te leggen op habitats

  ->'stress' can mean 'nadruk leggen op', but here 'onder druk zetten' is meant.

Part of Speech

Definition 'The Dutch word is semantically related to the word in the ST, but an incorrect grammatical category has been chosen.'

Annotation Select the word(s). If the translation is split up by non-relevant words, select the two parts separately and link them together with an arrow.

- e.g.: EN: Before, when the animals could talk...

  NL: Voordat, toen de dieren konden spreken...

  -> 'Before' is used as an adverb here, not as a conjunction.

Hyponymy

Definition 'The target text contains a hyponym of the word/expression used in the source text, even though it could have been translated in an equally specific way.'

Annotation Select the hyponym.

- e.g.: EN: he sent her love messages

  NL: hij stuurde haar liefdesbrieven

  -> 'love letters' is a hyponym of 'love messages' and there was no mention of 'letters' in the text

Hyperonymy

Definition 'The translation contains a superordinate term or expression even though it could have been translated in an equally specific/explicit way.'

Annotation Select the hypernym.

- e.g.: EN: infectious diseases

  NL: ziektes

- e.g.: EN: herbal potions

  NL: kruidengeneesmiddelen

  -> potions are liquids, 'geneesmiddelen' is a hypernym

Terminology

Definition 'Non-adherence to the specified terminology guidelines: The source text word was an entry in the glossary, yet the translator chose a different word as a translation. '

Annotation Select the term.

Be careful! If the translation is an incorrect translation of a word that was present in the glossary, annotate it both as a 'wrong word sense' or 'other' problem (depending on the nature of the incorrect translation) and as a 'terminology' problem.

Be careful! (2) If the translation is one of the possible translations of the word according to the glossary, but not the one needed in this context, select 'wrong word sense', not terminology.

Be careful! (3) If the translation is problematic when looking at the glossary, but it is also inconsistent when looking at previous occurrences of the same source text term, select both 'terminology' and 'inconsistent terminology'.

- e.g.: EN: application

  NL: applicatie

  Glossary: toepassing

  -> 'toepassing' is the appropriate term to use here, according to the glossary.

Quantity

Definition 'The number or amount mentioned in the target text is different from that in the source text.'

Annotation If the number is different, select the word. If the amount is different, select the quantifying element

- e.g.: EN: I brought the books back to the library

  NL: Ik bracht het boek terug naar de bibliotheek

- e.g.: EN: He paid her 500 euros.

  NL: Hij betaalde haar 50 euro.

Time

Definition 'The time of the target text is different from the one in the source text, which changes the meaning. This can be caused by an incorrect verb tense or temporal element.'

Be careful! Only select this category if there is a difference in meaning between source text and target text, not if the verb is grammatically incorrect.

Annotation Select the verb or the temporal element. If the verb is split up by non-relevant elements, select the two parts separately and connect them with an arrow.

- e.g.: EN: We receive a lot of requests

  NL: We kregen veel verzoeken

  -> The translation makes it sound as if it is something from the past, whereas the present tense in the source text indicates that the statement is still true.

- e.g.: EN: He always drinks coffee in the evening

  NL: Hij drinkt 's ochtends altijd koffie

Punctuation - meaning shift caused by punctuation

Definition 'The punctuation mark alters the meaning as expressed in the source text.'

Annotation Select the punctuation mark and specify the different meaning. If the punctuation mark is missing, select the sentence code at the beginning of the sentence and specify the change in meaning.

- e.g.: Exclamation/statement/question

Function_word - meaning shift caused by incorrect translation of function word

Definition 'The chosen translation for the function word leads to a change in meaning between source and target text.'

Annotation Select the function word.

- e.g.: EN: they went to the store.

  NL: ze gingen om de winkel.

  -> 'to' can mean 'om', but here it means 'naar'.

Misplaced_word - meaning shift caused by misplaced word

Definition 'The element in the target text modifies a different word/constituent than the one in the source text, thus changing the meaning.'

Annotation Select the entire constituent affected by the misplaced word or constituent and specify the problem in the notes.

- e.g.: EN: the reproduction of the handwritten report

  NL: de handgeschreven weergave van het rapport

  -> the report is handwritten, not the reproduction

Deletion

Definition 'A meaningful element of the source text or relationship present in the source text is deleted in the target text.'

Be careful! Cases of hyperonymy must be annotated as such, not as cases of deletion.

Annotation Select the elements that are deleted in the English source text. You cannot edit the source text in the comparison mode in brat, so do this on the source text in a separate browser tab. You can refresh the comparison page to make the source text annotations appear in the comparison screen as well. If two parts of deleted information are split up by non-relevant words, select the two parts separately and link them

together with an arrow. If the deletion is more complex, select the sentence code at the beginning of the sentence and specify the deletion in the 'notes' section.

Addition

Definition 'A meaningful element that cannot be derived from the source text is added in the target text.'

Annotation Select the added information. If the addition is split up by non-relevant words, select the two elements separately and link them together with an arrow.

Explicitation

Definition 'Information that can be derived from the source text but that is not necessary for the reader to understand the information presented in the text is added to the translation.'

Annotation Select the entire constituent / fragment that contains the information. If the fragment is split up by non-relevant words, select the two parts separately and connect them with an arrow.

- e.g.: EN: The United Nations Environment Programme has warned that changes to the environment bring about a rise in diseases.

  NL: Het VN-milieuprogramma (UNEP: United Nations Environment Programme) heeft gewaarschuwd dat milieuveranderingen een stijging van ziektes met zich meebrengen

  -> The English name does not give the reader extra information that he/she needs in order to understand what 'VN-milieuprogramma' is, and no meaningful elements are added, as 'UNEP' and 'VN-milieuprogramma' refer to the same organisation.

Coherence

Definition 'The conjunction expresses a different relation in the target text than in the source text.'

Annotation Select the conjunction and specify the difference in relation.

- e.g.: EN: new diseases arise as environments destroyed

  NL: nieuwe ziektes komen op terwijl/wanneer omgevingen worden kapotgemaakt

  -> shift of causal to temporal relationship

- e.g.: EN: the judge granted the request, noting the restraining order was granted merely as a matter of proper pleading

NL: de rechter keurde het verzoek goed, maar wees erop dat...

-> the judge simply gave an explanation; there is no relationship of contrast in the original text.

Inconsistent_terminology

Definition: 'The same term was used throughout the source text, but the translator uses different terms throughout the target text.'

Annotation: Select the first inconsistent occurrence of the term and specify the problem in the notes-section.

Other - other meaning shift

Choose this category when you encounter a meaning shift that does not belong to any of the abovementioned categories. Always explain why you choose this category. If two elements of the meaning shift are split up by non-relevant elements, select the two parts separately and connect them with an arrow. If the problem affects the entire sentence, select the sentence code at the beginning of the sentence and specify the problem in the notes.

Consolidation phase: identifying the origin of errors + creating 'gold standard'

Linking acceptability to adequacy issues

Some of the identified acceptability problems find their origin in adequacy problems. In order to make this relationship clear, the annotations need to be linked together.

Open the annotated translation in the brat tool and make sure you adjust the visibility again: at the top of the page, click 'visible layers' and make sure all entities are orange, not blue. You should now see all annotations (both acceptability and adequacy annotations).

Look at the acceptability annotations. If you find an adequacy error that seems to be the cause for the acceptability error, draw an arrow from the acceptability annotation to the adequacy annotation (the relationship should be 'caused_by').

For example: a part of the sentence did not make sense (acceptability - logical problem), because a word was mistranslated (adequacy - word sense error). Or a necessary part of the structure was missing (acceptability - missing) because a source text word was deleted (adequacy - deletion).

It is possible for multiple acceptability problems to be caused by the same adequacy problem, and for one acceptability problem to be caused by more than one adequacy problem.

Creating a gold standard

Though annotations made by one annotator can already provide a lot of information about the quality of a text, translation quality assessment is a rather subjective task. We therefore advise introducing a consolidation phase between the annotations of at least two different annotators. More might of course even be better, but is often not possible due to time or other constraints. The following paragraphs contain guidelines for the consolidation phase with two annotators.

Warning: If you want to keep the original annotations for future reference, make a copy of the annotation files before you begin with the consolidation phase.

Open the annotated texts from both translators in comparison mode. The editable text (right hand side of the screen) will become the gold standard text.

In a first step, annotator 1 (left hand side) looks at the annotations made by annotator 2 (right hand side) and discards the annotations from annotator 2 that he does not agree with. If both annotators highlighted the same problem, but with a different span, annotator 1 makes sure the span is corrected on the right hand side. In a second step, annotator 2 (right hand side) looks at the annotations made by annotator 1 (left hand side) and adopts the annotations from annotator 1 that he agrees with.

After these preliminary steps, the following annotations are left: annotations that both annotators agree on, and annotations of segments that both annotators believe are problematic, but where they highlighted different categories. Annotators should discuss the annotations that are still problematic together and come up with a consolidated version based on the guidelines and mutual agreement. If the same issues occur multiple times, it might be necessary to add rules to the guidelines or to add extra subcategories to accommodate these new issues.

Appendix 3: Texts and MT output main experiment

Source text 1

Listen up. This blue whale's earwax tells the story of its life and locale

A giant plug of earwax pulled from a dead blue whale is providing scientists with a detailed biography of the wild animal's life, from birth to death, in 6-month chapters.

The scientists' new technique is described in the journal Proceedings of the National Academy of Science.

It arms researchers with a tool to understand a whale's hormonal and chemical biography — and a window into how pollutants, some long discontinued, still remain in the environment today.

Whales are often called marine sentinels because they can reveal a lot about the waters they swim through, said Sascha Usenko.

He is an analytic environmental chemist at Baylor University.

"These types of marine mammals that are long-lived have a great ability to accumulate contaminants, and so they're often perceived as being sentinels of their ecosystem," said Usenko, who helped write the study.

MT output text 1

Luister. Oorsmeer Deze blauwe walvis vertelt het verhaal van zijn leven en locale.

Een gigantische stekker van oorsmeer getrokken uit een dode blauwe vinvis is het verstrekken van wetenschappers met een gedetailleerde biografie van het leven van het wilde dier, van geboorte tot dood, in 6 maanden hoofdstukken.

Nieuwe techniek van de wetenschappers wordt beschreven in het tijdschrift Proceedings van de National Academy of Science.

Het armen onderzoekers met een tool om hormonale en chemische biografie van een walvis te begrijpen - en een raam in hoe verontreinigende stoffen, sommige lang gestaakt, nog steeds in het milieu blijven vandaag.

Walvissen worden vaak genoemd zeemilieu, omdat ze veel over de wateren ze doorheen te zwemmen kan openbaren, zei Sascha Usenko.

Hij is een analytische milieuchemicus aan de Baylor University.

"Deze soorten zeezoogdieren die lange geleefd hebben een groot vermogen om verontreinigende stoffen zich ophopen, en dus zijn ze vaak gezien als schildwachten van hun ecosysteem," zei Usenko, die hielp schrijf de studie.

Source text 2

Done with mirrors: Bringing the sun to a small Norwegian town

Tucked between steep mountains, Rjukan is normally shrouded in shadow for almost six months a year.

Residents have to catch a cable car to the top of a nearby precipice to get a dose of midday vitamin D.

But on Wednesday, faint rays from the winter sun for the first time reached the town's market square, thanks to three 183-square-foot mirrors placed on a mountain.

Cheering families, some on sun loungers, drinking cocktails and waving Norwegian flags, donned shades as the sun crept from behind a cloud.

It hit the mirrors and reflected down onto the faces of delighted children below.

"Before when it was a fine day, you would see that the sky was blue and you knew that the sun was shining. But you couldn't quite see it. It was very frustrating," said Karin Roe, from the local tourist office.

MT output text 2

Gedaan met spiegels: Bringing de zon om een kleine Noorse stad

Verscholen tussen steile bergen, is Rjukan normaal gehuld in de schaduw voor bijna zes maanden per jaar.

Bewoners hebben een kabelbaan nemen naar de top van een nabijgelegen afgrond aan een dosis van de middag vitamine D te krijgen

Maar op woensdag, zwakke stralen van de winterzon voor het eerst bereikte het marktplein van de stad, dankzij drie 183-square-foot spiegels geplaatst op een berg.

Juichen families, sommige op de ligstoelen, cocktails drinken en zwaaien Noorse vlaggen, trok tinten als de zon kroop van achter een wolk.

Het raakte de spiegels en gereflecteerd naar beneden op de gezichten van opgetogen kinderen hieronder.

"Voordat toen het nog een mooie dag, zou je zien dat de lucht blauw was en je wist dat de zon scheen. Maar je kon niet goed zien. Het was erg frustrerend," zei Karin Roe, van het lokale VVV-kantoor .

Source text 3

China's "orphan grandparents" can sue absent children for not visiting

Yan Meiyue, 90, said her 72-year-old daughter rarely visited, even for the annual Spring Festival, when families traditionally reunite.

So Yan, a widow since her husband's death nearly a decade ago, spends every weekday at a modest community center near her home, where she plays mahjong and eats meals prepared by a volunteer staff.

"The volunteers keep us company," she said with a smile, her voice trailing off.

Yan is one of a rapidly growing number of self-described "orphan grandparents" who feel personally or financially abandoned.

It's a troubling trend for China where elders have traditionally been among the most respected members of society.

For centuries, Chinese households have included many generations, and Chinese elders could count on their children caring for them as they grew frail.

But today this ancient social contract is giving way.

The booming Chinese economy is prying apart families with job opportunities that lure adult children to distant cities or other countries.

MT output text 3

China's " orphan grootouders " kan afwezig kinderen aanklagen voor het niet bezoeken van

Yan Meiyue , 90 , zei dat haar 72 - jarige dochter zelden bezocht , zelfs voor de jaarlijkse Lente Festival , waarin families traditioneel herenigen .

Dus Yan , een weduwe sinds de dood van haar man bijna een decennium geleden , besteedt elke weekdag op een bescheiden buurthuis in de buurt van haar huis , waar ze speelt mahjong en eet maaltijden, bereid door een vrijwilliger personeel .

" De vrijwilligers houden ons bedrijf," zei ze met een glimlach , haar stem achterstand .

Yan is een van een snel groeiend aantal zelf-beschreven " wees grootouders " die het gevoel hebben persoonlijk of financieel verlaten .

Het is een verontrustende trend voor China, waar de oudsten van oudsher een van de meest gerespecteerde leden van de samenleving zijn geweest .

Eeuwenlang hebben Chinese huishoudens vele generaties inbegrepen , en Chinese ouderen konden rekenen op hun kinderen de zorg voor hen als ze groeiden broos .

Maar vandaag deze oude sociaal contract geeft manier .

De bloeiende Chinese economie nieuwsgierige apart gezinnen met vacatures die volwassen kinderen te lokken naar verre steden of andere landen .

Source text 4

Lie-detector test comes under fire as FBI hiring tool

Thousands of job seekers come to FBI offices all across the country every year, only to be turned away from the top U.S. law enforcement agency.

The reason is not because they do not have the right work experience or education, or because they have a criminal record.

They are turned down because they failed their polygraph tests.

The polygraph is also known as a lie detector.

Many scientists disagree with the FBI's policy of rejecting candidates who fail the tests.

They say government agencies should not rely solely on the tests to decide whether to hire or fire someone.

Experts say polygraph testing does not reliably show when somebody is actually lying, especially when they are applying for a job.

“I was called a lazy, lying, drug-dealing junkie by a man who doesn’t know me, my stellar background or my societal contributions,” wrote one applicant in Baltimore.

MT output text 4

Leugendetectortest komt onder vuur als FBI inhuren hulpmiddel

Duizenden werkzoekenden komen FBI kantoren over het hele land elk jaar, maar uit de buurt van de top van de Amerikaanse politie om worden gedraaid.

De reden is niet omdat ze niet over de juiste werkervaring of opleiding, of omdat ze een strafblad hebben.

Ze worden afgewezen omdat zij hun polygraaf testen mislukt.

De leugendetector is ook bekend als een leugendetector.

Veel wetenschappers het niet eens met de FBI beleid van afwijzing kandidaten die de tests mislukken.

Ze zeggen dat de overheid moet niet alleen vertrouwen op de tests om te beslissen of te huren of ontslaan iemand.

Experts zeggen leugendetector test niet betrouwbaar om als iemand daadwerkelijk ligt, vooral wanneer zij solliciteren naar een baan.

"Ik was een luie, liegen, drugsdealende junkie gebeld door een man die mij niet kent, mijn stellaire achtergrond of mijn maatschappelijke bijdragen," schreef een aanvrager in Baltimore.

Source text 5

Huge art pieces, done on an iPad, draw gasps at museum exhibit

Britain's most celebrated living artist, David Hockney, is pioneering in the art world again.

Happily hunched over his iPad, he is using his index finger like a paintbrush to create colorful landscapes and richly layered scenes on a touch screen.

"It's a very new medium," said Hockney.

So new, in fact, he wasn't sure what he was creating until he began printing his digital images a few years ago.

"I was pretty amazed by them actually," he said, laughing. "I'm still amazed."

A new exhibit of Hockney's work, including many iPad images, opened Saturday in San Francisco's de Young Museum.

Located in Golden Gate Park, the museum is just a short trip for Silicon Valley techies who created both the hardware and software for this 21st-century reinvention of finger-painting.

The show is billed as the museum's largest ever.

MT output text 5

Enorme kunstwerken, gedaan op een iPad, trekken hapt op museumstuk

Groot-Brittannië's meest gevierde levende kunstenaar, David Hockney, is een pionier in de kunstwereld weer.

Gelukkig gebogen over zijn iPad, is hij met behulp van zijn wijsvinger als een penseel aan kleurrijke landschappen en rijk gelaagde scènes op een touchscreen te maken.

"Het is een heel nieuw medium," zei Hockney.

Zo nieuw, in feite was hij niet zeker wat hij creëerde, totdat hij begon te drukken zijn digitale foto's een paar jaar geleden.

"Ik was behoorlijk verbaasd door hen eigenlijk," zei hij lachend. "Ik ben nog steeds verbaasd."

Een nieuwe tentoonstelling van Hockney's werk, waaronder veel iPad beelden, opende zaterdag in de Young Museum in San Francisco.

Gelegen in het Golden Gate Park, het museum is slechts een korte reis voor Silicon Valley techneuten die zowel de hardware als software gemaakt voor dit 21e-eeuwse heruitvinding van vingerverven.

De show wordt aangekondigd als museum de grootste ooit het.

Source text 6

Scared and scarred by the global crisis, families hoard their money

Although they speak different languages, live in wealthy countries and poor ones, face good job markets and bad, when it comes to money they are acting as one.

Families are holding tight to their cash, driven more by a fear of losing what they have than a desire to increase it.

An Associated Press study of households in the 10 biggest economies shows that families continue to spend cautiously.

They have pulled hundreds of billions of dollars out of the stock market and cut their borrowing for the first time in decades.

They are putting their money into savings and investments that offer low interest payments, often too small to keep up with the cost of living increase each year.

"It doesn't take very much to destroy confidence, but it takes an awful lot to build it back," says Ian Bright, a senior economist at ING, a global bank based in Amsterdam.

MT output text 6

Bang en getekend door de wereldwijde crisis, gezinnen hamsteren hun geld

Hoewel ze verschillende talen spreken, leven in rijke landen en arme landen, geconfronteerd met goede arbeidsmarkt en slecht, als het gaat om geld dat ze handelen als een.

Gezinnen zijn strak vast te houden aan hun geld, meer gedreven door een angst om te verliezen wat ze hebben dan een verlangen om het te verhogen.

Een Associated Press onderzoek van de huishoudens in de 10 grootste economieën blijkt dat gezinnen blijven voorzichtig door te brengen.

Ze hebben honderden miljarden dollars getrokken uit de aandelenmarkt en snijd hun leningen voor het eerst in decennia.

Ze zetten hun geld in sparen en beleggen dat betalingen lage rente, vaak te klein te houden met de kosten van levensonderhoud elk jaar bieden.

"Het maakt niet heel veel voor nodig om het vertrouwen te vernietigen, maar het kost heel veel om het terug te bouwen," zegt Ian Bright, een senior econoom bij ING, een wereldwijde bank gevestigd in Amsterdam.

Source text 7

Some young Iranians ignore officially enforced anger at the West

World leaders in Geneva negotiated the future of Iran's nuclear development program.

In exchange for limiting uranium enrichment, Iran will be freed from certain trade restrictions, known as sanctions.

But religious hard-liners here continued to warn of a deceitful West scheming to weaken the Islamic Republic.

Yet things are different for the mostly young, jeans-clad set in this busy capital city.

Among them, chanting denunciations of the United States is as out of date as 1970s fashion.

"In art, in fashion, in cinema and in our daily lifestyle, we copycat American culture," said Sarah.

She is the proprietor of a cozy cafe in the basement of a high-rise in northwestern Tehran.

"There is a big difference between the approved culture and the reality of urban lifestyles in big cities like Tehran."

Just as Western views of Iran are far from monolithic, the view here is diverse.

MT output text 7

Sommige jonge Iraniërs negeren officieel afgedwongen woede op het Westen

Wereldleiders in Genève onderhandeld over de toekomst van de nucleaire ontwikkeling van Iran.

In ruil voor de beperking van de verrijking van uranium, zal Iran bevrijd worden van bepaalde handelsbeperkingen, bekend als sancties.

Maar religieuze hardliners hier blijven om te waarschuwen voor een bedrieglijke West gekonkel om verzwakking van de Islamitische Republiek.

Toch liggen de zaken anders voor de veelal jonge, jeans geklede set in deze drukke hoofdstad.

Onder hen, zingen veroordelingen van de Verenigde Staten is zo verouderd als 1970 mode.

"In de kunst, in de mode, in de bioscoop en in ons dagelijks leven, we copycat Amerikaanse cultuur", zegt Sarah.

Zij is houdster van een gezellig cafe in de kelder van een hoogbouw in het noordwesten van Teheran.

"Er is een groot verschil tussen de goedgekeurde cultuur en de realiteit van de stedelijke levensstijl in grote steden als Teheran."

Net zoals westerse standpunten van Iran zijn verre van monolithische, het uitzicht hier is divers.

Source text 8

Climate change could lead to more wars and civil unrest, a study says

The theory that high temperatures fuel aggressive and violent behavior is only just beginning to be studied.

Using examples ranging widely from road rage, ancient wars and Major League Baseball, scientists have taken early steps to quantify the potential influence of climate warming on human conflict.

Three researchers at the University of California, Berkeley, have pulled together data from these and other studies.

They concluded that outbreaks of war and civil unrest may increase by as much as 56 percent by 2050 because of higher temperatures and extreme rainfall patterns predicted by climate change scientists.

Likewise, episodes of personal violence — murder, assault, rape, domestic abuse — could increase by as much as 16 percent.

Their study was published on Aug. 1 by the journalScience.

“We find strong causal evidence linking climatic events to human conflict ... across all major regions of the world,” the researchers concluded.

MT output text 8

Klimaatverandering kan leiden tot meer oorlogen en burgerlijke onrust , een studie zegt

De theorie dat hoge temperaturen brandstof agressief en gewelddadig gedrag is slechts het begin te worden bestudeerd .

De hand van voorbeelden op grote schaal , variërend van Road Rage , oude oorlogen en de Major League Baseball , hebben wetenschappers vroeg stappen om de mogelijke invloed van de klimaatopwarming op de menselijke conflicten kwantificeren genomen .

Drie onderzoekers van de Universiteit van Californië , Berkeley , hebben samen gegevens getrokken uit deze en andere studies .

Zij concludeerden dat het uitbreken van de oorlog en burgerlijke onrust kan toenemen met maar liefst 56 procent in 2050 als gevolg van hogere temperaturen en extreme regenval patronen voorspeld door klimaatverandering wetenschappers .

Ook afleveringen van persoonlijk geweld - moord , mishandeling , verkrachting , huiselijk geweld - zou kunnen toenemen met maar liefst 16 procent .

Hun studie werd gepubliceerd op 1 augustus door de journalScience .

" We zien een sterke causale bewijsmateriaalaaneenschakeling weersomstandigheden voor de menselijke conflicten ... in alle belangrijke regio's van de wereld ", concludeerden de onderzoekers .

Appendix 4: Participant survey before experiment

Q2 Thank you for wanting to participate in this experiment as part of the ROBOT-project. By answering these questions, you grant the project's researchers permission to use the information provided within the framework of the research project. If you wish, you may request a summary of the findings via e-mail. The information provided will be processed anonymously and will never be given to third parties without participants' express permission. Please answer truthfully. The information will only be used to get a general idea of the background of the participants in this experiment and will never be used to explicitly refer to one possible participant. Thus, your answers will not impact your relationship with other people (professors, employers, fellow students, colleagues, etc.). You will be able to add extra feedback to some of the questions, please use the space provided for any additional comments you may have regarding your answer to that question.

Q3 GENERAL

1. PERSONAL INFORMATION

Q4 Name:

Q34 Sex:
- ❍ male
- ❍ female
- ❍ other

Q5 Age:

Q36 I am a ...:
- ❍ student [if selected, participant sees Q6-Q7-Q8]
- ❍ professional translator [if selected, participant sees Q44-Q37-Q38-Q39-Q40]
- ❍ other, please specify _____ [if selected, participant sees Q41]

Q6 I study at (college/university):

Q7 Name of the study programme:

Q8 Most of the classes I take belong to:
- ❍ Bachelor 1
- ❍ Bachelor 2
- ❍ Bachelor 3
- ❍ Master 1
- ❍ Master 2
- ❍ other, please specify _____

Q44 I have obtained the following degree(s):
- ❑ Translation
- ❑ Interpreting
- ❑ Multilingual Communication
- ❑ Linguistics
- ❑ Other translation or linguistic degree, please specify: _____
- ❑ Other degree, please specify: _____
- ❑ I have not obtained a degree

[if Q44 not 'I have not obtained a degree'] Q45 Where did you obtain this degree?

[if Q44 not 'I have not obtained a degree'] Q46 When did you graduate? (last degree obtained)

Q37 Number of years experience as a translator:

Q38 I translate ... (number) words on average each year:

Q39 Specify what percentage of your total time translating is spent on the following subjects:
_____ Technical translation
_____ Judicial translation
_____ Medical translation
_____ Economic translation
_____ General translation
_____ Other (please specify)

Q40 You may provide additional information here (explicitation of the percentages, explanation of the category 'other', etc.):

Q41 Elaborate on your position and translation or linguistic experience with different text genres. Questions about language skills will follow.

Q10 2. LANGUAGE SKILLS

Q11 My native language:

Q13 I feel like I translate best from the following (not native) language:

Q14 I also translate well from the following language (not native, leave blank if not applicable):

Q15 I also translate from the following language(s) (not native, leave blank if not applicable):

Q16 Other languages I know (leave blank if not applicable):

Q42 I ...

- ❍ never translate from my native language
- ❍ sometimes translate from my native language
- ❍ translate from my native language just as often as from other languages
- ❍ usually translate from my native language
- ❍ always translate from my native language

[if Q42 not 'never'] Q43 When translating from your native language, what language(s) do you translate into?

Q17 Additional remarks:

Q18 MACHINE TRANSLATION AND POST-EDITING

Q19 I have heard of the following machine translation systems:
- ❑ *Google Translate*
- ❑ SYSTRAN
- ❑ Bing Translator
- ❑ Babylon
- ❑ Reverso
- ❑ Moses
- ❑ Language Studio (Asia Online)
- ❑ Apertium
- ❑ SDL
- ❑ Other (please specify) _____

Q20 When I translate, I ... (fill in the blank) use machine translation.
- ❍ always [participant sees Q21-Q47-Q22]
- ❍ often [participant sees Q21-Q47-Q22]
- ❍ sometimes [participant sees Q21-Q47-Q22]
- ❍ never [participant sees Q26-Q24-Q36-Q32-Q33]

Q26 Why not?

Q21 List the machine translation systems you use:

Q47 How do you use machine translation in your translations? Why do you use it?

Q22 I have post-edited an entire text before
- ❍ Yes [participant sees Q23-Q35-QQ25-Q27-Q28-Q29-Q30-Q31]
- ❍ No [participant sees Q24-Q36-Q32-Q33]

Q23 Which of the following best describes you?
- ❍ I find post-editing more rewarding than manual translation
- ❍ I find post-editing more rewarding than manual translation, but I do not mind manual translation.
- ❍ I find post-editing and manual translation equally rewarding.
- ❍ I find manual translation more rewarding, but I do not mind post-editing.
- ❍ I find manual translation more rewarding than post-editing.

Q35 You may provide additional information to the above question here.

Q24 Which of the following best describes you?

- ❍ I think I would find post-editing more rewarding than manual translation.
- ❍ I think I would find post-editing more rewarding than manual translation, but I would not mind manual translation.
- ❍ I think I would find post-editing and manual translation equally rewarding.
- ❍ I think I would find manual translation more rewarding, but I would not mind post-editing.
- ❍ I think I would find manual translation more rewarding than post-editing.

Q36 You may provide additional information to the above question here.

Q25 I find that the output of machine translation systems is ... (fill in the blank)

- ❍ ...always useful
- ❍ ...often useful
- ❍ ...sometimes useful
- ❍ ...never useful

Q27 You may provide additional information to the above question here.

Q28 I think I translate...

- ❍ ...faster when post-editing than when translating manually
- ❍ ...slower when post-editing than when translating manually
- ❍ ...just as fast when post-editing as when translating manually.

Q29 You may provide additional information to the above question here.

Q30 I think the quality of my translations is...

- ❍ ...better when post-editing than when translating manually.
- ❍ ...worse when post-editing than when translating manually.
- ❍ ... just as good when post-editing as when translating manually.

Q31 You may provide additional information to the above question here.

Q32 I think the quality of my translations would be...

- ❍ ...better when post-editing than when translating manually.
- ❍ ...worse when post-editing than when translating manually.
- ❍ ... just as good when post-editing as when translating manually.

Q33 You may provide additional information to the above question here.

Appendix 5: Participant survey after experiment

Q1 Thank you for participating in the experiment! This survey is about your personal experience. There are no right or wrong answers. Here as well, your answers will be processed anonymously.

Q13 Name:

Q3 For this text type, which translation method do you prefer?
- ❍ manual translation
- ❍ post-editing
- ❍ no preference

Q4 Why did you prefer this method?

Q5 Which translation method allowed you to translate the fastest?
- ❍ manual translation
- ❍ post-editing
- ❍ both methods were equally fast

Q6 Why was this method faster?

Q8 Which translation method did you find least tiring?
- ❍ manual translation
- ❍ post-editing
- ❍ both methods were equally tiring

Q9 What made translation or post-editing especially tiring?

Q10 Which suggestions from the machine translation did you find useful?

Q11 What could have been better during post-editing? (specific output problems, usability of the interface, possible irritations)

Q12 Please use this space to add any other remarks you still have about your experience during the experiment (process, manual translation versus post-editing, interface, etc.)

Appendix 6: Summary of Chapter 5 mixed models effects

Fig.	dependent variable	AIC null model	AIC predictor model	predictors	effect	sig.
Figure 27	Time per word (in ms)	26959	26937	method	-965 (± 227)	**< 0.001**
				experience	598 (± 702)	0.4
				method:exp	238 (± 332)	0.47
Figure 28	Average fixation duration (in ms)	12516	12492	method	-5.06 (± 1.24)	**< 0.001**
				experience	7.92 (± 15.63)	0.62
				method:exp	0.2 (± 1.84)	0.91
n/a	Average # fixations ST	8923	8871	method	-1.94 (± 0.36)	**< 0.001**
				experience	-0.66 (± 1.54)	0.67
				method:exp	-0.2 (± 0.53)	0.7
Figure 29	Average fixation duration ST	13292	13280	method	-0.6 (± 1.67)	0.72
				experience	12.34 (± 12.39)	0.33
				method:exp	-6.95 (± 2.48)	**0.005**
Figure 30	Average # fixations TT	9171	9161	method	-0.05 (± 0.4)	0.91
				experience	-0.68 (± 1.18)	0.57
				method:exp	1.77 (± 0.59)	**0.003**
n/a	Average fixation duration TT	13356	13251	method	-11.77 (± 1.64)	**< 0.001**
				experience	7.82 (± 17.1)	0.65
				method:exp	-2.43 (± 2.42)	0.31
n/a	Time spent in external resources per word (in ms)	27470	27474	n/a	n/a	n/a
n/a	Average # external resources per word	-1021	-1024	method	-0.02 (± 0.01)	0.07
				experience	0.03 (± 0.03)	0.36
				method:exp	-0.006 (± 0.02)	0.72
Figure 32	Average # external resources per word	-1021	-1027	method	-0.02 (± 0.01	**0.005**
Figure 34	Average time spent in dictionaries per word (in ms)	22571	22566	method	-14.66 (±41.18)	0.72
				experience	243.27(±109.81)	**0.036**
				method:exp	-109.46(60.66)	0.07
n/a	Average total error weight	-2806	-2804	n/a	n/a	n/a
Figure 37	Average adequacy error weight	-3579	-3582	experience	0.013(±0.006)	**0.035**