

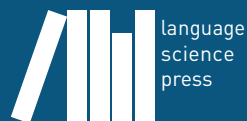
# Translation, interpreting, cognition

The way out of the box

Edited by

Tra&Co Group

Translation and Multilingual Natural  
Language Processing 15



## Translation and Multilingual Natural Language Processing

Editors: Oliver Czulo (Universität Leipzig), Silvia Hansen-Schirra (Johannes Gutenberg-Universität Mainz), Reinhard Rapp (Hochschule Magdeburg-Stendal), Mario Bisiada (Universität Pompeu Fabra)

In this series:

1. Fantinuoli, Claudio & Federico Zanettin (eds.). *New directions in corpus-based translation studies.*
2. Hansen-Schirra, Silvia & Sambor Gruzca (eds.). *Eyetracking and Applied Linguistics.*
3. Neumann, Stella, Oliver Čulo & Silvia Hansen-Schirra (eds.). *Annotation, exploitation and evaluation of parallel corpora: TC3 I.*
4. Czulo, Oliver & Silvia Hansen-Schirra (eds.). *Crossroads between Contrastive Linguistics, Translation Studies and Machine Translation: TC3 II.*
5. Rehm, Georg, Felix Sasaki, Daniel Stein & Andreas Witt (eds.). *Language technologies for a multilingual Europe: TC3 III.*
6. Menzel, Katrin, Ekaterina Lapshinova-Koltunski & Kerstin Anna Kunz (eds.). *New perspectives on cohesion and coherence: Implications for translation.*
7. Hansen-Schirra, Silvia, Oliver Czulo & Sascha Hofmann (eds.). *Empirical modelling of translation and interpreting.*
8. Svoboda, Tomáš, Lucja Biel & Krzysztof Łoboda (eds.). *Quality aspects in institutional translation.*
9. Fox, Wendy. *Can integrated titles improve the viewing experience? Investigating the impact of subtitling on the reception and enjoyment of film using eye tracking and questionnaire data.*
10. Moran, Steven & Michael Cysouw. *The Unicode cookbook for linguists: Managing writing systems using orthography profiles.*
11. Fantinuoli, Claudio (ed.). *Interpreting and technology.*
12. Nitzke, Jean. *Problem solving activities in post-editing and translation from scratch: A multi-method study.*
13. Vandevoorde, Lore. *Semantic differences in translation.*
14. Bisiada, Mario (ed.). *Empirical studies in translation and discourse.*
15. Tra&Co Group (ed.). *Translation, interpreting, cognition: The way out of the box.*

# Translation, interpreting, cognition

The way out of the box

Edited by

Tra&Co Group



Tra&Co Group (ed.). 2021. *Translation, interpreting, cognition: The way out of the box* (Translation and Multilingual Natural Language Processing 15). Berlin: Language Science Press.

This title can be downloaded at:

<http://langsci-press.org/catalog/book/288>

© 2021, the authors

Published under the Creative Commons Attribution 4.0 Licence (CC BY 4.0):

<http://creativecommons.org/licenses/by/4.0/> 

ISBN: 978-3-96110-304-1 (Digital)

978-3-98554-000-6 (Hardcover)

ISSN: 2364-8899

DOI: 10.5281/zenodo.4544686

Source code available from [www.github.com/langsci/288](http://www.github.com/langsci/288)

Collaborative reading: [paperhive.org/documents/remote?type=langsci&id=288](http://paperhive.org/documents/remote?type=langsci&id=288)

Cover and concept of design: Ulrike Harbort

Typesetting: Felix Kopecky

Proofreading: M. Bisiada, B. Beekhuizen, A. Burchardt, Aniefon D., A. Ghorbanpour, A.C. Gieshoff, O. Czulo, G. de Sutter, C. Fantinuoli, A. Garcia, A. Gros, S. Halverson, S. Hansen-Schirra, A. Hervais-Adelman, K. Hvelplund, P. Jerono, L. Mackenzie, J. Monti, R. Muñoz, M. Myers, J. Nitzke, E. Le Foll, K. Oster, B. Reynolds, A. Rojo, C. Rossi, M. Schaeffer, A. Schiefner, M. Stelling, A. Tardel, E. Tiselius, A. Toral, D. Uštulica, F. Vandervoort, J. van de Weijer

Fonts: Libertinus, Arimo, DejaVu Sans Mono

Typesetting software: Xe<sub>La</sub>T<sub>E</sub>X

Language Science Press

xHain

Grünberger Str. 16

10243 Berlin, Germany

[langsci-press.org](http://langsci-press.org)

Storage and cataloguing done by FU Berlin

Freie Universität  Berlin

# Contents

<b>Preface</b>	
Tra&Co Group	iii
<b>1 Multi-modal estimation of cognitive load in post-editing of machine translation</b>	
Nico Herbig, Santanu Pal, Antonio Krüger & Josef van Genabith	1
<b>2 Comparing NMT and PBSMT for post-editing in-domain formal texts: A case study</b>	
Sergi Álvarez, Toni Badia & Antoni Oliver	33
<b>3 German light verb construction in the course of the development of machine translation</b>	
Shaimaa Marzouk	47
<b>4 Dialogue-oriented evaluation of Microsoft's Skype Translator in the language pair Catalan-German</b>	
Felix Hoberg	67
<b>5 Investigating post-editing: A mixed-methods study with experienced and novice translators in the English-Greek language pair</b>	
Maria Stasimioti & Vilelmini Sosoni	79
<b>6 The processing of website contents in native and non-native language</b>	
Jean Nitzke	105
<b>7 Assessing indicators of cognitive effort in professional translators: A study on language dominance and directionality</b>	
Aline Ferreira, Stefan Th. Gries & John W. Schwieter	115
<b>8 Attention distribution and monitoring during intralingual subtitling</b>	
Anke Tardel, Silvia Hansen-Schirra, Moritz Schaeffer, Silke Gutermuth, Volker Denkel & Miriam Hagmann-Schlatterbeck	145

<b>9</b>	<b>Eye tracking study of reading for translation and English-Russian sight translation</b>	
	Elena Kokanova, Maya Lyutyanskaya & Anna Cherkasova	<b>163</b>
<b>10</b>	<b>Emotion and the social embeddedness of translation in the workplace</b>	
	Hanna Risku & Barbara Meinx	<b>173</b>
	<b>Index</b>	<b>189</b>

# Preface

Tra&Co Group

Johannes Gutenberg University Mainz

After the first successful International Congress on Translation, Interpreting and Cognition held at the University of Mendoza, Argentina in 2017, the second conference in this series has been hosted by the Tra&Co Center at the Johannes Gutenberg University of Mainz in Gernersheim, Germany in 2019. The predictive and explanatory power of studies investigating the translation process has been recognised by many in the field, so it is not surprising that cognitive aspects of the translation process have become central in many research endeavours in Translation and Interpreting Studies in recent years. Interdisciplinary paradigms have been useful in the field for a long time, but even more so in Cognitive Translation and Interpreting Studies. Interdisciplinarity has been useful in overcoming the limits of single disciplines, but also to shed light on hitherto hidden phenomena. The aim of the second International Congress on Translation, Interpreting and Cognition was therefore to call for interdisciplinary multi-method approaches. There were contributions on a range of topics which were held together by a central thread, i.e., by the study of cognitive aspects of translation and interpreting. In particular, there were studies which observed behaviour during translation and interpreting – with a focus on training of future professionals, on language processing more generally and language dominance in particular, on the role of working memory during simultaneous interpreting. In addition, there were studies on how to measure translation competence, on the role of technology in the practice of translation, on interpreting and subtitling, on translation of multimodal media texts, on aspects of ergonomics and usability, on emotions and the role they play in the translation process, on translators' self-concept and psychological factors, on writing in a foreign language and finally also on revision and post-editing. For the present publication, we selected a number of contributions, which showcase the breadth and depth of studies that have been presented at the conference. We are grateful for the general support from the



Tra&Co Group. 2021. Preface. In Tra&Co Group (ed.), *Translation, interpreting, cognition: The way out of the box*, iii–iv. Berlin: Language Science Press.

DOI: 10.5281/zenodo.4545027 

*Tra&Co Group*

Gutenberg Research College (GRC), the Freundeskreis FTSK, and the JGU Internal University Research Funding. Finally, we are extremely grateful for the many reviewers and their valuable suggestions and feedback to the individual contributions of the present proceedings.

Tra&Co Group

Germersheim, September 2020

(Silvia Hansen-Schirra, Anne-Kathrin Gros, Silke Gutermuth, Ann-Kathrin Habig, Jean Nitzke, Katharina Oster, Moritz Schaeffer, Anke Tardel, and Janna Verthein)



# Chapter 1

## Multi-modal estimation of cognitive load in post-editing of machine translation

Nico Herbig<sup>a,b,c</sup>, Santanu Pal<sup>a,b,c</sup>, Antonio Krüger<sup>a,b,c</sup> & Josef van Genabith<sup>a,b,c</sup>

<sup>a</sup>German Research Center for Artificial Intelligence (DFKI) <sup>b</sup>Saarland University  
<sup>c</sup>Saarland Informatics Campus

In this paper, we analyze a wide range of physiological, behavioral, performance, and subjective measures to estimate cognitive load (CL) during post-editing (PE) of machine translated (MT) text. To the best of our knowledge, the analyzed feature set comprises the most diverse set of features from a variety of modalities that has been investigated in the translation domain to date. Our focus lies on predicting the subjectively reported perceived CL based on the other measures, which could for example be used to better capture the usefulness of MT proposals for PE, including the mental effort required, or to develop cognition-aware translation environments that support human translators according to their current level of CL. Based on the data gathered from 10 professional translators, we show that feature sets from all different modalities outperform our baseline measures in terms of predicting the subjectively perceived level of CL, and that especially eye-, heart-, or skin-based features yield good results in a simple “top-down” regression analysis using feature selection. When passing the participant and segment to the regression models, other modalities like keyboard, text, body posture, or time, also perform well. An additional correlation analysis provides insights into redundancies among the features which may be used to further improve the currently achieved best regression score of 0.7 mean squared error (MSE) on a 9-point scale.



## **1 Introduction**

Even though machine translation (MT) systems are improving rapidly, the resulting translations currently still require manual post-editing (PE) to capture and correct errors and make the target texts conform to their intended objective. PE has the potential of inducing high cognitive load (CL) on the translator: it involves continuous scanning of texts, including source, the incrementally evolving final translation output and possible error-prone MT output for mistakes, (sub-)strings that can be reused, text that has already been translated, text that still needs to be translated, etc. When PE is required, we should therefore optimize for a low perceived CL during PE, and not only focus on MT quality in terms of automatic measures or time to post-edit. Here, we see CL as “a variable that attempts to quantify the extent of demands placed by a task on the mental resources we have at our disposal” (Chen et al. 2016).

While CL and MT quality are interrelated, they cannot be considered equal: for example, repeated mistakes that have been corrected by the translator again and again in the past may impact perceived CL, while the MT quality remains the same. Therefore, it has been argued that CL is a more decisive indicator of the overall effort expended by post-editors (Vieira 2016).

To investigate how computer-aided translation (CAT) tools could adapt when high cognitive loads are detected, Herbig, Pal, van Genabith, et al. (2019) interviewed professional translators. The most proposed and most liked idea was to provide alternative translations from MT, translation memories (TM), or a corpus; however, other adaptations like automatic proposals to encourage the translator to take a break, reordering segments to switch between highly and less demanding segments, user interface adaptations, or payment based on induced CL were also discussed.

Apart from these CAT adaptations based on CL, the automatic capture of CL without interfering in the PE process would further enable the creation of large datasets of CL scores for (source, MT, PE) tuples that could be used to optimize MT systems to produce output inducing lower CL on the post-editors.

To provide some first steps towards these goals, we are concerned with the question of how to actually estimate CL during PE. For this, (1) we present an approach based on a wide range of physiological, behavioral, performance, and subjective measures, yielding the so far most diverse set of features from a variety of modalities that has been investigated in the translation domain. (2) We analyze how well predictive models based on feature combinations from these modalities can predict perceived CL, as measured by subjective ratings on a well

established CL scale from psychology (Paas & van Merriënboer 1994). The different modalities and their combinations are then compared in terms of regression performance. (3) Similar to Vieira (2016), we investigate pairwise correlations between different interesting indicators of CL and also subjectively assessed CL and run a principal component analysis (PCA) to figure out which features capture similar or distinct underlying concepts. This step aims to help us understand the relation between the different CL estimators.

The results of our analyses indicate that heart, eye and skin, as well as combined measures perform very well, while text, keyboard, body posture, or time features only perform well when considering the individual participant and segment s/he is editing. Overall, the best predictive model achieved a regression score of 0.7 mean squared error (MSE) on a 9-point scale. However, the correlation analysis shows that our “top-down” regression approach, which uses a simple feature selection algorithm, sometimes chooses redundant features, suggesting that it might be possible to improve results by analyzing the features in more depth and combining them in a more sophisticated way.

## 2 Related work

This section discusses related studies by first giving an overview of CL measures and then presenting studies on measuring CL during translation.

### 2.1 Overview of cognitive load measures

Cognitive load theory (Paas & van Merriënboer 1994; Sweller et al. 1998) has been developed in psychology and is concerned with an efficient use of people’s limited cognitive resources to apply acquired knowledge and skills to new situations (Paas et al. 2003). Approaches to detect CL can be roughly divided into four categories: subjective measures, performance measures, behavioral measures, and physiological measures.

**SUBJECTIVE MEASURES** are based on the assumption that subjects can self-assess and report their cognitive processes after performing a task (Paas & van Merriënboer 1994). Several scales exist, and introspection is often used as a ground truth to evaluate how well CL can be assessed by other means, such as physiological measurements.

**PERFORMANCE MEASURES** such as the time required or the text quality achieved assume that when working memory capacity is overloaded, a performance drop occurs due to the increase in overall CL (Chen et al. 2016). However, by increasing their efforts, humans can compensate for the overload and maintain their

performance over a period of time, although this can lead to additional strain and fatigue (Hockey 1997).

BEHAVIORAL MEASURES can be extracted from user activity while performing a task. Especially interesting in the context of PE are mouse and keyboard input-based features, which were shown to correlate to CL (Arshad et al. 2013).

Last, a lot of research has been done on PHYSIOLOGICAL MEASUREMENTS, which assume that human cognitive processes can be observed in the human physiology (Kramer 1991). Eye-tracking is frequently used for physiological CL measurements: the pupil diameter increases with higher CL (Iqbal et al. 2004; O'Brien 2006a), the frequency of rapid dilations changes (Demberg & Sayeed 2016), and the blink behavior adapts (Van Orden et al. 2001). Furthermore, Chen & Epps (2013) as well as Stuyven et al. (2000) showed that fixations and saccades can also be used for CL predictions. Apart from the eyes, the skin also provides information about the user's cognitive state: galvanic skin response (GSR) can be used to determine whether a user feels stressed (Villarejo et al. 2012) and provides information about the CL (Shi et al. 2007). Remote measurements of the skin temperature have also been effective (Yamakoshi et al. 2008). Further commonly used indicators rely on the cardiovascular system: blood pressure (Yamakoshi et al. 2008), heart rate (Mulder 1992), and especially heart rate variability (HRV; Rowe et al. 1998) have been shown to correlate with CL. In addition, features such as the head pose also correlate to CL when learning (Asteriadis et al. 2009).

## **2.2 Cognitive load estimation in the translation domain**

Due to the parallel activation of two languages, reading for translation imposes more demand on the working memory than reading within a single language (Macizo & Bajo 2006), thus, making CL estimation particularly interesting in the translation domain. Therefore, a few, albeit seminal, publications relevant to the cognitive dimension of modeling PE have been presented:

Krings (2001) utilized think-aloud protocols to capture cognitive effort; however, as pointed out by O'Brien (2005), post-editors constantly reporting what they are doing (a) slows down the process and (b) changes the process itself.

O'Brien (2005) explored correlating pauses in typing behavior to potentially difficult source text features. In a follow-up analysis (O'Brien 2006b), she concluded that "while pauses provide some indication of cognitive processing, supplementary methods are required". Lacruz et al. (2012) and Lacruz & Shreve (2014) built upon this work, but instead of examining long pauses, they analyzed clusters of shorter pauses. Their metrics called average pause ratio (APR) and pause to word ratio (PWR) could be correlated to technical effort (the required mouse

and keyboard actions), arguing that “it is likely that in many situations technical effort and cognitive effort will be related”. Pause ratios were also shown to be more sensitive to grammatical, word order, or structure errors. For TMs, Mellinger (2014) was able to correlate keystroke logs and pause metrics to translation quality ratings. Last, the total pause duration was found to be smaller when post-editing than during manual translation of metaphors (Koglin 2015); however, this could be explained by the large time savings achieved through PE.

While pauses and technical effort relate to these MT quality measures, which are in turn related to perceived CL, CL and MT quality cannot be considered equal: consider very bad MT proposals that are still very easy to PE due to the simplicity of the segments or the contrary situation, a very high MT quality where spotting the error can remain difficult and induce a high CL. We will nevertheless integrate pause measures, as they are very easily applicable in TPR studies, but compare them to physiological and subjective measures of CL.

Among the physiological measures, eye-tracking has frequently been used as a means to capture CL during PE: O’Brien (2006a) proposed pupil dilation as a measure of CL and focused on correlations with different match types retrieved from a TM. Doherty et al. (2010) also explored eye-tracking by measuring different features while reading MT output. They found that gaze time and fixation count correlate with MT quality; however, fixation duration and pupil dilation were less reliable. Carl et al. (2011) found more fixations and longer gaze times on the target text when comparing PE to manual translation. Therefore, the authors argue that there is more effort in correcting MT outputs, whereas manual translation requires more effort for reading and understanding the source. This finding was also replicated by Koglin (2015). Moorkens et al. (2015) correlated ratings of expected PE effort with temporal, technical and cognitive effort, in terms of time, translation error rate (TER; Snover et al. 2006; 2009), and fixation counts and durations, respectively. Interestingly, the correlations between eye-tracking data and predicted effort were either very weak or weak, suggesting that human predictions of PE effort cannot be considered completely reliable. Furthermore, Daems (2016) found that fixations are mostly impacted by coherence and other meaning shifts. In contrast to these quality-, time-, and expectation-based measures, Vieira (2014) uses a psychology-motivated definition of CL. He linked average fixation duration, fixation counts, and a self-report scale measuring CL, which is frequently used in psychology (Paas & van Merriënboer 1994) to segments expected to pose different levels of translation difficulty and their corresponding Meteor (Lavie & Agarwal 2007) ratings.

As can be seen, a variety of approaches already exists linking different eye features to effort metrics, ranging from simply counting fixations on the source

and target to pupil diameter measures. However, the focus was again mostly on a link to translation quality, sentence features, or expected effort, with only one consideration of CL in the psychological sense. Furthermore, the works only investigated eye tracking, without considering other physiological or behavioral measures.

In contrast, the follow-up work by Vieira (2016) analyzes how all of the above measures, as well as pause metrics and editing time, relate to each other in a multivariate analysis. He found correlations between all measures; however, a PCA showed that they cluster in different ways. The work most related to this study is our previous study – Herbig, Pal, Vela, et al. (2019) – with translation master’s students, where we explored a vast variety of CL measures, including eye, skin, heart, and typing features that were previously unexplored in the translation domain, analyzed correlations, and investigated how well these can be used to predict the subjective CL ratings.

In this work, we built upon our previous findings (1) by conducting a similar study with professional translators instead of translation master’s students, (2) by incorporating even more sensors and features in the system, and (3) by not only analyzing predictive models of subjective CL or correlations to this subjective measure, but further by performing the multivariate analysis of Vieira (2016) to understand how the different measures relate to each other and how the features cluster together.

### **3 Method**

As stated earlier, we believe that the CL perceived by translators during PE should be considered more closely, since MT output often requires PE, and considering only the number of changes needed may not provide an accurate measure of the effort involved (Koponen 2016). Adding this CL-based perspective on PE of MT to the commonly used but oversimplifying BLEU (Papineni et al. 2002) perspective on MT quality should lead to a better approximation of actual PE cost.

To test which measuring approaches can actually reflect different levels of CL in PE, we perform a user study<sup>1</sup> to gather data from a variety of sensors, which can be combined in a multi-modal fashion. For the analysis, we conduct a hybrid of the approaches by Herbig, Pal, Vela, et al. (2019) and Vieira (2016). That is, we aim to predict subjectively assessed CL based on the captured multi-modal sensor data by training regression models and we further perform a multivariate analysis and a PCA to find pairwise correlations and clusters of different features.

---

<sup>1</sup>The study was approved by the university’s ethical review board.

The goal of the regression analysis is to automatically infer the CL from the raw sensor data, ideally using as few and as commonly used sensors as possible. The multivariate analysis should then provide more detailed insights into why some measuring approaches perform well while others contribute little.

### 3.1 Analyzed measures of cognitive load

Compared to Vieira (2016), Herbig, Pal, Vela, et al. (2019) already increased the amount of analyzed features significantly by adding heart-, skin-, and camera-based features. In this work, we add even more and higher quality sensors and add further high-level features.

#### 3.1.1 Subjective measures

Subjective measures are based on the assumption that subjects can self-assess and report their cognitive processes after performing a task. For this, we adapted a CAT tool to ask for a subjective CL rating (SubjCL) using the scale proposed by Paas & van Merriënboer (1994) after every single segment. This scale was chosen because it focuses on CL and not on quality, and further since it was used in the two most related studies by Vieira (2016) and Herbig, Pal, Vela, et al. (2019). The single 9-point question is “In solving or studying the preceding problem I invested” with a choice of answers ranging from “very, very low mental effort” to “very, very high mental effort”.

#### 3.1.2 Performance measures: Text and time

The usual performance measures based on the required time or achieved quality are not as easily accessible in PE as in other cognitive tasks, since it is possible to trade of quality for time and because translation quality is a partly subjective measure. Nevertheless, we integrate the following simple time and text measures:

For the `TIME FEATURES` we integrate PE time (PeTime) and length-normalized PE time which also considers the segment length (LNPeTime).

The `TEXT FEATURES` consist of smoothed BLEU, HBLEU (Lin & Och 2004), TER, HTER (Snover et al. 2009), and sentence length (SL). Note that the difference between the non-H- and H-based measures lies in the choice of the reference translation and hypothesis: BLEU and TER take the MT output as hypothesis and the independently provided human translation as reference and calculate  $n$ -gram overlap (BLEU) or the amount of necessary edits (TER) to transform the hypothesis into the reference, while HBLEU and HTER perform the same calculations, but this time between the MT output and the post-edited translation.

### **3.1.3 Behavioral measures: Keyboard typing and body posture**

Behavioral measures can be extracted from user activity while performing a task. Especially interesting in the context of PE, where the translator does not move a lot, is focused on the screen, does not speak, etc., are `MOUSE AND KEYBOARD INPUT-BASED FEATURES`. Therefore, our most basic sensor is a key logger storing all keyboard and mouse input during PE. The higher-level pause features `APR` and `PWR` by Lacruz et al. (2012), which were shown to correlate with PE effort, are automatically calculated from the keyboard events.

Furthermore, the `BODY POSTURE` is captured by a Microsoft Kinect v2. We hypothesize that post-editors come closer to the screen for hard-to-edit translations, so we calculate the distance to the head and normalize it per participant (`HeadDist`).

### **3.1.4 Physiological measures: Eyes, heart, and skin**

As physiological measurements, we integrate eye-, heart-, and skin-based measures in our experiment.

For `EYE-BASED FEATURES`, we use a web-cam and an eye tracker. The web-cam, which is naturally not as precise as the eye tracker but easily accessible on most modern devices, is used to calculate the eye aspect ratio (`EAR`), which indicates the openness of the lids (Soukupova & Cech 2016). The remote Tobii eye tracker 4C with the Pro SDK records the raw gaze data. Based on this raw data, we calculate the amount of blinking (of less than 2 s length; `BlinkAmount`) and also normalize this by the PE time (`NormBlinkAmount`) (Van Orden et al. 2001). Similarly, we calculate the number of fixations (`FixAmount`) and normalize it by PE time (`NormFixAmount`). We further compute the fixation durations (`FixDur`) and saccade durations (`SaccDur`) (Doherty et al. 2010; Moorkens et al. 2015), all of which have been shown to be indicators of CL. Furthermore, we reimplemented the work by Goldberg & Kotval (1999) to calculate the probability of visual search based on the eye movements (`SearchProb`), which was proposed to determine whether a user is searching within a user interface and could therefore also be an indication of a user feeling “lost” while PE. Last, and as the main distinction from Herbig, Pal, Vela, et al. (2019), we also capture the pupil diameter (`PupilDiameter`, O’Brien 2006a). For calculating higher-level features on the sensor output, we first replace blinks from the signal by linear interpolation. Then, the index of cognitive activity (`ICA`), which is the frequency of small rapid dilations of the pupil (Demberg & Sayeed 2016) that was shown to be more robust to changes in illumination, is calculated based on this signal. Two approaches are



implemented: one uses a wavelet transformation to calculate the number of rapid dilations (ICA<sup>wave</sup>), while the other simply counts how often a sample deviates by more than 5 times the rolling standard deviation from the rolling mean of the signal (ICA<sup>count</sup>). Last, we also implemented the work of Hossain & Yeasin (2014), which checks for sharp changes and continuations of the ramp in the Hilbert unwrapped phase of the pupil diameter signal (Hilbert).

For HEART MEASURES, we integrate three devices: a Polar H7 heart belt, a Garmin Forerunner 935 sports watch, and the Empatica E4 wristband. That way, we have two sports devices (Polar and Garmin) and one CE certified medical device (type 2a) offering an early glimpse of the data quality achieved by future consumer devices. From both the Polar belt and the Garmin watch, we capture the heart rate (HR).

The Polar belt, as well as the Empatica wristband, further capture the RR interval (RR), which is the length between two successive Rs (basically the peaks) in the ECG signal. Based on this, we calculate the often-used CL measures of heart rate variability (HRV, Rowe et al. 1998), in particular the root mean square of successive RR interval differences (RMSSD) and the standard deviation of NN intervals (SDNN). Here, the SDNN uses NN intervals, which normalize across the RR intervals and thereby smooth abnormal values. Furthermore, we add the HRV features NN50 and pNN50, which are the number and percentage of successive NN intervals that differ by more than 50 ms (Shaffer & Ginsberg 2017), for both the Empatica and the Polar to the analysis.

Furthermore, the Empatica measures the blood volume pulse (BVP), which is the change in volume of blood measured over time. Based on it, we calculate the BVP amplitude (BVPamp, Iani et al. 2004), which contains the amplitude between the lowest (diastolic point) and highest (systolic point) peak in a one second interval. Last, we also calculate the median absolute deviation (BVPMedAbsDev) and the mean absolute difference (BVPMeanAbsDiff) among the BVP values (Haapalainen et al. 2010). Here, BVPMedAbsDev is the median of the absolute differences between individual measurements and the median of all measurements. BVPMeanAbsDiff is simply the mean of absolute differences of each pair of measurements. Both these features are calculated per interval of 125 ms.

The main difference compared to Herbig, Pal, Vela, et al. (2019) regarding heart features is that we additionally included the Garmin and Empatica devices, which allowed us to also integrate BVP-related measures. Furthermore, we extended the set of considered HRV measures to also include NN50 and pNN50.

For SKIN-BASED FEATURES, we integrate the Microsoft Band v2 and again use the Empatica and the Garmin devices. The MSBand and Empatica both measure the commonly used galvanic skin response (GSR) which is an indicator of CL.

We also transform this signal to the frequency domain (FreqGSR) as described in Chen et al. (2016). In accord with their work, we also calculate data frames of length 16, 32, and 64 samples, which are similarly transformed to the frequency domain and normalized by the participant average (FreqFrameGSR).

Furthermore, we use the Ledalab software<sup>2</sup> to calculate higher level skin conductance features on the Empatica raw data. It provides us with “global” features, namely the mean value (Leda<sub>avg</sub>) and the maximum positive deflection (Leda<sub>MaxDefl</sub>), and “through-to-peak (TTP)/min-max” analysis, namely the number of significant (i.e. above-threshold) skin conductance responses (SCRs) (Leda<sub>TTP.nSCR</sub>), the sum of SCR amplitudes (Leda<sub>TTP.AmpSum</sub>) of significant SCRs, and the response latency (Leda<sub>TTP.Lat</sub>) of the first significant SCR. Furthermore, and most interestingly, we use Ledalab to perform a continuous decomposition analysis (CDA, Benedek & Kaernbach 2010), which separates skin conductance data into continuous signals of tonic (background) and phasic (rapid) activity. The features based on this CDA analysis again include the number of significant SCRs, the SCR amplitudes of significant SCRs, and the latency of the first SCR (Leda<sub>CDA.nSCR</sub>, Leda<sub>CDA.AmpSum</sub>, Leda<sub>CDA.Lat</sub>). Furthermore, the average phasic driver (Leda<sub>CDA.SCR</sub>), the area of phasic driver (Leda<sub>CDA.ISCR</sub>), as well as the maximum value of phasic activity (Leda<sub>CDA.PhasMax</sub>) and the mean tonic activity (Leda<sub>CDA.Ton</sub>) features are created by the Ledalab software.

The Empatica and Garmin devices also measure the skin temperature, which we use as a feature (SkinTemp).

The differences from Herbig, Pal, Vela, et al. (2019) for the skin features are as follows: we further use the skin resistance data delivered by the Empatica E4, on which we calculate the same features as in their work, but additionally add the Ledalab features. Furthermore, we integrate the skin temperature features.

### 3.1.5 Data normalization and segment-wise feature calculation

The features described above can be categorized into two classes: *global features* and *continuous features*.

By GLOBAL FEATURES we mean features that yield only one value per segment: this class comprises subjective measures (SubjCL), time measures (PeTime, LNPeTime), text measures (BLEU, HBLEU, TER, HTER, SL), keyboard measures (APR, PWR), the amount-based eye features (BlinkAmount, FixAmount, NormBlinkAmount, NormFixAmount), and all Ledalab skin features. However, one should note that the time and text features here really only can be calculated on the whole segment, while the amount-based eye features or the skin-based Ledalab features could also be calculated over shorter periods of time.

---

<sup>2</sup><http://www.ledalab.de/>

Apart from these global features, all other features are basically just a CONTINUOUS SIGNAL (of different sampling rates) that we still need to transform to a directly usable set of values per segment: Each signal is first normalized as described in Chen et al. (2016) by dividing it by the participant’s mean value. Then 6 very simple features are calculated from this normalized signal: the accumulated, average, standard deviation, minimum, maximum, and range (max – min). As an example, this means that GSR, actually consists of the 6 features  $GSR_{acc}$ ,  $GSR_{avg}$ ,  $GSR_{std}$ ,  $GSR_{min}$ ,  $GSR_{max}$ , and  $GSR_{range}$ .

We manually inspected the data distribution per segment and participant for outliers and overall data quality. First of all, the Empatica E4 sensor, which claims clinical quality observations, indeed shows the fewest outliers and nicely bell shaped data distributions. In contrast, the Polar H7 sports sensor and the Microsoft Band v2 showed much more noisy data. Therefore, we filtered values according to visual inspection and related literature: data above 100,000 k $\Omega$  for the raw Microsoft Band GSR was removed. Furthermore, Polar RMSSD and SDNN values above 1000 (van den Berg et al. 2018) as well as  $HR^{Polar}$  and  $RR^{Polar}$  samples which fall outside the acceptable 50–120 beats per minute or 500–1200 ms ranges were ignored (Shaffer & Ginsberg 2017).

### 3.2 Text and apparatus used for the experiment

Apart from the sensors, we need to generate translations for our experiments that contain realistic error types. For this, we use the same 30 sentences as Herbig, Pal, Vela, et al. (2019), which are chosen as follows: A neural MT system (Gehring et al. 2017) was trained on the English-German parallel data from the WMT 2017 news translation task and provided translation candidates on the respective test data set. Then 30 sentences were chosen from this test set by (a) using sentences of different TER intervals, (b) reducing the number of possible candidates based on manual error analysis, and (c) further shrinking the set based on subjective CL ratings from two translation master’s students in a pre-study. For details regarding the selection of sentences please refer to Herbig, Pal, Vela, et al. 2019. All participants used these same 30 segments; however, the order is randomized to avoid ordering effects.

For the study, the post-editor is equipped with a Microsoft Band v2 on her right wrist, the Garmin Forerunner 935 and Empatica E4 on the left wrist (the Garmin is further up), the heart belt on her chest, and an eye tracker, as well a web-cam and a Microsoft Kinect v2 camera facing her. As input possibilities, a standard keyboard and mouse are attached, and a 24-inch monitor displays the translation environment. We chose SDL Trados Studio 2017 for this study as it is by far the most used CAT tool in professional applications.

### 3.3 Data analysis approach

First, we analyze the subjective ratings provided by our participants. Then, similar to Herbig, Pal, Vela, et al. (2019), we estimate the subjective ratings of perceived CL based on a combination of different features. Last, we use the approach by Vieira (2016) and investigate correlations between our measures to understand how they relate to each other.

For all analyses, we discuss the features in terms of the feature sets described in Section 3.1: *subjective, time, text, keyboard, body posture, heart, eye, and skin* features. Finally, we also investigate *combinations* of these sets.

#### 3.3.1 Subjective ratings

We start by reporting and analyzing the subjective ratings provided by our participants. As this is our target measure, it is important to understand the distribution of our dataset as well as inter-rater differences.

#### 3.3.2 Multi-modal CL regression analysis

The goal of this stage is to investigate the feasibility of automatically gathering CL values for segments through different sensors. For this, we learn a function that fits our features to the subjective CL as reported by each participant on the rating scale after each segment; thus, the output space is 1 to 9. We consider each segment of each participant an individual sample with the corresponding subjective rating as a label. Please note that neither a manual annotation of the segments nor an average CL rating across participants is used here.

The reason why we focus on subjectively assessed CL is that it is good at capturing inter-translator differences. This is important because the task difficulty by itself is of a subjective nature, as it depends on the translator’s experience with similar texts, vocabulary, etc. Thus, we also do not normalize our target variable, because the lowest rating assigned by one participant is not necessarily comparable to the lowest rating assigned by another participant due to prior experience, which in turn could also result in different physiological responses. Thus, instead of potentially biasing our data by transforming the target variable, we keep it as is and perform a comparison between models with a random effect for participant and those without such knowledge, as described in further detail below. Apart from subjectively assessed CL we could also have chosen quality or time measures as the target, however, as discussed above, quality and CL cannot be considered equal, and time could be traded off for quality, thereby limiting findings based solely on these measures.

We compare the different regression models based on different feature sets against each other, but also compare each model to a very simple baseline: always predicting the mean subjective rating ( $\text{SubjCL}_{\text{avg}}$ ).

Overall, we compare two approaches for training regression models.

The first approach uses only the above measures to predict  $\text{SubjCL}$ , and has no knowledge about which participant the data comes from or which segment was post-edited while recording the data. Thus, it is a very generic approach that learns one set of parameters across all participants, thereby exploring the feasibility of applying CL adaptations during PE in practice, e.g. for automatically providing alternative proposals when loaded. Since different features and their combinations require different types of functions to best approximate them locally, we train not only one, but several regression algorithms making different assumptions about the underlying function space: linear models with different regularizers, namely a stochastic gradient descent regressor (SGD), a lasso model (Lasso), an elastic net (ENet), and a ridge regressor (Ridge), as well as a non-linear random forest regressor (RF), all provided in the `scikit-learn` library using the default parameters and feature normalization. This analysis is very similar to Herbig, Pal, Vela, et al. (2019), except that our previous analysis additionally used a support vector regression (SVR) model.<sup>3</sup>

As a second approach, which is an extension to the first approach, we further integrate linear mixed-effect models (LMEMs) using R (version 3.6.0, `lme4` package version 1.1-21), as these can effectively capture inter-participant as well as segment-dependent differences by adding a random effect for subject and a random effect for item.<sup>4</sup> To make the comparison between LMEMs and the other models fair, we also provide the `scikit` models with the participant and segment ID; thus, all models can learn to act differently depending on this information. While the normalization of the signal discussed above already normalizes the data such that each participant's average heart rate is at value 1, some participants might still react more strongly to CL, e.g. one participant might increase his heart rate by 10%, while another's might increase by 20%. By incorporating the participant and segment as a feature into the models, we ensure that they can learn such individual difference. This is also a major distinction from Herbig, Pal, Vela, et al. (2019), who did not incorporate these measures. However,

---

<sup>3</sup>Since SVR does not support our selected feature selection approach, and since it never performed best in tests without feature selection, we decided to not use it for this experiment.

<sup>4</sup>Since the R package used for LMEMs does not support our feature selection approach either, we decided to instead perform feature selection with a normal linear regression model with L2 regularization.

this approach of training the models is only relevant for strictly controlled experiments, because in practice no two translators will PE the same segment.

By training multiple regression models, we obtain locally optimal results before comparing them and drawing conclusions on the usefulness of the features involved. That way, our results are not biased or distorted by the use and limitations of a single classifier (and with it the class of functions that can be learned). While we do not fine-tune hyper-parameters of the models and might therefore miss some ideal hyper-parameter combination, our approach offers a reasonably wide range of function spaces to choose from.

To avoid over-fitting, all regression functions use regularization or averaging, and we perform cross-validation (CV). Before passing a feature to a regression model, we apply a z-transformation to achieve 0 mean and unit variance. For combining individual features within a modality or across modalities, we then use simple vector concatenation. As a feature selection approach we use recursive feature elimination with CV (RFECV in `scikit-learn`) to decide on how many and which features to select.

For all of these feature combinations, we train each of the above regressors using a 10-fold stratified CV, which is better suited for an imbalanced distribution of the target variable (that we happen to have, see Section 4.1). We further perform a 5 by 2-fold stratified CV which we use to statistically compare the different models. This method has been suggested by Dietterich (1998) as it ensures that each sample only occurs in the train or test dataset for each estimation of model skill, thereby reducing inter-dependencies. Naturally, every regression model is trained on the same folds, to make results comparable. For each regressor, the average test MSE is computed across the 10 folds and is then compared across regressors as it is a good measure for our actual goal: predicting the subjective CL as well as possible. We choose the MSE as the main metric, since the error squaring strongly penalizes large errors, which are particularly undesirable for our goal.

### **3.3.3 Pairwise correlations and PCA**

Vieira (2016) argues that “using a large number of different measures in the hope that together they will provide a more accurate parameter might be an inefficient approach”, especially when the measures are correlated. Our above approach uses a well established feature selection mechanism to select a good feature subset and thereby automatically reduces redundancies and removes inconclusive features. However, this “top-down” experimental approach still does not provide

any insight into how all the different features correlate and which features reflect the same underlying construct.

To target these shortcomings, Vieira (2016) inspects a correlation matrix visualizing pairwise feature correlations. To further investigate why some measures seem to be more related to each other than others, suggesting that there is also a great degree of redundancy involved, he then used a PCA. As Vieira (2016) nicely puts it, “informally, PCA transforms a group of variables into a group of orthogonal principal components (PC) containing linear combinations of the original variables”. Usually a small number of PCs is enough to explain most of the original data, which is especially important for our data consisting of a huge amount of features.

To keep the reporting concise, we only report PCs that together explain 95% of the variance. Since we have many more features than Vieira (2016), a plot including all features would become very messy and unreadable. Therefore, we create a separate plot per modality to investigate within-modality correlations and further report an across-modality plot. For modalities with more than 5 features, we reduce this set based on the MSE a regressor that was trained solely on each single feature would achieve in a 5 by 2-fold CV. While this does not give us a full picture, it remains interpretable and provides interesting insights.

### 3.4 Participants and user evaluation procedure

The experiment participants were 10 professional translators (8 female), aged 28–62 (mean = 40.4, SD = 9.7). Half of them were freelance translators, while the other half worked for a translation company. All of them were native Germans and had studied translation from English. Their professional experience ranged from 3 to 30 years (mean = 12.1, SD = 3). All of them have worked with Trados SDL Studio, which is the CAT tool we also used for our experiment. However, on average they have used 4.4 distinct CAT tools (SD = 2.1, min = 1, max = 9). On a 5-point scale ranging from very bad to very good, they judged their knowledge of CAT tools as good (mean = 4.2, SD = 0.9), their experience with Trados as good (mean = 4.4, SD = 0.7), their general knowledge of translation as very good (mean = 4.8, SD = 0.4), and their PE knowledge as good (mean = 3.8, SD = 1.0).

After signing a data protection form and filling out the above demographics questionnaire, they were given written instructions explaining that they should (1) post-edit the proposed translations and not translate from scratch, and (2) focus on grammatical and semantic correctness while avoiding stylistic changes. Concrete time limits were not stated. The reason for clear instructions was to ensure a similar PE process across participants; other specifications would also

have been valid for such an experiment. We further allowed but did not require participants to look up terms in a corpus or dictionary online. Before starting the actual PE process, they were given time to familiarize themselves with the environment, e.g. to adjust the chair and adapt the Trados view settings. They then each post-edited the 30 text segments described above in random order while wearing all the sensors. For one participant the USB hub we used broke after post-editing 9 segments, thereby reducing the gathered amount of data for this participant.

## 4 Results and discussion

In this section, we present and discuss the results of each individual step of our data analysis.

### 4.1 Subjective ratings

All 9 CL ratings were used during the experiment; however, 90.3% of the ratings were within the range 3 to 7 (inclusive) while the extreme cases were only rarely chosen (see Figure 1.1). We also observe rating differences between post-editors, with an average standard deviation across segments of 1.2 on our 9-point scale. In general, the rating distribution and the inter-rater differences are strongly comparable to the results of Herbig, Pal, Vela, et al. (2019). As argued in this work, a reason for the non-uniform, rather normal rating distribution could be the strong wording of the used rating scale (Paas & van Merriënboer 1994): “very, very high/low mental effort” is something that we believe users simply do not identify themselves with often.

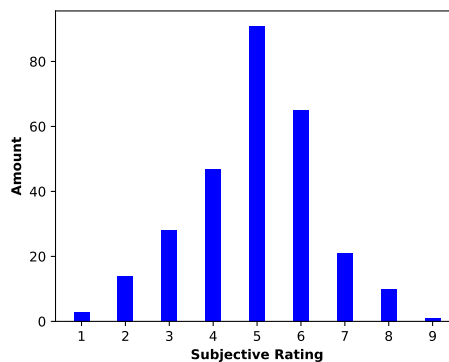


Figure 1.1: Rating distribution across subjective CL scale.



Note that we use these individual CL ratings (without any aggregation on segment level) for the remaining analyses to also capture inter-participant differences. Inspecting the data in further detail, we find 80 out of 151 cases where multiple participants rated the same segment as equally tough while having an editing difference of more than 20 HTER. This supports our above argument that strong differences in editing behavior do not necessarily impact the CL.

## 4.2 Multi-modal CL regression analysis

The results of the first regression analysis approach, that is *without passing the participant and segment* alongside the features to the model, are reported in Table 1.1. It shows the MSE achieved in 1 by 10- and 5 by 2-fold CV, once for the baseline, and further for each category of features described above. For each feature category, we report the results achieved by a model trained on all features (ALL) of that category, and the results achieved by a model trained using feature selection (FS). The features are ordered by their regression performance (MSE) when training a model solely on this single feature. Next to each MSE score, we report the type of model (e.g. Ridge). Last, we also report the standard deviation of the 10 runs within 5 by 2-fold CV.

The first thing one should note when looking at Table 1.1 is that only ridge and random forest models were chosen, and that the results for 1 by 10-fold and 5 by 2-fold CVs are rather similar. We compare each 5 by 2-fold MSE score using a univariate ANOVA with all models as conditions and calculate the contrasts to the mean baseline as references. The ANOVAs violated the sphericity assumption but still showed strong significance ( $p < 0.01$ ) after Greenhouse-Geisser correction of the degrees of freedom. Table 1.1 shows that all models are significantly better than the mean baseline (after Bonferroni correction).

When looking at the individual results in Table 1.1, one can see that already this baseline is actually quite good, with a MSE of 2.045 on a 9-point scale, which comes from the rather normally distributed ratings. Among our considered categories, text is the worst, followed by keyboard, body posture, and time, which show similar results. Much better and more interesting results are obtained in the three categories skin, eye, and heart measures, which again show similar results. When combining multiple modalities, the results improve a bit further.

Table 1.2 shows how the results change when including LMEMs and *adding the participant and segment* as additional features to the other regression models. This time only LMEMs and random forest models were chosen, and again the 1 by 10-fold and 5 by 2-fold scores are roughly comparable. We again use a univariate ANOVA (including Greenhouse-Geisser correction) and find that all models are significantly better than the baseline (after Bonferroni correction).

Table 1.1: Feature evaluation results *without considering LMEMs/without adding participant and segment*. For 10-fold and 5 by 2-fold CV with standard deviation (SD). Asterisk (\*) in the right column indicates a significant difference ( $p < 0.01$ ) from  $\text{SubjCL}_{\text{avg}}$  after Bonferroni correction.

		MSE	
	Features	1x10-CV↓(Reg.)	5x2-CV↓ (SD)
Baseline	$\text{SubjCL}_{\text{avg}}$	2.045 (-)	2.045 (0.04)
Time Features	ALL: PeTime, LNPeTime	1.457 (Ridge)	1.487 (Ridge) (0.11)*
	FS: PeTime	1.453 (Ridge)	1.490 (Ridge) (0.11)*
Text Features	ALL: TER, HTER, HBLEU, BLEU, SL	1.756 (Ridge)	1.764 (Ridge) (0.07)*
	FS: TER, HTER, SL	1.736 (Ridge)	1.747 (Ridge) (0.07)*
Keyboard	ALL: PWR, APR	1.551 (Ridge)	1.577 (Ridge) (0.08)*
	FS: PWR	1.554 (Ridge)	1.568 (Ridge) (0.07)*
Body Posture	ALL: HeadDist	1.471 (Ridge)	1.487 (RF) (0.11)*
	FS: HeadDist	1.456 (Ridge)	1.474 (RF) (0.12)*
Eyes	ALL: SearchProb, FixAmount, ICA, FixDur, SaccDur, Hilbert, EAR, BlinkAmount, PupilDiameter, NormFixAmount, NormBlinkAmount	0.965 (RF)	1.086 (RF) (0.08)*
	FS: FixAmount, ICA, FixDur, SaccDur, SearchProb, Hilbert, EAR, PupilDiameter	0.918 (RF)	1.029 (RF) (0.09)*
Heart	ALL: NN50, pNN50, BVPMedAbsDev, HR, SDNN, RMSSD, RR, BVPMeanAbsDiff, BVPamp, BVP	1.073 (RF)	1.130 (RF) (0.13)*
	FS: BVPMedAbsDev, NN50, SDNN, RMSSD, HR, RR, BVPamp, BVP	1.004 (RF)	1.117 (RF) (0.11)*
Skin	ALL: SkinTemp, Ledalab, FreqFrameGSR, GSR, FreqGSR	0.942 (RF)	1.148 (RF) (0.17)*
	FS: SkinTemp, FreqFrameGSR, Ledalab, GSR	0.858 (RF)	1.033 (RF) (0.14)*
Combined Features	ALL	0.857 (RF)	0.984 (RF) (0.15)*
	FS: FixAmount, ICA, SaccDur, NN50, SDNN, FixDur, RMSSD, FreqFrameGSR, HR, HeadDist, Ledalab, SearchProb, Hilbert, SkinTemp, EAR, GSR, PupilDiameter	0.718 (RF)	0.886 (RF) (0.12)*

When comparing the results of Table 1.2 to Table 1.1, we see that the results with participant and segment improved substantially for the time, text, keyboard, and body posture categories. For the other modalities – eyes, heart, skin, as well as combinations – the results are roughly comparable. Even though the performance improved, the text features remain the worst category, followed by the keyboard features. All other modalities now show similar results.

We also perform pairwise comparisons between the feature selection models of each individual category against the feature selected version of *combinations*, which we report in Table 1.3. Note that these results are using the models without incorporating participant and segment (Table 1.1), as we found these results more interesting. For the pairwise comparisons we use the 5 by 2-fold CV results in combination with a modified *t*-test (Dietterich 1998) followed by Bonferroni-Holm corrections.

As expected, the *combined* model is indeed significantly better than *time*, *text*, *keyboard*, and *body posture*; however, it is not significantly better compared to *eyes*, *heart*, and *skin*, which are already very good by themselves.

Summarizing, Tables 1.1 and 1.3 suggest that CL measurement without special adaptations per participant and segment work best when combining multiple modalities; however, using skin, eye, or heart measures also works similarly well. The often used keyboard features based on typing pauses, as well as time and body posture measures perform worse. The text metrics, which include common quality measures, are the worst among our explored predictors of subjective CL.

When the models can adapt to participant and segment (Table 1.2), the often used text and keyboard features remain the worst; however, all other categories (time, body posture, eyes, heart, skin, as well as combinations) now perform similarly well.

### 4.3 Pairwise correlations and PCA

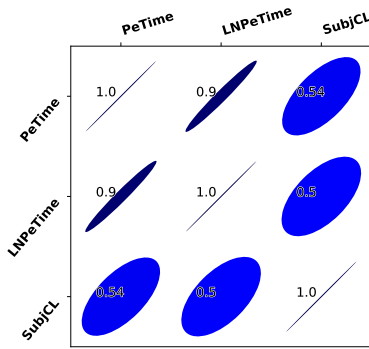
Similar to Vieira (2016), we analyze pairwise correlations between our measures of CL. For each modality, we report a maximum of 5 best features, which we compare to each other and to the subjective rating.

Figures 1.2, 1.3 and 1.4 depict the pairwise Pearson correlations alongside the PCA loadings, as described above. Narrower ellipses indicate stronger correlations; however, the correlation coefficient is also given numerically and encoded through coloring. Blue and upward-oriented ellipses indicate positive correlations, while red and downward-oriented ellipses indicate negative correlations. The PCA plot shows which feature loads on which PC. Here, the line thickness

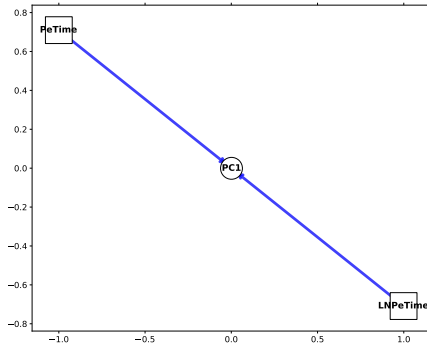
Table 1.2: Feature evaluation results when *considering LMEMs/adding participant and segment*. For 10-fold and 5 by 2-fold CV with standard deviation (SD). Asterisk (\*) in the right column indicates a significant difference ( $p < 0.01$ ) from  $\text{SubjCL}_{\text{avg}}$  after Bonferroni correction.

		MSE (L: LMEM, R: RF)	
	Features	1x10-CV↓(Reg.)	5x2-CV↓(SD)
Baseline	$\text{SubjCL}_{\text{avg}}$	2.045 (-)	2.045 (0.04)
Time Features	ALL: PeTime, LNPeTime	0.856 (L)	0.886 (L) (0.04)*
	FS: PeTime	0.868 (L)	0.891 (L) (0.05)*
Text Features	ALL: TER, HTER, HBLEU, BLEU, SL	1.126 (L)	1.219 (L) (0.07)*
	FS: TER, HTER, SL	1.121 (L)	1.193 (L) (0.04)*
Keyboard	ALL: PWR, APR	1.075 (L)	1.158 (L) (0.06)*
	FS: PWR	1.055 (L)	1.136 (L) (0.06)*
Body Posture	ALL: HeadDist	0.890 (L)	0.963 (L) (0.06)*
	FS: HeadDist	0.872 (L)	0.896 (L) (0.05)*
Eyes	ALL: SearchProb, FixAmount, ICA, FixDur, SaccDur, Hilbert, EAR, BlinkAmount, PupilDiameter, NormFixAmount, NormBlinkAmount	0.924 (R)	0.968 (R) (0.07)*
	FS: FixDur, SearchProb	0.882 (R)	0.938 (L) (0.09)*
Heart	ALL: NN50, pNN50, BVPMedAbsDev, HR, SDNN, RMSSD, RR, BVPMeanAbsDiff, BVPamp, BVP	0.921 (R)	1.057 (R) (0.11)*
	FS: HR	0.820 (L)	0.859 (L) (0.06)*
Skin	ALL: SkinTemp, Ledalab, FreqFrameGSR, GSR, FreqGSR	0.860 (R)	1.018 (R) (0.16)*
	FS: SkinTemp, GSR	0.816 (L)	0.919 (L) (0.16)*
Combined Features	ALL	0.801 (R)	0.962 (R) (0.12)*
	FixAmount, ICA, SaccDur, NN50, SDNN, FixDur, RMSSD, FreqFrameGSR, HR, HeadDist, Ledalab, SearchProb, Hilbert, SkinTemp, EAR, GSR, PupilDiameter	0.703 (R)	0.867 (R) (0.13)*

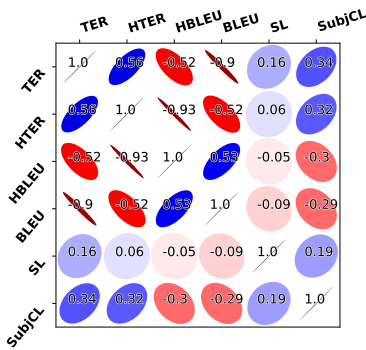
# 1 Multi-modal estimation of cognitive load in post-editing of MT



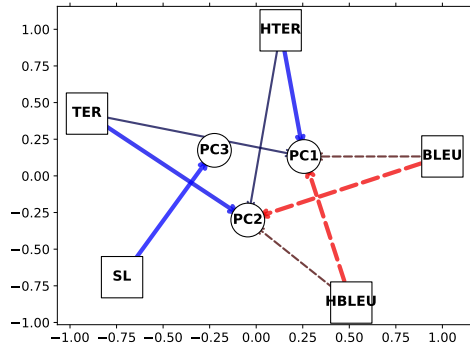
(a) Time – Pearson



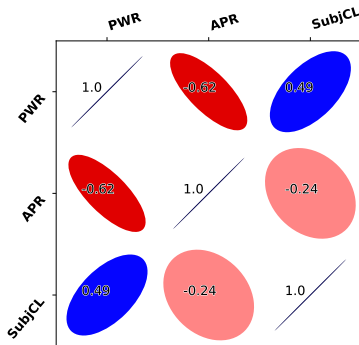
(b) Time – PCA



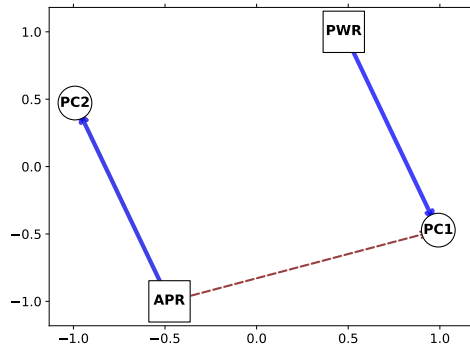
(c) Text – Pearson



(d) Text – PCA

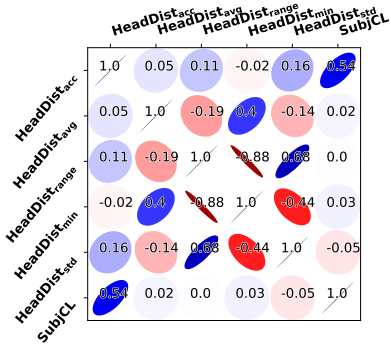


(e) Keyboard – Pearson

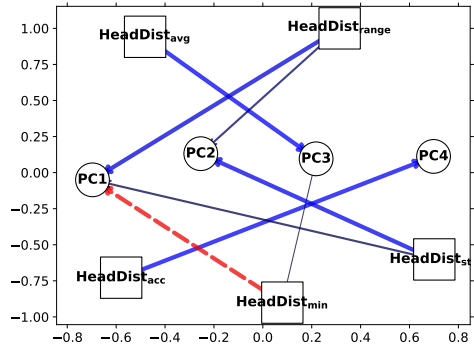


(f) Keyboard – PCA

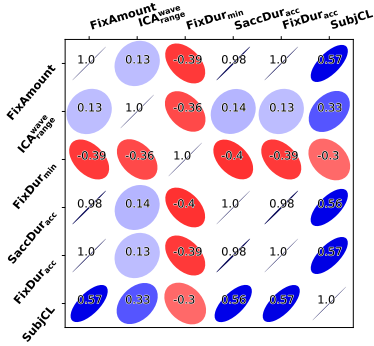
Figure 1.2: Correlations and PCA for time, text, and keyboard modalities.



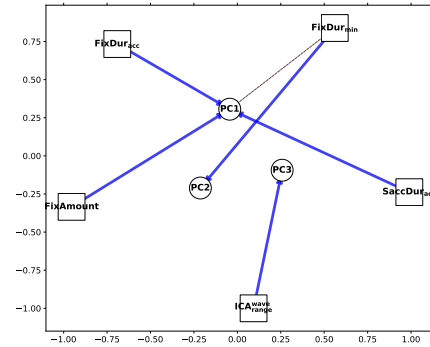
(a) Body Posture – Pearson



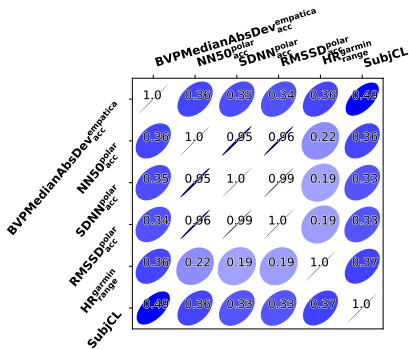
(b) Body Posture – PCA



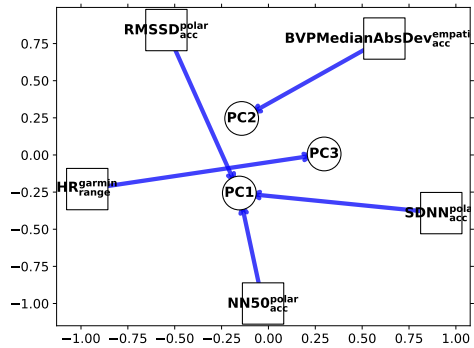
(c) Eyes – Pearson



(d) Eyes – PCA



(e) Heart – Pearson



(f) Heart – PCA

Figure 13: Correlations and PCA for body posture, eye, and heart modalities.

## 1 Multi-modal estimation of cognitive load in post-editing of MT

Table 1.3: Pairwise comparisons between the feature selected models *without LMEM/without participant and segment* (Table 1.1). \* shows significance with  $p < 0.05$  after Bonferroni-Holm correction.  $\tilde{t}$  is the test statistics for the modified paired  $t$ -test (Dietterich 1998).

Features	$\tilde{t}$
Time vs. combined	-4.06 *
Text vs. combined	-6.03 *
Keyboard vs. combined	-5.35 *
Body posture vs. combined	-6.32 *
Eyes vs. combined	-0.98
Heart vs. combined	-1.42
Skin vs. combined	-1.34

and color shows the strength of the loading; blue continuous lines represent positive loadings, while red dashed lines indicate negative loadings. For space reasons, we only summarize the most interesting results, which are all statistically significant.

For the *time features*, we see that PeTime and LNPeTime correlate very strongly and load on the same PC, but also that both show strong correlations to SubjCL.

For the *text features*, there expectedly are very strong correlations ( $-0.9$ ) between TER and BLEU and between HTER and HBLEU, where each pair also loads on the same PC. Furthermore, strong correlations can be observed between TER and HTER, as well as between BLEU and HBLEU.

For the *keyboard features*, we see a very strong correlation between APR and PWR, however, both load on distinct PCs. PWR correlates more strongly to SubjCL than APR, indicating that PWR is by itself a better estimator of SubjCL than APR.

As expected, the most relevant *eye features* FixAmount, SaccDur<sub>acc</sub>, and FixDur<sub>acc</sub> correlate by almost 1, load on the same PC, and strongly relate to SubjCL.

For the *heart features*, the correlations between NN50<sub>acc</sub><sup>polar</sup>, SDNN<sub>acc</sub><sup>polar</sup>, and RMSSD<sub>acc</sub><sup>polar</sup> are again very close to 1, and the PCA plot nicely visualizes that they cluster together. BVPMedAbsDev shows the strongest correlation to SubjCL.

Inspecting the most relevant *skin features*, we see very strong correlations between FreqFrameGSR<sub>avg</sub><sup>64,Empatica</sup> and Leda<sub>avg</sub>, as well as medium to strong correlations between the frequency frame and SkinTemp<sub>acc</sub><sup>Garmin</sup> features.

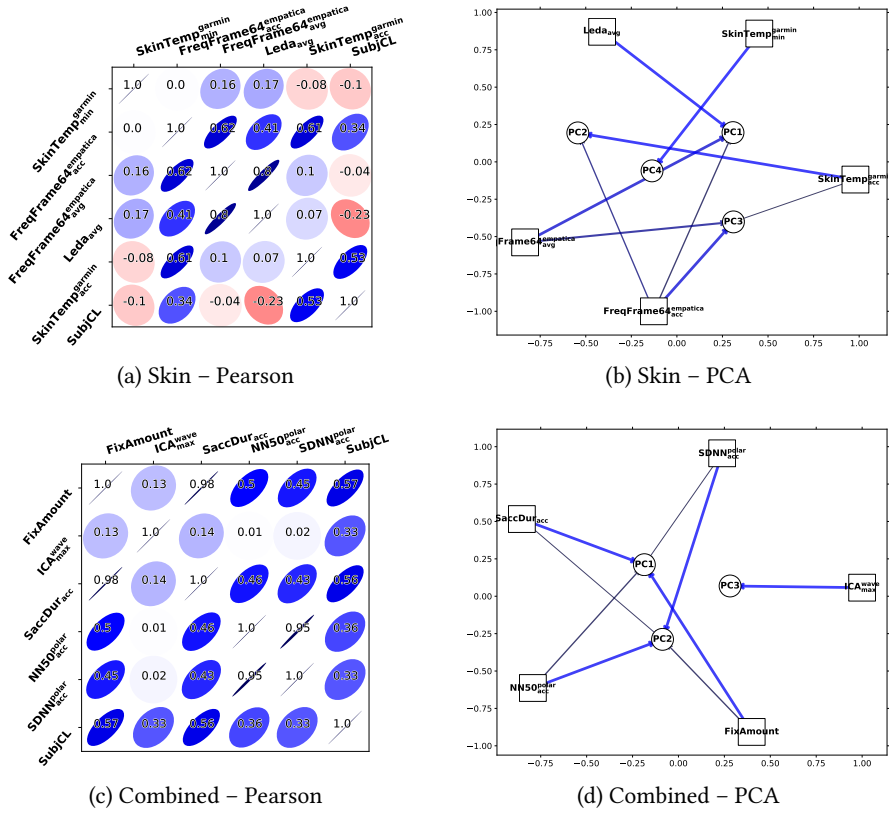


Figure 1.4: Correlations and PCA for the skin and combined modalities.

Most interestingly, for the *combined features* we can again see that  $SDNN_{acc}^{polar}$  and  $NN50_{acc}^{polar}$ , as well as  $FixAmount$  and  $SaccDur_{acc}$ , correlate with almost a value of 1. There also seems to be a strong link between the HRV measures and the eye measures  $SaccDur_{acc}$  and  $FixAmount$ . The PCA further shows that there is one PC for the HRV measures, one for the ICA, and another one for the eye features  $FixAmount$  and  $SaccDur_{acc}$ .

## 4.4 Discussion

Overall, very good regression results of up to 0.7 MSE on a 9-point scale were achieved by our regression models. This amount of error should be acceptable for most possible applications discussed in Herbig, Pal, van Genabith, et al. 2019. While the 5 by 2-fold CV results are often slightly worse, which might be be-



cause less training data was seen, the results of 1 by 10-fold and 5 by 2-fold are comparable, and the very small standard deviations indicate model robustness.

When comparing the regression results without adding participant and segment to Herbig, Pal, Vela, et al. (2019), whose approach is almost the same apart from having fewer sensors and features, we note a few similarities and differences: first of all, we found consistently better results across all modalities; however, already the baseline yields better results on our dataset. While the time features in Herbig, Pal, Vela, et al. (2019) were rather good, they are among the worst modalities here. A reason might be that we considered many more features, that helped the other modalities improve over the time as a feature. Furthermore, while in Herbig, Pal, Vela, et al. (2019) the eyes were by far the best among the three main categories eye, skin, and heart, all three show similar results here. This could be due to the numerous additional skin and heart features considered in our analysis. Whereas in both studies the combined approach leads to the best results, the performance gains when combining multiple modalities were much stronger in Herbig, Pal, Vela, et al. (2019), probably again because the three main categories are already very good by themselves.

So when we do not consider the individual participant and the segment they are post-editing (Table 1.1 or Herbig, Pal, Vela, et al. 2019), we can achieve the best results only with our main categories, eyes, heart, skin, or by combining features from several modalities. This is relevant for less controlled and more practical applications, e.g. adapting the user interface to perceived CL, where it is impossible to use participant and segment information, as ideally no two translators should post-edit the same sentence (which would otherwise be contained in TM).

In contrast, when we do consider participant and segment (Table 1.2), modalities of lesser quality, like time, text, keyboard, or body posture can also achieve good results. So considering *who is editing what* seems to yield enough information to learn from when combined with these features, while without considering participant and segment, the generalization is impeded. However, if the goal is to conduct a controlled experiment, e.g. to investigate the impact of different sentence features on subjectively felt CL, integrating participant and segment into the models allows to also achieve valuable estimates with these other modalities. The above experiment therefore also suggests that text quality, keyboard, and time measures, which are frequently used in the literature to estimate effort, only work well in controlled settings.

While we cannot compare all our correlation and PCA results to Vieira (2016), since we considered many more features, there is still some interesting overlap: The time features in both studies correlated strongly to SubjCL. Furthermore, the link between the PWR and SubjCL also seems comparable, while that between

APR and SubjCL appears weaker in our dataset. However, the correlation between these two keyboard features is similarly strong in both studies. The eye features FixAmount and FixDur also correlate to a similar extent with SubjCL in both studies. To summarize, we could both reproduce (except APR vs. SubjCL) and extend the findings by Vieira (2016), which strengthens our results.

The correlation and PCA especially revealed that many highly redundant features were selected by the feature selection approach (e.g. the HRV measures). The reason for this probably is their strong correlation to SubjCL; however, due to the redundancy, it is unclear whether incorporating multiple such features really helps. Therefore, we want to explore if handcrafting a set of features with fewer redundancies, or using a more sophisticated feature selection approach than RFECV, could boost the performance further. Since space constraints allowed us to analyze only very few features in terms of correlations and PCA, we also plan to investigate the link to the non-selected features, as well as a PCA including more features from all different modalities than the few reported here.

#### **4.5 Limitations**

The results presented in this study are subject to the following limitations: The data sample is relatively small, since only 10 subjects participated in our study. Next, while we performed CV and only report results on segments unseen during training, we did not completely leave out participants and then predict those participants' perceived CL from the data gathered by the other participants. Thus, to achieve these results in practice one may need to fine-tune and train for new users. Moreover, one should also note that our eye tracker only samples at 90 Hz, which could affect the peak velocity reconstruction and thereby saccades (Mack et al. 2017). Last, while our predictive approach yields interesting first insights, it is only an automatic "top-down" approach that might be improved by selecting an optimal set of features and tuning the hyper-parameters.

## **5 Conclusions and future work**

In this paper, we have focused on perceived cognitive PE effort and argued for the need to robustly measure CL during PE. In contrast to most related work, we investigated whether and how multiple modalities to measure CL can be combined and used for the task of predicting the level of perceived CL during PE of MT. To the best of our knowledge, our analyzed feature set comprises the most

diverse set of features from a variety of modalities that has to date been investigated in the translation domain, considering even more factors than Herbig, Pal, Vela, et al. (2019).

Based on the data gathered from 10 professional translators, we report how well subjective CL can be predicted depending on the various features: When the models are unaware of which participant and segment the data belongs to, eye, skin, and heart features, or a combination of different modalities, performed best. In contrast, for regression models that can react differently depending on participant and segment, the less well performing categories time, text, keyboard, and body posture also achieved good results, probably due to overfitting on the participant. While this finding is very interesting for controlled experiments, it is less relevant for practical use, where no two participants should PE the same segment. Overall, the trained models can estimate CL during PE without interrupting the actual process through manual ratings with comparably low error of at best 0.7 MSE on a 9-point scale. However, further data analysis is needed to understand the required steps to achieve such results in practice.

We also report how strongly the different measures correlate and which features cluster together, where we reproduce almost all the findings of Vieira (2016) and extend them further by considering many more features.

In the future, we want to conduct more detailed investigations, e.g. in terms of a more complex feature selection approach or hand-crafting a subset of features based on the correlation and PCA findings, in combination with hyper-parameter tuning, to make better use of the available data than the chosen “top-down” regression approach. Furthermore, we want to use the captured continuous signals to already predict perceived CL while still editing the segment (i.e. based on a time window of the data), to allow for more real-time applications.

The long-term goal is to be able to decrease the perceived CL, and thereby stress and exhaustion, during PE. As discussed in Herbig, Pal, van Genabith, et al. (2019), this could be achieved by fine-tuning MT systems on the user’s CL measurements to produce less demanding outputs, or by automatically showing alternative translations or other forms of assistance. The measurement techniques explored within this paper form the basis for future research towards this goal.

## Acknowledgments

This research was funded in part by the Deutsche Forschungsgemeinschaft (DFG) under grant number GE 2819/2-1/AOBJ: 636684. The responsibility lies with the authors.

## References

- Arshad, Syed, Yang Wang & Fang Chen. 2013. Analysing mouse activity for cognitive load detection. In *Proceedings of the 25th Australian computer-human interaction conference: Augmentation, application, innovation, collaboration*, 115–118.
- Asteriadis, Stylianos, Paraskevi Tzouveli, Kostas Karpouzis & Stefanos Kollias. 2009. Estimation of behavioral user state based on eye gaze and head pose: Application in an e-learning environment. *Multimedia Tools and Applications* 41(3). 469–493.
- Benedek, Mathias & Christian Kaernbach. 2010. A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods* 190(1). 80–91.
- Carl, Michael, Barbara Dragsted, Jakob Elming, Daniel Hardt & Arnt Lykke Jakobsen. 2011. The process of post-editing: A pilot study. *Copenhagen Studies in Language* 41. 131–142.
- Chen, Fang, Jianlong Zhou, Yang Wang, Kun Yu, Syed Z. Arshad, Ahmad Khawaji & Dan Conway. 2016. *Robust multimodal cognitive load measurement*. Cham: Springer International Publishing.
- Chen, Siyuan & Julien Epps. 2013. Automatic classification of eye activity for cognitive load measurement with emotion interference. *Computer Methods and Programs in Biomedicine* 110(2). 111–124.
- Daems, Joke. 2016. *A translation robot for each translator?: A comparative study of manual translation and post-editing of machine translations: Process, quality and translator attitude*. Ghent University. (Doctoral dissertation).
- Demberg, Vera & Asad Sayeed. 2016. The frequency of rapid pupil dilations as a measure of linguistic processing difficulty. *PloS One* 11(1). 1–29.
- Dietterich, Thomas G. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation* 10(7). 1895–1923.
- Doherty, Stephen, Sharon O’Brien & Michael Carl. 2010. Eye tracking as an MT evaluation technique. *Machine Translation* 24(1). 1–13.
- Gehring, Jonas, Michael Auli, David Grangier, Denis Yarats & Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th international conference on machine learning*, vol. 70, 1243–1252.
- Goldberg, Joseph H. & Xerxes P. Kotval. 1999. Computer interface evaluation using eye movements: Methods and constructs. *International Journal of Industrial Ergonomics* 24(6). 631–645.
- Haapalainen, Eija, SeungJun Kim, Jodi F. Forlizzi & Anind K. Dey. 2010. Psychophysiological measures for assessing cognitive load. In *Proceedings of the 12th ACM international conference on ubiquitous computing*, 301–310.

- Herbig, Nico, Santanu Pal, Josef van Genabith & Antonio Krüger. 2019. Multi-modal approaches for post-editing machine translation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–11.
- Herbig, Nico, Santanu Pal, Mihaela Vela, Antonio Krüger & Josef van Genabith. 2019. Multi-modal indicators for estimating perceived cognitive load in post-editing of machine translation. *Machine Translation* 33(1–2). 91–115.
- Hockey, Robert. 1997. Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. *Biological Psychology* 45(1). 73–93.
- Hossain, Gahangir & Mohammed Yeasin. 2014. Understanding effects of cognitive load from pupillary responses using Hilbert analytic phase. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 375–380.
- Iani, Cristina, Daniel Gopher & Peretz Lavie. 2004. Effects of task difficulty and invested mental effort on peripheral vasoconstriction. *Psychophysiology* 41(5). 789–798.
- Iqbal, Shamsi T., Xianjun Sam Zheng & Brian P. Bailey. 2004. Task-evoked pupillary response to mental workload in human-computer interaction. In *Extended abstracts on human factors in computing systems*, 1477–1480.
- Koglin, Arlene. 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* 7(1). 126–141.
- Koponen, Maarit. 2016. Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. *The Journal of Specialised Translation* 25. 131–148.
- Kramer, Arthur F. 1991. Physiological metrics of mental workload: A review of recent progress. In Diane L. Damos (ed.), *Multiple-task performance*, chap. 11, 279–328. Oxfordshire: Taylor & Francis.
- Krings, Hans P. 2001. *Repairing texts: Empirical investigations of machine translation post-editing processes*. Vol. 5. Kent, OH: Kent State University Press.
- Lacruz, Isabel & Gregory M. Shreve. 2014. Pauses and cognitive effort in post-editing. In Sharon O'Brien, Laura W. Balling, Michael Carl, Michel Simard & Lucia Specia (eds.), *Post-editing of machine translation: Processes and applications*, chap. 11, 246–274. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Lacruz, Isabel, Gregory M. Shreve & Erik Angelone. 2012. Average pause ratio as an indicator of cognitive effort in post-editing: A case study. In *AMTA workshop on post-editing technology and practice*, 21–30.

- Lavie, Alon & Abhaya Agarwal. 2007. Meteor: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the second workshop on statistical machine translation*.
- Lin, Chin-Yew & Franz J. Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd annual meeting of the Association for Computational Linguistics*. Barcelona, Spain.
- Macizo, Pedro & M. Teresa Bajo. 2006. Reading for repetition and reading for translation: Do they involve the same processes? *Cognition* 99(1). 1–34.
- Mack, David J., Sandro Belfanti & Urs Schwarz. 2017. The effect of sampling rate and lowpass filters on saccades—a modeling approach. *Behavior Research Methods* 49(6). 2146–2162.
- Mellinger, Christopher Davey. 2014. *Computer-assisted translation: An empirical investigation of cognitive effort*. Kent, OH: Kent State University.
- Moorkens, Joss, Sharon O’Brien, Igor A. L. Da Silva, Norma B. De Lima Fonseca & Fabio Alves. 2015. Correlations of perceived post-editing effort with measurements of actual effort. *Machine Translation* 29(3). 267–284. DOI: 10.1007/s10590-015-9175-2.
- Mulder, Lambertus J. M. 1992. Measurement and analysis methods of heart rate and respiration for use in applied environments. *Biological Psychology* 34(2). 205–236.
- O’Brien, Sharon. 2005. Methodologies for measuring the correlations between post-editing effort and machine translatability. *Machine Translation* 19(1). 37–58.
- O’Brien, Sharon. 2006a. Eye-tracking and translation memory matches. *Perspectives: Studies in Translatology* 14(3). 185–205.
- O’Brien, Sharon. 2006b. Pauses as indicators of cognitive effort in post-editing machine translation output. *Across Languages and Cultures* 7(1). 1–21.
- Paas, Fred G.W.C., Juhani E. Tuovinen, Huib Tabbers & Pascal W.M. Van Gerwen. 2003. Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist* 38(1). 63–71.
- Paas, Fred G.W.C. & Jeroen J.G. van Merriënboer. 1994. Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review* 6(4). 351–371.
- Papineni, Kishore, Salim Roukos, Todd Ward & Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the ACL*, 311–318. Philadelphia, Pennsylvania.

- Rowe, Dennis W., John Sibert & Don Irwin. 1998. Heart rate variability: Indicator of user state as an aid to human-computer interaction. In *Proceedings of the conference on human factors in computing systems*, 480–487.
- Shaffer, Fred & Jay P. Ginsberg. 2017. An overview of heart rate variability metrics and norms. *Frontiers in Public Health* 5. 1–7.
- Shi, Yu, Natalie Ruiz, Ronnie Taib, Eric Choi & Fang Chen. 2007. Galvanic skin response (GSR) as an index of cognitive load. In *Extended abstracts on human factors in computing systems*, 2651–2656.
- Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla & John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the association for machine translation in the Americas*, 223–231.
- Snover, Matthew, Nitin Madnani, Bonnie Dorr & Richard Schwartz. 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the 4th workshop on statistical machine translation*, 259–268.
- Soukupova, Tereza & Jan Cech. 2016. Real-time eye blink detection using facial landmarks. In *21st computer vision winter workshop*, 1–8.
- Stuyven, Els, Koen Van der Goten, André Vandierendonck, Kristl Claeys & Luc Crevits. 2000. The effect of cognitive load on saccadic eye movements. *Acta Psychologica* 104(1). 69–85.
- Sweller, John, Jeroen J.G. van Merriënboer & Fred G.W.C. Paas. 1998. Cognitive architecture and instructional design. *Educational Psychology Review* 10(3). 251–296.
- van den Berg, Marten, Peter Rijnbeek, Maartje Niemeijer, Albert Hofman, Gerard van Herpen, Michiel Bots, Hans Hillege, Kees Swenne, Mark Eijgelsheim, Bruno Stricker & Jan Kors. 2018. Normal values of corrected heart-rate variability in 10-second electrocardiograms for all ages. *Frontiers in Physiology* 9. 1–9.
- Van Orden, Karl F., Wendy Limbert, Scott Makeig & Tzyy-Ping Jung. 2001. Eye activity correlates of workload during a visuospatial memory task. *Human Factors* 43(1). 111–121.
- Vieira, Lucas Nunes. 2014. Indices of cognitive effort in machine translation post-editing. *Machine Translation* 28(3–4). 187–216.
- Vieira, Lucas Nunes. 2016. How do measures of cognitive effort relate to each other? A multivariate analysis of post-editing process data. *Machine Translation* 30(1–2). 41–62.

- Villarejo, María Viqueira, Begoña García Zapirain & Amaia Méndez Zorrilla. 2012. A stress sensor based on galvanic skin response (GSR) controlled by ZigBee. *Sensors* 12(5). 6075–6101.
- Yamakoshi, Takehiro, Ken-ichi Yamakoshi, Shinobu Tanaka, Masamichi Nogawa, Sang-Bum Park, Mariko Shibata, Yoshito Sawada, Peter Rolfe & Yasuo Hirose. 2008. Feasibility study on driver's stress detection from differential skin temperature measurement. In *Engineering in medicine and biology society*, 1076–1079.



## Chapter 2

# Comparing NMT and PBSMT for post-editing in-domain formal texts: A case study

Sergi Álvarez<sup>a</sup>, Toni Badia<sup>a</sup> & Antoni Oliver<sup>b</sup>

<sup>a</sup>Universitat Pompeu Fabra <sup>b</sup>Universitat Oberta de Catalunya

This paper details a comparative analysis between phrase-based statistical machine translation (PBSMT) and neural machine translation (NMT) for English-Spanish in-domain medical documents using human rankings, fluency and adequacy, and post-editing (technical and temporal) effort, performed by professional translators. When MT output is ranked against translations performed by professional translators, results show a clear preference for human translations, with NMT in the second position. Regarding MT outputs, NMT is perceived as more fluent and conveying better the meaning of the source sentence. Despite this preference, post-editing temporal effort does not improve significantly in NMT compared to PBSMT, although technical effort is reduced.

## 1 Introduction

Over the last years, post-editing of machine translation (PEMT) has become common practice in the translation industry. It has been included as part of the translation workflow because it increases productivity and reduces costs (Guerberof 2009a). A recent survey showed that more than half of the language service providers (LSPs) offered PEMT as a service (Lommel & DePalma 2016). Post-editors “edit, modify and/or correct pre-translated text that has been processed by an MT system from a source language into (a) target language(s)” (Allen 2003).



Yet, many professional translators state that after post-editing a few MT segments, they delete the remaining segments and translate everything from scratch if they consider it will take them less time (Parra Escartín & Arcedillo 2015).

Effective PE, therefore, requires sufficient quality of the MT output. The issue, then, is how to detect that a machine translation output is good enough to serve as input to PE. Very often, the usual automatic metrics do not always correlate to PE effort (Koponen 2016). Even translators' perception does not always match PE effort (Koponen 2012; Moorkens 2018). Research in this field has mainly focused on measuring the PE effort related to MT output quality (Guerberof 2009a,b; Specia 2011; 2010), productivity (O'Brien 2011; Parra Escartín & Arcedillo 2015; Plitt & Masselot 2010; Sanchez-Torron & Koehn 2016), translator's usability (Castilho et al. 2014; Moorkens & O'Brien 2013) and perceived PE effort (Moorkens et al. 2015).

Statistical machine translation (SMT) has been well established as the dominant approach in machine translation for many years. However, in the last few years, research has become more interested in neural machine translation after the computational limitations have been solved (Bahdanau et al. 2018; Cho et al. 2014). The first results obtained have been very successful in terms of quality, for example in WMT 2016 (Bojar et al. 2016), WMT 2017 (Bojar et al. 2017), and WMT 2018 (Bojar et al. 2018). These promising results have driven a technological shift from (phrase-based) statistical machine translation (SMT) to neural machine translation (NMT) in many translation industry scenarios.

All of the current research on post-editing machine translation output uses the division established by Krings (2001) regarding PE effort: temporal effort (time spent PE), technical effort (number of edits, often measured using keystroke analysis), and cognitive effort (usually measured with eye-tracking or think-aloud protocols). Even though no current measure includes all three dimensions, cognitive effort correlates with technical and temporal PE effort (Moorkens et al. 2015). In our experiments, we use automatic measures of both temporal and technical effort.

As this new approach to MT becomes more popular among LSPs and translators, it is essential to test what NMT can offer for PE in terms of quality compared to the results of PBSMT. Recent studies (Bentivogli et al. 2016; Castilho, Moorkens, Gaspari, Sennrich, et al. 2017; Toral & Sánchez-Cartagena 2017) have stated an improved quality of NMT for PE. In this paper, we continue in this direction, but we focus on in-domain formal documents, which are the ones usually post-edited by professional translators.

Our objectives with these experiments are threefold:

## 2 Comparing NMT and PBSMT for post-editing in-domain formal texts

- Determine which MT method (PBSMT or NMT) yields better results for PE in-domain formal texts.
- Analyze the relation between human and automatic metrics for PE.
- Study translators perception as a prospective measure of PE effort.

In Section 2, we review previous work comparing SMT and NMT approaches. In Section 3 we describe the MT systems and the training corpus used. In Section 4 we include the automatic evaluation of the MT systems used. We give details about the methodology used for our experiments in Section 5. We explain the results obtained in Section 6 and, finally, we state the main conclusions and our plans for future work in Section 7.

## 2 Previous work

One of the first complete papers studying the impact of SMT and NMT in PE was Bentivogli et al. (2016). In it, they carry out a small scale study on post-editing NMT and SMT outputs of English to German translated TED talks. They conclude that NMT generally decreases the PE effort, but degrades faster than SMT with sentence length. One of the main strengths of NMT is the reordering of the target sentence.

Wu et al. (2016) evaluate the quality of NMT and SMT, in this case using BLEU (Papineni et al. 2002) and human scores for machine-translated Wikipedia entries. Results show that NMT systems outperform and improve the quality of MT results. Other studies have confirmed this diagnostics (Junczys-Dowmunt et al. 2016; Isabelle et al. 2017), as have the results of the automatic PE tasks at the Conference on Machine Translation (Bojar et al. 2016; 2017).

Toral & Sánchez-Cartagena (2017) broaden the scope of Bentivogli et al. (2016) adding different language combinations and metrics, and they conclude that although NMT yields better quality results in general, it is negatively affected by sentence length, and the improvement of the results is not always perceivable in all language pairs.

Castilho et al. (2017) discuss three studies using automatic and human evaluation methods. One of them includes in-domain formal texts for chemical patent titles and abstracts. In addition to the automatic metrics, two reviewers assess 100 random segments to rank the translations and to identify translation errors. Automatic evaluation doesn't give clear results, but the SMT system is ranked higher than NMT in human evaluation.

Castilho et al. (2017) report on a comparative study of PBSMT and NMT, with four language pairs and different automatic metrics and human evaluation methods. It highlights some strengths and weaknesses of NMT, which in general yields better results. The study focuses especially on PE and uses the PET interface (Aziz et al. 2012) to compare educational domain output from both systems using different metrics. They conclude that NMT reduces word order errors and improves fluency for certain language pairs, so fewer segments require PE, especially because there is a reduction in the number of morphological errors. However, they don't detect a decrease in PE effort nor a clear improvement in omission and mistranslation errors.

Our experiments study the differences of post-editing NMT and SMT outputs for formal in-domain texts. We compare the usual automatic scores for MT with direct and indirect PE effort metrics. Mainly, we study translators' perception regarding quality, and fluency and accuracy, and analyze temporal and technical post-editing effort.

## **3 MT systems and training corpus**

### **3.1 MT systems**

In order to help contextualise the results in our experiments, we have decided to use two MT systems as references to compare their results with the ones of the systems we trained. As reference MT systems, we have chosen Apertium (Forcada et al. 2011), a shallow transfer MT system, and Google Translate, a neural MT system for the English-Spanish language pair, which is the one we use in our experiments.

For training the PBSMT and neural MT systems we have used ModernMT (Germann et al. 2016) version 2.4. This version allows to train both statistical and neural MT systems. We have used the default options for this version. One of the salient characteristics of ModernMT is the fact that it can take into account the context of the sentence to be translated. In the evaluation results, we show figures for both cases: with and without taking the context into account. In the experiments we take context to be the previous and the next segment (except for the first and last segment, where we have taken into account the next and the previous segment only, respectively). Short contexts are usually enough to calculate the context vector used by ModernMT.

### 3.2 Data: Medical corpus

To train the system, we have compiled all of the publicly available corpora in the English-Spanish pair known to us. We have also created several corpora from websites with medical content:

- The EMEA<sup>1</sup> (*European Medicines Agency*) corpus.
- The IBECs<sup>2</sup> (*Spanish Bibliographical Index in Health Sciences*) corpus.
- Medline Plus:<sup>3</sup> we have compiled our own corpus from the web and we have combined this with the corpus compiled in MeSpEn<sup>4</sup>.
- MSDManuals<sup>5</sup> English-Spanish corpus, compiled for this project under permission of the copyright holders.
- Portal Clínic<sup>6</sup> English-Spanish corpus, compiled by us for this project.
- The PubMed<sup>7</sup> corpus.
- The UFAL Medical Corpus<sup>8</sup> v1.0.

We have also treated as a corpus glossaries and glossary-like databases containing a lot of useful terms and expressions in the medical domain. Namely, we have used the English-Spanish glossary from MeSpEn, the 10th revision of the international statistical classification of ICD and SnowMedCT.

With all the corpora and glossaries we have created an in-domain training corpus of 2,836,580 segments and entries. We have split the corpus in two parts: 99% of the segments for training, and the remaining 1% for testing.

We have also used other general corpora for training the MT systems, namely the Scielo corpus, the Europarl corpus<sup>9</sup> (Koehn 2005), Global Voices corpus<sup>10</sup> and

---

<sup>1</sup><http://opus.npl.eu/EMEA.php>

<sup>2</sup><http://ibecs.isciii.es>

<sup>3</sup><https://medlineplus.gov/>

<sup>4</sup><http://temu.bsc.es/mespen/>

<sup>5</sup><https://www.msdmanuals.com/>

<sup>6</sup><https://portal.hospitalclinic.org>

<sup>7</sup><https://www.ncbi.nlm.nih.gov/pubmed/>

<sup>8</sup>[https://ufal.mff.cuni.cz/ufal\\_medical\\_corpus](https://ufal.mff.cuni.cz/ufal_medical_corpus)

<sup>9</sup><http://www.statmt.org/europarl/>

<sup>10</sup><https://globalvoices.org/>

News Commentary. The IBECs, Scielo, Pubmed and a part of the MedlinePlus corpus have been obtained from the MeSpEn corpus<sup>11</sup> (Villegas et al. 2018).

In Table 2.1 the size of all corpora and glossaries used for training the MT systems are shown. The figures are calculated after eliminating all the repeated source segment – target segment pairs in the corpora.

Table 2.1: Size of the corpora and glossaries used to create the corpus to train the MT systems.

Corpus	Segments/Entries	Tokens eng	Tokens spa
EMEA	366,769	5,327,963	6,008,543
IBECs	628,798	13,432,096	14,879,220
MedLine Plus	15,689	209,074	234,660
MSD Manuals	241,336	3,719,933	4,467,906
Portal Clinic	8,797	159,717	169,294
PubMed	320,475	2,752,139	3,035,737
UFAL	258,701	3,202,162	3,437,936
Glossary MeSpEn	125,645	286,257	348,415
ICD10-en-es	5,202	25,460	30,580
SnowMedCT Denom.	887,492	3,509,062	4,457,681
SnowMedCT Def.	4,268	177,861	184,574
In-domain	2,836,580	32,479,955	36,893,257
Scielo	741,407	17,464,256	19,305,165
Europarl	1,961,672	50,008,219	52,489,142
Global Voices	559,418	10,717,938	11,496,683
News Commentary	259,412	5,898,912	6,903,975
Out-of-domain	3,521,363	84,087,899	90,193,659

## 4 Automatic evaluation of the MT systems

In Table 2.2 we can observe the evaluation values of the trained systems using MTEval<sup>12</sup> along with Apertium and Google Translate. This software allows to calculate BLEU, NIST, RIBES and WER using only one reference. We have used all

<sup>11</sup><http://temu.bsc.es/mespen/>

<sup>12</sup><https://github.com/odashi/mteval>

the test sets of the corpus. As shown in the table, the systems trained in the experiment obtain better results in all metrics than the reference systems used, except for the Google Translate system, which obtains a slightly better NIST result than the MMT Phrase-Based system without context and a better WER result than the two MMT Phrase-Based systems. The MMT Neural system performs consistently better than the MMT Phrase-Based system. In the MMT Neural system, we do not see any significant difference between the results obtained when trained with or without context.

Table 2.2: Results of the automatic evaluation using mteval.

MT system	BLEU	NIST	RIBES	WER
Apertium	0.192577	6.442539	0.713117	0.702716
Google T.	0.402497	9.632268	0.809469	0.530053
MMT P.B. no context	0.424183	9.536248	0.814425	0.637821
MMT P.B. context	0.444832	9.801466	0.819303	0.621032
MMT Neural no context	0.503935	11.106222	0.836954	0.485474
MMT Neural context	0.505778	11.141294	0.836313	0.481039

## 5 Experiments

We carried out three different experiments with English-Spanish medical texts to assess human perception and evaluation of both PBSMT and NMT systems.

### 5.1 Translation ranking

In the first part, participants had to answer some questions about their previous experience in the translation industry. The survey was open both to students and professional translators as we were mainly interested in the perception of quality. In the second part of the survey, participants had to rank the translation of 40 segments (human translation, NMT and PBSMT), which had no context and were randomized to avoid bias. They were selected so there were no repeated translations and all had a minimum length of 100 characters. Then we applied a script to ensure there was a minimum editing distance of 15% between the human-PBSMT, human-NMT and PBSMT-NMT solutions. This reduced the number of segments from 230 to 145. We hand-picked 40 segments without typos nor any other problem.

## 5.2 Fluency and adequacy

We presented a survey with the same English segments as in the previous experiment. In the first part, participants (both students and professional translators) had to answer some questions about their previous experience in the translation industry. Afterwards, they had to evaluate the fluency and adequacy of the proposed translation on a four-point Likert scale. The translation was either PBSMT or NMT chosen randomly without any knowledge of the participants. The goal was to assess fluency and adequacy for in-domain formal texts.

## 5.3 PE time and technical effort

Finally, in the third experiment, participants had to post-edit 41 segments from a 2018 medical paper. They had to carry out the task in PET (Aziz et al. 2012)<sup>13</sup>, a computer-assisted translation tool that supports PE. It was used with its default settings. It logged both PE time and edits (keystrokes, insertions and deletions, that is, technical effort). Four professional translators with more than two years of experience post-editing carried out the task: two of them post-edited the PBSMT output and the other two post-edited the NMT output.

# 6 Results

## 6.1 Translation ranking

29 people answered the survey. From those, 86.21% had previous experience as translators and 58.62% had worked on PE tasks. Confirming the initial hypothesis, most respondents preferred the human translation. However, this percentage was only of 60.52%. The second most preferred translation was NMT, with 25.17%, and PBSMT was only considered the best translation for 14.31% of the segments. We calculated inter annotator agreement using Fleiss' kappa (Fleiss 1971), which showed a fair agreement among the annotators ( $\kappa = 0.36$ ). These results were statistically significant in a one-way ANOVA comparison ( $p < 0.05$ ).

Although the survey was conducted on a fairly small number of sentences, it seems to point in two directions: NMT is far from achieving the quality of human translation for medical texts, and NMT yields better translations than PBSMT. We conducted a manual analysis of the sentences in which NMT or PBSMT were selected as the best translation. It was observed the main reason for the selection was terminology precision and fluency of the MT output.

---

<sup>13</sup><http://wilkeraziz.github.io/dcs-site/pet/index.html>



Table 2.3: Results of the human-NMT-PBSMT ranking survey.

Evaluation	Human	NMT	PBSMT
EN-ES (40)	60.52%	25.17%	14.31%

## 6.2 Fluency and adequacy

In the second experiment, eleven people answered the survey. Seven of them were translators with more than two years of experience and only four of them were students. Both fluency and adequacy obtained a higher rate for NMT after calculating the mean for both MT systems. We calculated inter annotator agreement using Fleiss’ kappa (Fleiss 1971). For fluency, it showed poor agreement among the annotators ( $\kappa = 0.01$ ). Results were statistically significant in a one-way ANOVA comparison, with an  $F$ -ratio value of 2.75586 and a  $p$ -value of 0.04856 (significance at  $p < 0.05$ ). For adequacy, there was also poor agreement among annotators. These results weren’t statistically significant, with an  $F$ -ratio value of 0.96767 and a  $p$ -value of 0.412816 ( $p < 0.05$ ).

If we take a closer look at the sentences that had to be assessed, PBSMT segments often contain morphological problems (e.g. concordance) that we cannot spot in NMT segments, as in example (1). This way the generally higher ratings for fluency and adequacy of the NMT system are confirmed.

- (1) Source: Craniopharyngioma had more hormone deficiencies  
 Gloss: Craneofaringioma tenían más déficits hormonales  
 PBSMT: ‘Craneofaringioma/had (plural)/more/deficits/hormonal’

Table 2.4: Results of the ranking survey.

System	Fluency	Adequacy
PBSMT	2.28	2.24
NMT	2.46	2.50

## 6.3 PE time and technical effort

Results for the PE task by professional translators have been grouped in temporal effort and technical effort (see Tables 2.5 and 2.6). In both cases, the mean for

PBSMT is higher, though only technical effort shows a statistically significant difference (in a  $t$ -test with a  $p$ -value of 0.002054). It is worth highlighting that there was a considerable difference in time and keylogging between the translators, especially for the two professionals who post-edited PBSMT (as indicated by the standard deviation in Tables 2.5 and 2.6).

Table 2.5: Temporal PE effort (secs/segment).

System	Mean	SD
PBSMT	88.75	44.59
NMT	79.25	33.43

Table 2.6: Technical effort (keystrokes/segment).

System	Mean	SD
PBSMT	130.68	39.63
NMT	54.99	16.90

## 7 Conclusions and future work

Although the number of segments analyzed is quite small, for this language combination and text type, there seems to be a clear preference for human translations, which are considered better in more than half of the cases. Regarding MT engines, NMT presents more fluency and adequacy. This corresponds with the higher results in all automatic metrics. However, the results for the perception and automatic assessments do not correlate with PE time, even though there is a reduction in technical effort when post-editing NMT outputs. Thus, even though NMT produces more fluent results, this improvement does not always entail a reduction of the PE effort for professional translators, probably due to the added difficulty of error spotting in more fluent outputs.

In future research, we intend to further analyze PE, increasing the number of segments and language combinations to assess the correlation between automatic metrics and PE (technical and temporal) effort.

## Acknowledgements

We want to thank the copyright holders for granting permission for the MSD-Manuals website and for using these texts to create an English-Spanish parallel corpus. The training of the neural MT systems has been possible thanks to the NVIDIA GPU grant programme.

## References

- Allen, Jeffrey H. 2003. Post-editing. In Harold Sommer (ed.), *Computers and translation: A translator's guide*, 297–317. Amsterdam: John Benjamins. DOI: 10.1075/btl.35.19all.
- Aziz, Wilker, Sheila C. M. De Sousa & Lucia Specia. 2012. PET: A tool for post-editing and assessing machine translation. In *Proceedings of the eight international conference on language resources and evaluation (LREC'12)*, 3982–3987.
- Bahdanau, Dzmitry, Felix Hill, Jan Leike, Edward Hughes, Pushmeet Kohli & Edward Grefenstette. 2018. Jointly learning “what” and “how” from instructions and goal-states. In *6th international conference on learning representations, ICLR 2018, Vancouver, BC, Canada, April 30–May 3, 2018, workshop track proceedings*. OpenReview.net. <https://openreview.net/forum?id=BkmZvdkPM>.
- Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo & Marcello Federico. 2016. Neural versus phrase-based machine translation quality: A case study. In *Proceedings of the 2016 conference on empirical methods in Natural Language Processing*, 257–267. Austin, Texas: Association for Computational Linguistics. DOI: 10.18653/v1/D16-1025.
- Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia & Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Second conference on machine translation*, 169–214. <http://www.aclweb.org/anthology/W17-4717>.
- Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor & Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. *Proceedings of the First Conference on Machine Translation 2*. 131–198. <http://www.aclweb.org/anthology/W16-2301>.
- Bojar, Ondřej, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn & Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the third conference on machine translation*, 272–303. <http://aclweb.org/anthology/W18-6401.pdf>.
- Castilho, Sheila, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley & Andy Way. 2017. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics* 108(1). 109–120. DOI: 10.1515/pralin-2017-0013.

- Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilemini Sisoni, Yota Georgakopoulou, Pintu Lohar, Andy Way, Miceli Barone & Maria Gialama. 2017. A comparative quality evaluation of PBSMT and NMT using professional translators. In *Proceedings of MT summit XVI, vol. 1: Research track*, 116–131.
- Castilho, Sheila, Sharon O’Brien, Fabio Alves & Morgan O’Brien. 2014. Does post-editing increase usability? A study with Brazilian Portuguese as target language. In *Proceedings of the 17th annual conference of the European association for machine translation*, 183–190. Dubrovnik, Croatia: European Association for Machine Translation. <https://www.aclweb.org/anthology/2014.eamt-1.40>.
- Cho, Kyunghyun, Bart van Merriënboer, Dzmitry Bahdanau & Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8: Eighth workshop on syntax, semantics and structure in statistical translation*, 103–111. Doha, Qatar: Association for Computational Linguistics. DOI: 10.3115/v1/W14-4012.
- Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5). 378–382.
- Forcada, Mikel L., Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O’Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez & Francis M. Tyers. 2011. Apertium: A free/open-source platform for rule-based machine translation. *Machine translation* 25(2). 127–144.
- Germann, Ulrich, Eduard Barbu, M. Bentivoglio, Nikolay Bogoychev, C. Buck, D. Caroselli, L. Carvalho, A. Cattelan, R. Cattoni, Mauro Cettolo, Marcello Federico, Barry Haddow, David Madl, L. Mastrostefano, Prashant Mathur, A. Ruopp, A. Samiotou, V. Sudharshan, M. Trombetti & Jan van der Meer. 2016. Modern MT: A new open-source machine translation platform for the translation industry. *Baltic Journal of Modern Computing* 4. 397–397.
- Guerberof, Ana. 2009a. Productivity and quality in MT post-editing. In *Proceedings of MT Summit XII: Beyond translation memories: New tools for translators MT*, 1–9. Ottawa, Canada: AMTA. <http://www.mt-archive.info/MTS-2009-Guerberof.pdf>.
- Guerberof, Ana. 2009b. Productivity and quality in the post-editing of outputs from translation memories and machine translation. *The International Journal of Localisation* 7(1). 11–21. <https://www.tdx.cat/bitstream/handle/10803/90247/GuerberofThesis%20Final.pdf?sequence=1&isAllowed=y>.
- Isabelle, Pierre, Colin Cherry & George F. Foster. 2017. A challenge set approach to evaluating machine translation. *Computing Research Repository* abs/1704.07431. <http://arxiv.org/abs/1704.07431>.

## 2 Comparing NMT and PBSMT for post-editing in-domain formal texts

- Junczys-Dowmunt, Marcin, Tomasz Dwojak & Hieu Hoang. 2016. Is neural machine translation ready for deployment? A case study on 30 translation directions. *CoRR* abs/1610.01108. <http://arxiv.org/abs/1610.01108>.
- Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Xth MT summit*, vol. 5, 79–86.
- Koponen, Maarit. 2012. Comparing human perceptions of post-editing effort with post-editing operations. In *Proceedings of the seventh workshop on statistical machine translation (WMT '12)*, 181–190. Montréal, Canada: Association for Computational Linguistics. <https://www.aclweb.org/anthology/W12-3123>.
- Koponen, Maarit. 2016. Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. *The Journal of Specialised Translation* 25. 131–148.
- Krings, Hans P. 2001. *Repairing texts: Empirical investigations of machine translation post-editing processes*. Vol. 5. Kent, OH: Kent State University Press.
- Lommel, Arle & Donald A. DePalma. 2016. *Europe's leading role in machine translation: How Europe is driving the shift to MT*. Tech. rep. Boston. <http://cracker-project.eu>.
- Moorkens, Joss. 2018. Eye tracking as a measure of cognitive effort for post-editing of machine translation. In Walker Calum & Federico M. Federici (eds.), *Eye tracking and multidisciplinary studies on translation*, 55–70. Amsterdam. DOI: 10.1075/btl.143.04moo.
- Moorkens, Joss & Sharon O'Brien. 2013. User attitudes to the post-editing interface. In *Proceedings of machine translation summit XIV: Second workshop on post-editing technology and practice, Nice, France*, 19–25. <http://www.mt-archive.info/10/MTS-2013-W2-Moorkens.pdf>.
- Moorkens, Joss, Sharon O'Brien, Igor A. L. Da Silva, Norma B. De Lima Fonseca & Fabio Alves. 2015. Correlations of perceived post-editing effort with measurements of actual effort. *Machine Translation* 29(3). 267–284. DOI: 10.1007/s10590-015-9175-2.
- O'Brien, Sharon. 2011. Towards predicting post-editing productivity. *Machine Translation* 25(3). 197–215.
- Papineni, Kishore, Salim Roukos, Todd Ward & Wj Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In vol. July, 311–318. DOI: 10.3115/1073083.1073135.
- Parra Escartín, Carla & Manuel Arcedillo. 2015. A fuzzier approach to machine translation evaluation: A pilot study on post-editing productivity and automated metrics in commercial settings. In vol. 1, 40–45. <https://aclweb.org/anthology/W/W15/W15-4107.pdf>.

- Plitt, Mirko & François Masselot. 2010. A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague bulletin of mathematical linguistics* 93. 7–16. <https://ufal.mff.cuni.cz/pbml/93/art-plitt-masselot.pdf>.
- Sanchez-Torrón, Marina & Philipp Koehn. 2016. Machine translation quality and post-editor productivity. In *Proceedings of AMTA 2016*, 16–26. <https://researchspace.auckland.ac.nz/handle/2292/31486>.
- Specia, Lucia. 2010. Combining confidence estimation and reference-based metrics for segment-level MT evaluation. In *The ninth conference of the association for machine translation in the Americas*. <https://amta2010.amtaweb.org/AMTA/papers/2-03-BanerjeeDuEtal.pdf>.
- Specia, Lucia. 2011. Exploiting objective annotations for measuring translation post-editing effort. In *Proceedings of the 15th conference of the European association for machine translation*, 73–80. <http://www.mt-archive.info/EAMT-2011-Specia.pdf>.
- Toral, Antonio & Víctor M. Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th conference of the European chapter of the Association for Computational Linguistics*, vol. Volume 1: Long papers, 1063–1073. Valencia, Spain: Association for Computational Linguistics.
- Villegas, Marta, Ander Intxaurre, Aitor Gonzalez-Agirre, Montserrat Marimon & Martin Krallinger. 2018. The MeSpEN resource for English-Spanish medical machine translation and terminologies: Census of parallel corpora, glossaries and term translations. *Language Resources and Evaluation* 52. 32–39.
- Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes & Jeffrey Dean. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. *CoRR* abs/1609.08144. <http://arxiv.org/abs/1609.08144>.

## Chapter 3

# German light verb construction in the course of the development of machine translation

Shaimaa Marzouk

Johannes Gutenberg University Mainz

The German light verb construction (LVC) is commonly used despite its relative complexity. Different writing guidelines recommend avoiding LVCS and replacing them with the base verb constructions (BVCs). However, since not every LVC has an equivalent BVC, replacement is not always possible. The present study addresses two aspects: first, how the machine translation (MT) of LVC has evolved in light of recent progress in MT and the increasing dominance of neural machine translation (NMT), and second, whether the use of BVCs improves MT output compared to LVCS. The analysis of the MT output of both scenarios, LVC and BVC, is performed for different MT approaches in terms of number and types of MT errors, style and content quality ratings, and scores from two automatic evaluation metrics (AEMS). For this, a mixed-methods triangulation approach that includes error annotation, human evaluation, and automatic evaluation was applied and five MT systems were examined: a rule-based system (RBMT), a statistical system (SMT), two differently constructed hybrid systems (HMT), and a neural system (NMT). The study is conducted for the language pair German-to-English in the technical domain. The results show that systems that employ earlier MT approaches (RBMT, SMT, HMT) benefited from replacing the LVC with the corresponding BVC as their output was improved (i.e., MT errors were reduced; quality and AEMS scores were increased). On the contrary, the NMT system was able to produce MT with minimal number of errors both for LVCS and BVCs and recorded the highest quality levels in both scenarios among the analyzed MT systems.



## 1 Introduction

The German term *Funktionsverbgefüge* was coined by von Polenz (1963: 26); the English counterpart, *function verb constructions* or *light verb constructions*, hereafter LVCS, goes back to the linguist Jespersen (1942: 117). With this, Jespersen (ibid.) distinguishes between a light verb and a heavy verb (a.k.a. *full verb*, a verb that emphasizes the full meaning). Some examples of LVCS are: *eine Frage stellen* ('to pose a question'), where *stellen* is a light verb, as opposed to *fragen* ('to ask'), which acts as a full verb; *eine Handlung ausführen* ('to perform an action') to replace the full verb *handeln* ('to act'); *etw. zu Papier bringen* ('to put sth. on paper') instead of *schreiben* ('to write'), and *eine Entscheidung treffen* ('to make a decision') instead of *sich entscheiden* ('to decide').

As illustrated by these examples, an LVC is simply a combination of a verb and a noun that can only be correctly understood with both components. Strictly speaking, it is a complex predicate that consists of a semantically light verb and a deverbal noun (Jespersen 1964: 117). The verb in the LVC acts merely as a functional element, letting the noun represent the main predicate (Grimshaw & Mester 1988). The LVC is not just found in German, but in many other languages as well. In English, *make a decision* is sometimes used instead of *decide*. Similarly, in Arabic, one might say *yakhudh qrarāan* ('make a decision') or *yuqrīr* ('decide').<sup>1</sup> Both variants also exist in Spanish *tomar una decisión* ('make a decision') and *decidir* ('decide').

The present study focuses on German LVCS that can take on one of the following forms: a verb plus a noun in the accusative case (e.g., *eine Handlung ausführen*) or a verb plus a prepositional phrase (e.g., *zu Papier bringen*). German LVCS are used predominantly in technical, scientific, legal, and official texts (Bruker 2013: 38f.), but despite their widespread use, they are criticized both in linguistics and translation. In linguistics, they are seen as a sign of "Umschreibungssucht" (addiction to reformulating) and "Verbaphobie" (verbaphobia) (Daniels 1963: 9f.) and have been described as "unnecessarily complicated" and "inelegant" (Storrer 2006). Because of the relative complexity of LVCS, several Controlled Language varieties and writing guidelines prompt writers to avoid them: (1) The rule "Avoid light verb constructions" is found in *Leichte Sprache* (Easy German Language), which is increasingly being applied to simplify legal, political, and administrative texts for people with low language skills or cognitive limitations (Hansen-Schirra & Gutermuth 2018). Here, the rule is included to reduce sentence complexity (Bredel & Maaß 2016). (2) The rule is also applied in Controlled

---

<sup>1</sup>The Arabic examples were transliterated by <https://de.glosbe.com/transliteration/Arabic-Latin>.



### 3 German light verb construction in the course of the development of MT

Languages used in technical documentation in order to keep sentences more concrete and direct (Gesellschaft für Technische Kommunikation, Tekom e.V. 2013: 107). (3) The same rule is present in the guidelines from the weekly German magazine *Die Zeit* entitled *Recommendations for prospective journalists*, which advise journalists to use full verbs instead of LVCS, as full verbs are usually clearer and more efficient (Die Zeit 2007).

However, despite their structural complexity, LVCS are widely used. This is partially due to the fact that some LVCS completely lack any equivalent, e.g., *in Ordnung halten* ('to keep in order'). Other LVCS have a more nuanced meaning that can be difficult to express using the base verb construction, hereafter BVC. One example of this kind of LVC is *eine Maschine in Betrieb setzen* ('to put a machine into operation'); this is a process that usually includes several different procedures depending on the complexity of the machine and is therefore much more than simply (p. 107) *eine Maschine einschalten* ('to turn on a machine') (Baumert & Verhein-Jarren 2012). Concretely, the LVCS can influence meaning in four ways, which are known as "action types" (Zifonun, Hoffmann & Strecker 1997: 704):

*Causative*: emphasize the initiator of an action, e.g., *der Starter setzt den Motor in Gang* ('the starter sets the engine in motion').

*Inchoative*: mark the beginning of an action, e.g., *endlich geht das Buch in Druck* ('finally, the book goes to press').

*Durative*: emphasize the duration of an action, e.g., *ein neues Modell ist bereits in Arbeit* ('a new model is already in production').

*Passive*: form a distinct passive meaning variation, e.g., *die neue Methode findet Anwendung bei dem Versiegelungsprozess* ('the new method is applied in the sealing process').

## 2 Machine translation of LVCS

As discussed, the usage of the LVCS can be indispensable in conveying a distinct nuance of meaning or because there is no BVC equivalent. Despite the existence of LVCS in several languages, there are a number of difficulties in MT of LVCS. Heine (2017) describes LVCS as "a typical example of phenomena that are neither explainable with (exclusively) grammatical rules nor lexical units" and how the sentence syntax as well as the lexical components of the MT system are decisive for an error-free MT output. Therefore, depending on the complexity of the sentence syntax and the MT system approach as well as the system capacity, the

primary challenge of translating an LVC is that the parser must first identify it as such. The system needs to be able to distinguish between *stellen* as a full verb, as in *etwas auf den Tisch stellen* ('to put something on the table') as opposed to *stellen* as a light verb in *etwas zur Verfügung stellen* ('to make something available'), *eine Frage stellen* ('to ask a question'), or *etwas in Rechnung stellen* ('to invoice something').

After identifying the LVC in the source language, a transfer problem between the source and target language may appear. Depending on the language pair, the LVC might be (best) translated using an equivalent LVC, a different function verb, a BVC, or a completely different construction (Bruker 2013: 96), e.g., translating *zur Verfügung stehen* as 'are available'. In addition, a syntactic translation problem may arise while translating LVCS with prepositional phrases that have no articles or with a preposition that can require different cases, such as *in*, *an*, *auf*, or *unter*, e.g., *in Betrieb nehmen* (accusative) vs. *in Betrieb bleiben* (dative) or *in Verhandlungen treten* (accusative) vs. *in Verhandlungen stehen* (dative). Such cases cannot be strictly morphologically differentiated. The correct case for the output of the syntax information for the nominal phrase in the LVC can probably only be determined by using appropriate lexicon entries for the function verb, as the verb selects the case of the prepositional phrase. (ibid.: 75) Another potential problem can be encountered on a morpho-syntactic level in processing LVCS that include compounds, e.g., *Verstellung vornehmen* in *Höhenverstellung vornehmen* or *Behandlung durchführen* in *Fleckenbehandlung durchführen*. In such cases, the LVC with the compound must first be morpho-syntactically analyzed and broken down into its component parts (Winhart 2005). For this, an exact semantic analysis of the compound is required for a correct processing of the LVC (Bruker 2013: 97).

The difficulties in the MT of LVCS as well as their frequent use in the German language make its relevance for Natural Language Processing evident. Nonetheless, the LVC has not yet received the attention it needs in computational linguistics, particularly in MT research. There is a number of linguistic studies that closely investigate the linguistic differences between LVCS and BVCS on the basis of corpora (Glatz 2006; Storrer 2007; 2006). Storrer (2006) shows how the influence of both constructions goes beyond their different pragmatic and stylistic impacts. Others investigated the properties of multiword predicates and developed automatic methods for distinguishing among literal, metaphorical, and idiomatic multiword predicates (Fazly & Stevenson 2005). North (2005) examined the productivity of LVCS that include predicative nouns and developed computational measures for quantifying the acceptability of LVCS. Kuhn (1994) analyzed how

the HPSG-based translation approach<sup>2</sup> handles LVCS. Marzouk & Hansen-Schirra (2019) analyzed the impact of avoiding LVCS among other German Controlled Language rules on machine translatability across different MT approaches and found out that the NMT system delivers in comparison to RBMT, SMT and hybrid MT systems mostly error-free output both before and after the application of the rules showing even a decrease in quality after applying the rules. Further studies on MT of German LVCS across different MT approaches including the NMT have not yet been conducted, to the best of my knowledge. In light of the proven linguistic differences between LVCS and BVCS (Glatz 2006; Storrer 2007; 2006) and the success achieved by NMT in improving MT output compared to earlier approaches (Bentivogli et al. 2016; Marzouk & Hansen-Schirra 2019; Popović 2018; Toral & Sánchez-Cartagena 2017), this study aims to track MT's progress in translating LVCS.

The remainder of this paper is organized as follows: Section 3 provides an overview of the empirical study including the dataset and the MT systems used. Section 4 outlines the methodology applied. Results are presented in Section 5 followed by a discussion in Section 6. Finally, Section 7 provides the conclusion, mentions the limitations of the study, and gives an overview of future work.

### 3 Description of study

The study analyzes two aspects of MT with regard to LVCS: (1) to what degree different MT approaches are able to translate LVCS and (2) whether the use of BVCS improves the MT compared to using LVCS. In the analysis, MT outputs of LVCS and BVCS are contrasted across four MT approaches, and the impact of each construction is measured in terms of number and types of MT errors, style and content quality ratings, and AEMS scores. The examined MT approaches are represented by five MT systems: Google Translate (an NMT system), Lucy LT KWIK Translator (an RBMT system), SDL Free Translation (an SMT system), and Bing by Microsoft and Systran (two differently constructed HMT systems).<sup>3</sup> The selection criteria of the systems were (1) to be an online freely available system, (2) to offer the language pair German-to-English, and (3) to cover different MT approaches.

For the analysis, a test suite was constructed that consists of 24 source sentences extracted from a corpus of German technical user manuals using the Con-

---

<sup>2</sup>HPSG: Head-driven Phrase Structure Grammar (Pollard & Sag 1994) was considered the best available grammar formalism at that time.

<sup>3</sup>The MT step was performed at the end of 2016. At that time, Bing became an HMT system by adding language-specific rule components to its original SMT system, and Systran was also developed from an RBMT system into a hybrid system.

trolled Language checker CLAT (Rösener 2010).<sup>4</sup> The 24 analyzed LVCS were as follows: twelve accusative LVCS and twelve prepositional LVCS. After reformulating the sentences using the BVCS, both versions (LVCS and BVCS) were machine translated into English using the aforementioned five MT systems, resulting in a dataset of 240 MT sentences (24 source sentences × 2 versions × 5 systems). In the source sentences, company-specific and specialist terms were replaced with common terms (e.g. *Gerät* instead of *Feinstzerkleinerer*; *Steckdose* instead of *Schutzkontaktsteckdose*). This modification was necessary for two reasons: (1) the MT systems used in the study were not trained in advance with specific relevant corpora; (2) to avoid human evaluators investing too much time investigating the translation of these types of uncommon terms during the human evaluation. Source sentences that included more than two specific terms were excluded entirely from the analysis to avoid application of multiple changes to the original source sentences.

## 4 Methodology

A mixed methods triangulation approach was applied that incorporates three evaluation methods: error annotation, human evaluation, and automatic evaluation. The analyses were conducted in a black box context, as the focus is on the comparison of the MT outputs of the LVC and BVC scenarios (and not on the internal processes of the systems). In the following, the analyses are demonstrated in detail.

### 4.1 Error annotation

The goal of the error annotation is to identify the MT errors in the use of LVCS (LVC scenario) and BVCS (BVC scenario) and compare them in terms of their number and type. The annotation was conducted by a qualified experienced German–English translator and checked by two professional German–English translators. Further, based on the existence or non-existence of MT errors, the data were divided into four groups, referred to as “annotation groups”. These are: FF (for false–false): translation contains error(s) in both scenarios; FR (for false–right): translation contains error(s) only in the LVC scenario; RF (for right–false): translation contains error(s) only in the BVC scenario; RR (for right–right): no errors

---

<sup>4</sup>CLAT is one of the most well-known Controlled Language checkers in Germany developed by the society for the promotion of applied information sciences (IAI) at Saarland University; see: <http://www.iai-sb.de/de/produkte/clat>.

### 3 German light verb construction in the course of the development of MT

in either scenario. The error classification applied is mainly based on Vilar et al. (2006) and encompasses the error types shown in Table 3.1.<sup>5</sup>

Table 3.1: Error classification applied in the annotation

Category	No.	Type
Orthography	OR.01	Punctuation error
	OR.02	Capitalization error
Lexis	LX.03	Omission
	LX.04	Addition
	LX.05	Untranslated
	LX.06	Consistency error (a word is repeated in the sentence and translated differently each time)
Grammar	GR.07	Wrong word class
	GR.08	Wrong verb tense / composition / person
	GR.09	Wrong agreement gender / number / person
	GR.10	Wrong word order
Semantics	SM.11	Confusion of sense (output translation is possible, but not in the given context)
	SM.12	Wrong choice (output translation is apparently wrong)
	SM.13	Collocation error

The error taxonomy of Vilar et al. (2006) was used as a basis for the error annotation due to its explicitness, clarity and appropriate degree of granularity. However, further more extensive taxonomies, such as the multidimensional quality metrics (MQM) framework can be also used for the analysis. This would be particularly useful in case of examining fine-grained or more specific types of errors.

#### 4.2 Human evaluation

The goal of the human evaluation is to compare the content and style quality of the MT in the LVC and BVC scenarios. Following the quality definition of Hutchins & Somers (1992), the *content quality* is the extent to which the translation reflects

<sup>5</sup>As the analysis of the LVCs and BVCs was part of a large-scale study that aimed to examine different German Controlled Language rules, it was necessary to add two further relevant error types to Vilar et al. (2006)'s taxonomy (capitalization and consistency) and to exclude two error types in Vilar et al. (ibid.) that were irrelevant for the study (idioms and style).

the information in the source text accurately; and the extent to which the translation is easy to understand (ibid.). The *style quality* is the extent to which the translation sounds natural and idiomatic in standard written English, is appropriate for the intention of its content (ibid.) as well as presented clearly in terms of orthography. The definition covers the orthography as an instrument for presenting the content in an adequate way that serves its intention.

Based on these definitions, the content quality (CQ) covers the criteria accuracy and clarity; the style quality (SQ) encompasses the criteria idiomaticity, appropriateness to the content intention as well as correctness and clarity of the orthographic presentation.

The human evaluation Figure 3.1 consisted of (1) evaluating the SQ and CQ of the MT (\*) on two 5-point Likert scales; (2) selecting the relevant quality criteria that justify the assigned quality scores: accuracy and clarity under the CQ; idiomaticity, appropriateness to the content intention as the content well as correctness and clarity of the orthographic presentation under the SQ; (3) providing the word or part of the translation relevant to each chosen criterion; (4) where many modifications were necessary, the participant had to enter an alternative translation for the whole sentence.

Ist die Seriennummer des Gerätes bekannt, kann im Feld Seriennummer diese Nummer eingegeben werden. (\*)  
 Is the serial number of the device is known, this number can be entered in the "Serial Number" field.

<p><b>Style Quality</b></p> <p style="font-size: small;">very low <span style="float: right;">very high</span></p> <p style="text-align: center;">○ 1   ○ 2   ○ 3   ○ 4   ○ 5</p>	<p><b>Content Quality</b></p> <p style="font-size: small;">very low <span style="float: right;">very high</span></p> <p style="text-align: center;">○ 1   ○ 2   ○ 3   ○ 4   ○ 5</p>
---	---

(1)

<p><b>Style Quality</b></p> <p><input type="checkbox"/> I have an alternative translation that is <b>presented correctly</b> or (more) <b>clearly</b>, i.e. <b>orthographically</b></p> <hr/> <p><input type="checkbox"/> I have an alternative translation that is (more) <b>appropriate to the intention of the sentence</b>, e.g. <b>motivates the user to act, draws the user's attention, etc.</b></p> <hr/> <p><input type="checkbox"/> I have an alternative translation that <b>sounds (more) natural and idiomatic</b></p>	<p>Reason - word/part of the translation caused this problem. If applicable, please delete this text and replace it with: "word/part" is "...reason"... I suggest "..." instead.</p> <hr/> <p>Reason - "word/part" is "...reason"... I suggest "..." instead. (replace, if applicable)</p> <hr/> <p>Reason - "word/part" is "...reason"... I suggest "..." instead. (replace, if applicable)</p>
---	--

(2) (3)

<p><b>Content Quality</b></p> <p><input type="checkbox"/> I have an alternative translation that <b>reflects the information</b> in the source text (more) <b>accurately</b></p> <hr/> <p><input type="checkbox"/> I have an alternative translation that is <b>easier to understand</b>, i.e. <b>better formulated and/or presented</b></p>	<p>Reason - "word/part" is "...reason"... I suggest "..." instead. (replace, if applicable)</p> <hr/> <p>Reason - "word/part" is "...reason"... I suggest "..." instead. (replace, if applicable)</p>
--	---

Many modifications are necessary; I have the following **alternative translation** for the whole sentence:  
 Please replace this text with your alternative translation!

(4)

Figure 3.1: Interface of the human evaluation

Concerning the participants, different studies recommend recruiting more than 3–4 participants (Fiederer & O’Brien 2009). In this study, five participants initially conducted the tests and the number of participants was successively increased until the accumulated average of the quality values stabilized. After the eighth participant, the accumulated quality averages hardly changed. Accordingly, the number of participants was not increased anymore. The participants

### 3 German light verb construction in the course of the development of MT

are native English speakers and hold a bachelor's degree in translation. In addition, all participants were students in the last or penultimate semester of the master's degree program in translation. Participation was remunerated.

Regarding the test procedure, the analysis of the LVCS and BVCS was part of a large-scale study that aimed to examine different German Controlled Language rules (Marzouk & Hansen-Schirra 2019). Within the scope of the study, each participant evaluated in total 1,100 MT sentences that were randomized and split into 44 tests (the analysis of the LVC vs. BVC was a subset of this dataset). Each participant had the opportunity to choose whether to rate one, two or three tests per day, depending on his or her availability. The basic requirement was to evaluate at least one test daily, thus avoiding interruptions that could possibly have a negative effect on the intra-rater agreement. In addition, the participants were asked to take a break between the tests. The 44 tests were sent in a different randomized order to the participants (e.g. the 1st participant received test 40, test 8, test 5 consecutively). A decreasing motivation over a 3–4 week evaluation period is unavoidable. Therefore, this randomization ensured that no particular sentences were evaluated by all participants at the end of the evaluation. The tester received the answered tests every day and checked them for completeness (i.e. all sentences were rated and commented on if necessary). In case of any missing data, the participant was asked to complete them, then he or she received the new tests for the next day.

### 4.3 Automatic evaluation

The alternative translation obtained from the human evaluation acted as a reference translation for the automatic evaluation metrics (AEMS) in order to compare their scores in the LVC and BVC scenarios. Two reference translations per sentence were randomly selected for the comparison. The study applied the evaluation metrics TERbase and hLEPOR. The former is a basic edit distance metric that calculates the minimum number of edits needed to change the evaluated MT so that it exactly matches the reference translation and works without stemming, synonymy lookup and paraphrase support (Snover et al. 2006; González & Giménez 2014). It was necessary to consider the use of synonyms as an edit, as the participants quite often recommended the use of a certain synonym while evaluating the translation accuracy. At the same time, hLEPOR was applied as one of the advanced metrics that has proven to have a state-of-the-art correlation with human evaluation compared with metrics like BLEU, TER, and METEOR among others (Han et al. 2013). The calculation model of hLEPOR is based on three factors: an enhanced length penalty, an  $n$ -gram position difference penalty and the harmonic mean of precision and recall (ibid.).

## 5 Results

### 5.1 Analysis of the annotation groups (FF, FR, RF, RR) based on the error annotation

Comparing the LVC and BVC scenarios (Figure 3.2) showed that 42% of the sentences were translated correctly in both scenarios (group RR), while half of this percentage (21%) was translated incorrectly in both scenarios (group FF). At the same time, 29% of the sentences were translated incorrectly while using the LVCs and correctly after using the BVCs (group FR). On the other hand, 8% were only translated incorrectly while using the BVCs (group RF).

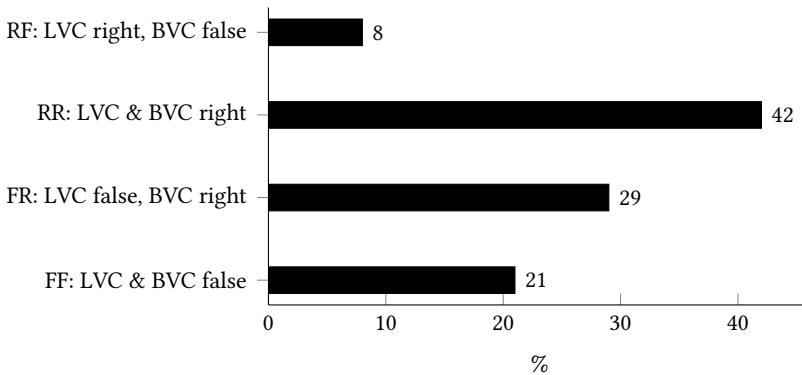


Figure 3.2: Distribution of annotation groups for all the MT systems

Based on the existence and non-existence of MT errors, the impact of using the BVC instead of the LVC on the MT output cannot be considered effectively positive. The only positive impact can be observed in the FR group (false in case of LVC – right in case of BVC). This group amounts to 29%. At the same time, the groups RF and FF together amount to 29%: In RF (right in case of LVC – false in case of BVC), there is a clear negative impact of using the BVC and in FF the usage of the BVC did not help produce an error-free MT.

Considering the groups RR and FF, since the translations were both in the LVC scenario and the BVC scenario correct (RR group) or incorrect (FF group), a positive impact of a certain scenario can only be justified if its quality values in these two groups were higher. In order to explore quality changes in each annotation group, the results of the error annotation and human evaluation were triangulated. The triangulated results showed no significant quality changes in the RR and FF groups. The only significant quality change was in a few cases of the



### 3 German light verb construction in the course of the development of MT

group FR, indicating that getting an incorrect MT of the LVC and a correct MT of the BVC led to significantly higher quality in case of the BVC.

#### 5.2 Analysis of the error types

On the semantic level, the three semantic error types SM.11 confusion of sense, SM.12 wrong choice and SM.13 collocation error were affected in both scenarios. However, a significant change in the number of errors was only observed in error type SM.13 collocation error; this decreased significantly after replacing the LVC with a BVC. Furthermore, in few cases, the grammatical error types GR.08 wrong verb and GR.10 wrong word order and the lexical error type LX.04 addition were differently affected in both scenarios without showing a significant increase or decrease in a certain scenario. The remaining error types were not relevant.

#### 5.3 Analysis of the quality changes based on the human and automatic evaluations

Although the analysis of the annotation groups did not reflect a substantial quality increase after using the BVC except for the aforementioned significant quality change in group FR, a significant increase in the MT quality in terms of style and content quality (SQ and CQ) as well as AEMS scores was detected based on the human and automatic evaluations where the BVC was used.

Furthermore, the Spearman test was conducted to investigate the correlation between the difference in the overall quality<sup>6</sup> and the differences in the AEMS scores in both scenarios. The test showed a significant positive strong correlation ( $\rho > 0.5$ ,  $p < 0.001$ ). Accordingly, the quality changes detected in both analyses (human and automatic evaluation) were in line with each other.

#### 5.4 Comparison of the impact of replacing the LVC with the BVC at MT system level

So far, the results show that the MT of BVCS had a significant higher quality in terms of human scores of the SQ and CQ as well as AEMS scores. Subsequently, an analysis at MT system level was conducted in order to explore which MT systems exhibited these higher quality levels. The general positive impact of using BVCS instead of LVCS on the MT output at system level is shown in Figure 3.3 and Figure 3.4.

---

<sup>6</sup>The overall quality is the mean of SQ and CQ, as analyzing the correlation here requires no distinction between the quality parameters.

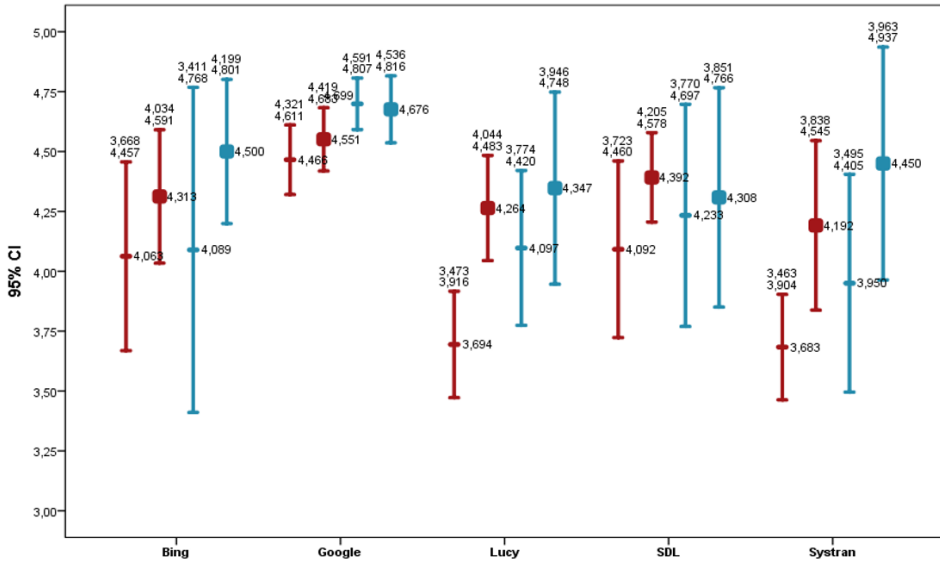


Figure 3.3: Style and content quality in case of using LVCs as opposed to BVCs

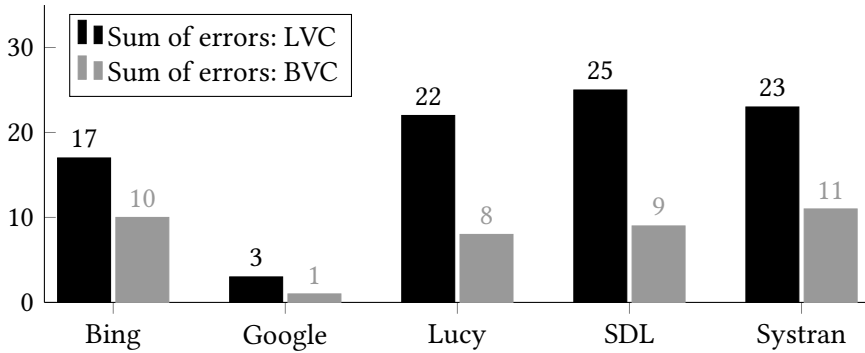


Figure 3.4: Number of MT errors in case of using LVCs as opposed to BVCs

### 3 German light verb construction in the course of the development of MT

For the RBMT system (Lucy) and one hybrid system (Systran), using the BVC was very advantageous in reducing the number of errors and increasing SQ significantly. In the other hybrid system (Bing) and the SMT system (SDL), the number of errors decreased and the SQ and CQ increased after using the BVC; however, the changes were not significant. The NMT system (Google Translate) showed distinct results: the number of errors was minimal (three errors in the LVC scenario; one error in the BVC scenario). GNMT was able to translate 88% of the sentences in both scenarios correctly, followed by Bing with 46%, and recorded the highest SQ and CQ among all systems in both scenarios as well.

#### 5.5 Correlation between the error types and the quality values

The earlier MT approaches showed the following significant strong correlations between a decreased number of errors of the different error types and increased quality values when using a BVC: In Lucy, the decrease in the semantic errors SM.11 confusion of sense and SM.12 wrong choice correlated with the increase in SQ ( $\rho = -0.521, p = 0.027$ ) and CQ ( $\rho = -0.537, p = 0.021$ ) respectively. In Bing, there was a correlation between the error type LX.03 omission and CQ ( $\rho = -0.565, p = 0.035$ ). In SDL, the correlation was observed between each of the error types LX.04 addition and GR.10 wrong word order and the SQ (for LX.04:  $\rho = -0.594, p = 0.020$ ; for GR.10:  $\rho = -0.641, p = 0.010$ ) as well as between each of the error types LX.04 addition and SM.12 wrong choice and the CQ (for LX.04:  $\rho = -0.646, p = 0.009$ ; for SM.12:  $\rho = -0.593, p = 0.020$ ). Finally, in Systran, the error type GR.07 wrong word class correlated with the CQ ( $\rho = -0.511, p = 0.018$ ).

## 6 Discussion

The results show that using BVCS instead of LVCS enhanced the MT of the systems that apply earlier MT approaches (RBMT, SMT, and HMT). It was observed that BVCS simplified the sentence structure and provided an equivalent for German LVCS, which do not have an English counterpart. This section discusses some examples and contrasts the output of the earlier MT approaches with that of the NMT approach in order to gain a deeper insight into the quantitative results.

The first LVC *Höhenverstellung vornehmen* (example 1 in Table 3.2) poses two challenges for MT: including the compound *Höhenverstellung* and having no counterpart for *Verstellung vornehmen* in English. The usage of the BVC *Höhe verstellen* led to breaking down the compound *Höhenverstellung* and solved the collocation problem in English for the RBMT system Lucy. Concretely, it was associated with a correction of the collocation error (SM.13) and thus facilitated producing an error-free MT.

In example 2 in Table 3.3 and example 3 in Table 3.4, the LVCs include prepositional phrases. The LVC *zur Verfügung stehen* is a common German LVC. Although the SMT system SDL was able to parse it correctly, the MT included a wrong word order error (GR.10). The usage of the BVC simplified the sentence structure and was associated with a correction of the word order. The LVC in example 3 (Table 3.4) *zur Anwendung kommen*, on the contrary, is not as common as *zur Verfügung stehen* and was associated with a wrong verb error (GR.08) in the MT of the HMT system Systran. This error was corrected when the BVC was used.

In translating the LVC *zur Verfügung stellen* in example 4 (Table 3.5), the HMT system Bing exhibited semantic and lexical difficulties: a wrong choice error (SM.12) in ‘represents’ and an addition error (LX.04) in ‘available’. Such semantic and lexical errors occur when the system translates the LVC literally (e.g., translating *zur Verfügung stellen* as ‘represent available’ instead of ‘provide’). Using the BVC resolved these MT difficulties and was associated with a correction of both errors.

According to the human evaluation, correcting the MT errors in the abovementioned examples made the translation more appropriate for its intention, more attention-grabbing, and easier to understand, which led to the enhancement of the SQ and CQ.

While systems that apply earlier approaches were not able to identify the LVC in the source language as such in some cases and in other cases faced different transfer problems in the translation from German to English, GNMT was able to overcome these difficulties and handle all the aforementioned MT issues that the other systems encountered. As a result, GNMT produced translations with a minimal number of errors, if any, and recorded the highest SQ and CQ levels both in the LVC and the BVC scenarios.

## 7 Conclusion

The German LVC is a relatively complex construction on both a linguistic and translational level. In this study, I analyzed to which degree different MT approaches (RBMT, SMT, HMT, and NMT) are able to translate the LVC, and whether replacing LVCs with BVCs improves the MT output. The analysis was conducted based on a comparison of the number and types of MT errors, style and content quality ratings, and AEMS scores in the LVC vs. BVC scenario for five MT systems. The study focused on the target language English in the technical domain.

### 3 German light verb construction in the course of the development of MT

Table 3.2: Example 1. The LVC and BVC are presented in **bold**. *Italic* is used for correct tokens of the translation; underlining for the incorrect tokens.

LVC	Die Höhen <b>verstellung</b> der Fronten können Sie mittels eines Schraubendrehers <b>vornehmen</b> .
Lucy	You can <u>carry out</u> the height <i>adjustment</i> of the fronts using a screwdriver.
GNMT	You can <i>adjust</i> the height of the fronts using a screwdriver.
BVC	Die <b>Höhe</b> der Fronten können Sie mittels eines Schraubendrehers <b>verstellen</b> .
Lucy	You can <i>adjust</i> the height of the fronts using a screwdriver.
GNMT	The height of the fronts can be <i>adjusted</i> by means of a screwdriver.

Table 3.3: Example 2

LVC	Auf der Startseite <b>stehen</b> die folgenden Funktionen zur Auswahl <b>zur Verfügung</b> .
SDL	On the Start page, <u>are</u> the following functions <i>available</i> to choose from..
GNMT	The following functions <i>are available</i> for selection on the start page.
BVC	Auf der Startseite <b>sind</b> die folgenden Funktionen zur Auswahl <b>vorhanden</b> .
SDL	On the Start page, the following functions <i>are available</i> to choose from.
GNMT	The following functions <i>are available</i> for selection on the start page.

Table 3.4: Example 3. The LVC and BVC are presented in **bold black**. *Italic* is used for correct tokens of the translation; underlining for the incorrect tokens.

LVC	Somit kann die Fluggesellschaft nicht garantieren, dass die Gepäckregeln immer <b>zur Anwendung kommen</b> .
Systran	Thus, the airline cannot guarantee that the baggage rules always <u>apply</u> .
GNMT	Thus, the airline cannot guarantee that the baggage rules <i>are</i> always <i>applied</i> .
BVC	Somit kann die Fluggesellschaft nicht garantieren, dass die Gepäckregeln immer <b>angewendet werden</b> .
Systran	Thus, the airline cannot guarantee that the baggage rules <i>are</i> always <i>applied</i> .
GNMT	Thus, the airline cannot guarantee that the baggage rules <i>are</i> always <i>applied</i> .

Table 3.5: Example 4. The LVC and BVC are presented in **bold black**. *Italic* is used for correct tokens of the translation; underlining for the incorrect tokens.

LVC	Der Navigationsbaum <b>stellt</b> alle vorhandenen Seiten der Konfigurierung <b>zur Verfügung</b> .
Bing	The navigation tree <u>represents</u> all existing pages of the configuration <u>available</u> .
GNMT	The navigation tree <i>provides</i> all existing pages of the configuration.
BVC	Der Navigationsbaum <b>stellt</b> alle vorhandenen Seiten der Konfigurierung <b>bereit</b> .
Bing	The navigation tree <i>provides</i> all the existing configuration pages.
GNMT	The navigation tree <i>provides</i> all existing pages of the configuration.

### 3 German light verb construction in the course of the development of MT

The results of the earlier MT approaches (RBMT, SMT, and HMT) confirmed the complexity of LVCS on a translational level: the MT of LVCS was more error-prone, and the MT quality (SQ, CQ, and AEMS scores) increased with the usage of BVCs. For the RBMT, SMT, and HMT systems, if there were an equivalent BVC for each LVC, the MT problem would be eliminated. However, not all LVCS have an equivalent BVC. In addition, an LVC is, in some cases, needed to express a certain nuanced meaning that the BVC cannot convey as effectively. Since the LVC cannot always be avoided, there is a need to translate it properly. According to the results, the NMT approach provides a capable architecture that can handle the complexity of LVCS: GNMT system was able to translate 88% of the sentences correctly in both the LVC and the BVC scenarios. This was followed by Bing's mere 46%. GNMT system also recorded the highest SQ and CQ values of all systems (> 4.4 out of 5 points) in both scenarios. Therefore, using an NMT system, such as GNMT, allows for the flexibility to choose between LVC and BVC. This, in turn, gives room for the author to prioritize sentence semantics and focus more on the pragmatics.

This study has explored the MT of German LVC for different MT architectures, including NMT, which – to the best of my knowledge – has not yet been examined. However, the following limitations should be mentioned: The study was conducted only for one target language. Although the number of the source sentences was not high, the sentences were translated by five different MT systems, and the MT output was evaluated by eight subjects. In future work, I plan to explore how the NMT architecture tackles further common complex constructions in German based on a corpus analysis of different target languages.

## Abbreviations

LVC	light verb construction	SQ	style quality
BVC	base verb construction	CQ	content quality

## References

- Baumert, Andreas & Annette Verhein-Jarren. 2012. *Texten für die Technik*. Berlin, Heidelberg: Springer Vieweg. DOI: 10.1007/978-3-662-47410-5.
- Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo & Marcello Federico. 2016. Neural versus phrase-based machine translation quality: A case study. In *Proceedings of the 2016 conference on empirical methods in Natural Language Processing*, 257–267. Austin, Texas: Association for Computational Linguistics. DOI: 10.18653/v1/D16-1025.

- Bredel, Ursula & Christiane Maaß. 2016. *Leichte Sprache: Theoretische Grundlagen. Orientierung für die Praxis* (Sprache im Blick). Berlin: Dudenverlag.
- Bruker, Astrid. 2013. *Funktionsverbgefüge im Deutschen: Computerlexikographische Probleme und Lösungsansätze*. Hamburg: Bachelor Master Publ.
- Daniels, Karlheinz. 1963. *Substantivierungstendenzen in der deutschen Gegenwartssprache: Nominaler Ausbau des verbalen Denkkreises* (Sprache und Gemeinschaft). Düsseldorf: Schwann.
- Die Zeit. 2007. *Der Weg zum journalistischen Schreiben*. Tech. rep. 4. <https://gymwil.files.wordpress.com/2007/10/zeit-journalistentipps5.pdf>.
- Fazly, Afsaneh & Suzanne Stevenson. 2005. Automatic acquisition of knowledge about multiword predicates. In vol. 1, 31–42. ACL.
- Fiederer, Rebecca & Sharon O'Brien. 2009. Quality and machine translation: A realistic objective? *The Journal of Specialised Translation* 11(11). 52–74. [http://www.jostrans.org/issue11/art\\_fiederer\\_obrien.pdf](http://www.jostrans.org/issue11/art_fiederer_obrien.pdf).
- Gesellschaft für Technische Kommunikation, Tekom e.V. 2013. *Leitlinie "Regelbasiertes Schreiben" Deutsch für die Technische Kommunikation*. Tech. rep. Stuttgart.
- Glatz, Daniel. 2006. Funktionsverbgefüge – semantische Doubletten von einfachen Verben oder mehr? In Kristel Proost, Gisela Harras & Daniel Glatz (eds.), *Domänen der Lexikalisierung kommunikativer Konzepte*, 129–178. Tübingen: Narr.
- González, Meritxell & Jesús Giménez. 2014. *An open toolkit for automatic machine translation (meta-) evaluation*. Tech. rep. [http://asiya.lsi.upc.edu/Asiya\\_technical\\_manual\\_v3.0.pdf](http://asiya.lsi.upc.edu/Asiya_technical_manual_v3.0.pdf).
- Grimshaw, Jane & Armin Mester. 1988. Light verbs and  $\theta$ -marking. *Linguistic Inquiry* 19. 205–232.
- Han, Aaron Li-feng, Derek F. Wong, Lidia S. Chao, Liangye He, Yi Lu, Junwen Xing & Xiaodong Zeng. 2013. Language-independent model for machine translation evaluation with reinforced factors. In Khalil Sima'an, Mikel L. Forcada, Daniel Grasmick, Heidi Depraetere & Andy Way (eds.), *Proceedings of the XIV machine translation summit*, 215–222.
- Hansen-Schirra, Silvia & Silke Gutermuth. 2018. Modellierung und Messung Einfacher und Leichter Sprache. In Susanne Jekat, Martin Kappus & Klaus Schubert (eds.), *Barrieren abbauen, Sprache gestalten*, 7–23. Winterthur: ZHAW Zürcher Hochschule für Angewandte Wissenschaften.
- Heine, Antje. 2017. Zwischen Grammatik und Lexikon: Ein forschungsgeschichtlicher Blick auf Funktionsverbgefüge. In *International conference – light verb constructions in Germanic languages*. Brussels: Université Saint-Louis.



### 3 German light verb construction in the course of the development of MT

- Hutchins, William J. & Harold L. Somers. 1992. *An introduction to machine translation*. Cambridge: Academic Press.
- Jespersen, Otto. 1942. *A modern English grammar on historical principles*. Part VI. Copenhagen: Ejnar Munksgaard.
- Jespersen, Otto. 1964. *Essentials of English grammar*. Alabama: University of Alabama Press.
- Kuhn, Jonas. 1994. *Die Behandlung von Funktionsverbgefügen in einem HPSG-basierten Übersetzungsansatz*. University of Stuttgart. (Doctoral dissertation).
- Marzouk, Shaimaa & Silvia Hansen-Schirra. 2019. Evaluation of the impact of controlled language on neural machine translation compared to other MT architectures. *Machine Translation* 33(1-2). 179–203. DOI: 10.1007/s10590-019-09233-w.
- North, Ryan. 2005. *Computational measures of the acceptability of light verb constructions*. University of Toronto. (Doctoral dissertation).
- Pollard, Carl J. & Ivan A. Sag. 1994. *Head-driven phrase structure grammar* (Studies in contemporary linguistics). Chicago: University of Chicago Press.
- Popović, Maja. 2018. Language-related issues for NMT and PBMT for English–German and English–Serbian. *Machine Translation* 32(3). 237–253. DOI: 10.1007/s10590-018-9219-5.
- Rösener, Christoph. 2010. Computational linguistics in the translator’s workflow: Combining authoring tools and translation memory systems. In *Proceedings of the NAACL HLT 2010 workshop on computational linguistics and writing*, 1–6. Association for Computational Linguistics.
- Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla & John Makhoul. 2006. A study of translation edit rate with targeted human annotation. *AMTA 2006: Proceedings of the 7th Conference of the Association for Machine Translation of the Americas: Visions for the Future of Machine Translation* 1. 223–231.
- Storrer, Angelika. 2006. Funktionen von Nominalisierungsverbgefügen im Text: Eine corpusbasierte Fallstudie. In Kristel Prost & Edeltraud Winkler (eds.), *Von der Intentionalität zur Bedeutung konventionalisierter Zeichen*, 147–178. Tübingen: Narr.
- Storrer, Angelika. 2007. Corpus-based investigations on German support verb constructions. In Christiane Fellbaum (ed.), *Collocations and idioms: Linguistic, lexicographic, and computational aspects*, 164–187. London: Continuum Press. [http://www.studiger.tu-dortmund.de/images/Storrer\\_2007\\_Corpus-based\\_investigations\\_on\\_german\\_support-verb\\_constructions.pdf](http://www.studiger.tu-dortmund.de/images/Storrer_2007_Corpus-based_investigations_on_german_support-verb_constructions.pdf).

- Toral, Antonio & Víctor M. Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th conference of the European chapter of the Association for Computational Linguistics*, vol. Volume 1: Long papers, 1063–1073. Valencia, Spain: Association for Computational Linguistics.
- Vilar, David, Jia Xu, Luis Fernando D’Haro & Hermann Ney. 2006. Error analysis of statistical machine translation output. In 697–702. Genoa, Italy: European Language Resources Association (ELRA).
- von Polenz, Peter. 1963. *Funktionsverben im heutigen Deutsch (Wirkendes Wort)*. Düsseldorf: Schwann.
- Winhart, Heike. 2005. *Funktionsverbgefüge im Deutschen: Zur Verbindung von Verben und Nominalisierungen*. Universität Tübingen. (Doctoral dissertation). <https://publikationen.uni-tuebingen.de/xmlui/bitstream/handle/10900/46248/pdf/Dissertation-Drive.pdf>.
- Zifonun, Gisela, Ludger Hoffmann & Bruno Strecker. 1997. *Grammatik der deutschen Sprache*. Berlin: Walter de Gruyter.

# Chapter 4

## Dialogue-oriented evaluation of Microsoft's Skype Translator in the language pair Catalan-German

Felix Hoberg

Leipzig University

This paper presents preliminary results of the work on Microsoft's Skype Translator. Tackling the question of how to evaluate such technology on a dialogue-oriented level, a case study on 21 German-speaking participants was conducted. Despite not having any proficiency in Catalan, these participants had to text-chat with Catalan native speakers via Skype, while the Skype Translator was activated. The sessions were observed by means of an eye tracking system. The collected data thus represents a naturalistic starting point to evaluate how users structure computer-mediated communication situations when real-time machine translation is involved and thereby they have to rely on that output.

### 1 Introduction

Automatic language processing, auto speech recognition and machine translation (MT) are considered valuable innovations by the language industry. However, progress in this field is still viewed skeptically, which in turn calls for continuous evaluation of the aforementioned systems (i.e. Ramlow 2009; Bowker & Ciro 2019), especially when it comes to dialogic interactions between humans and MT. Microsoft's Skype Translator will thus serve as a central element in this case study, as it offers real-time machine translation in 10 languages in voice and video chats and 60 languages in text chats.

To highlight how MT evaluation can be applied to services like the Skype Translator and how it has to be modeled on the dialogue-oriented level, the



project combines research in the fields of communication research (Beißwenger 2007) and machine translation. Additionally, this project aims to examine the behaviour of conversation's participants when an MT engine is involved (Fišer & Beißwenger 2017).

To achieve these goals, an exploratory eye-tracking-based case study was carried out. In that study, Skype Translator-mediated text chats between German and Catalan native speakers were captured in order to investigate the fixation duration and count on characteristic areas of interest of the Skype Translator.

This paper thus aims exclusively at giving a first impression on which aspects to analyse in the above mentioned context. For that reason, Section 2 introduces the theoretical background in terms of research on dialogue and conversation in the context of computer-mediated communication. Section 3 gives insights on the overall project conception, before explaining in detail to which extent the collected data is used for this analysis. Then, Section 4 presents early findings of the eye tracking data and situates them along the theoretical background, before the conclusion in Section 5 sums up the analysis, going back to the overall project.

## **2 Background**

### **2.1 Research on dialogue and conversation**

Since the early 1990s, various concepts in communication research have been modelled and restructured to fit on modern computer-mediated communication (cf. Fišer & Beißwenger 2017: 7). Apart from taking a look at global concepts such as text, sender, recipient or conversation, the interest in research has now passed on to questions which reflect the transitional processes web-based communication has undergone over the last two decades: How do we interact online? How does online interaction change our ways of communicating? Can we still speak of sender and recipient after all? How do we cope with this great amount of data and the rising machine learning technologies? (cf. Beißwenger 2007).

These questions also implicitly refer to the phenomena of turn-taking and speaker switch or the rising use of the term *hypertext* to describe digital textual behaviour (cf. Storrer 2001), central elements which have already been extensively studied regarding analog, face-to-face and monolingual web-based communication, but so far have not been adopted to bilingual, machine-translated, web-based conversations such as presented in this paper. This gap might be attributed to the fact that online communication follows different rules than offline communication. There are two obvious differences between oral, face-to-face and chat communication. The latter appears in written or typed form and lacks

of mostly all the non- and paraverbal elements like gesture, intonation or eye contact etc. which usually help to structure the communication act (cf. Beißwenger 2007: 172).

In contrast, an online chat message passes through more sections between a sender and an addressee than an oral, face-to-face talk. From the sender's mind, it goes from typing on the keyboard to the computer's short-term memory and from there to the server the software in use is connected to. From that server it goes to the addressee's software and it is subsequently processed by the computer to be displayed on screen before the addressee can spend cognitive resources on it (cf. Kienle et al. 2017: 146). In the case of the Skype Translator, one additionally has to take into account the time it takes to send, machine-translate and receive the original message. In the case of high latency, this time gap can have a severe impact on communication – while the person on the receiving end is still answering one incoming message, the other may already have sent another text. This can result in an asynchronous communication.

Thus, the use of computer-mediated communication technology, and in this case, to be precise, the Skype Translator, leads to a change in the communication process of sending and receiving messages. A text chat message has to be completely written before it can be sent<sup>1</sup> and it has to be received and read before it can be reacted to. At the same time, apart from in oral communication, the communication partners are not necessarily in the same location, nor near at all (cf. *ibid.*: 146).

Storrer (cf. 2001: 3) points out another important feature: even though online chatting appears mostly in written form, it follows the rules of oral production. The relationship of officially standardised language and its informal, but also widely accepted online communication use, which follows its own rules, has been an object of many research projects ever since, as for example in Verheijen (2017) in the context of Dutch. This relationship might helpfully be investigated by an eye tracking study. Consequently, the indicators explained below in Section 2.2 can be taken as initial points of reference on how the participants process the information on screen when text-chatting with people, whose language they do not speak.

## 2.2 Eye tracking and the Skype Translator

Parting from the communication research background above, it has to be made clear that this article focuses on Skype's text chat function, that is, on written

---

<sup>1</sup>Real-time text chat, where the text is transmitted immediately so that every user can observe the production process, will not be considered here.

communication. Voice- and video-chats are probably the main features Skype is known for. Skype Translator is also supported by those types of chat, but they will not be discussed here, since Catalan is not supported in those modes. That being stated, the focus passes on to written texts and their perception by its readers (or users), which have already been investigated in eye tracking studies. Two examples shall depict possible ways of combining eye tracking and reading perception at the background of translation studies.

In his article on the types of reading during translation, Hvelplund (2017: 63) investigates how reading leads to understanding a text “according to the purpose of the reading task” (ibid.: 55), exploring eye tracking data of different reading tasks (source text and target text reading while and when not typing, respectively). The mean fixation duration for the different categories is between 256ms for source text reading and 432ms for target text reading. The overall mean of the fixation duration is 332ms. Having reliable data on the cognitive workload of the participants that are reading the original and machine translated messages in Skype and on the accuracy the participants are reading with, some of the findings can be used for comparison in this study.

An investigation tackling the distribution of accurately reading and superficially scanning in information retrieval tasks has been conducted by Everdell (2014). That study took a closer look at how users are guided by the visual elements of web content and where their the main portion of their attention is drawn onto. Since the present article is also dealing with some sort of information retrieval in MT-mediated communication, some of the findings can be applied, too.

For that reason, *fixation count* and *fixation duration* represent two common but useful indicators to start with. More precisely, fixation count reflects how deeply the observed participants are reading, whereas fixation duration, most of the time measured in milliseconds, is taken as an indicator of cognitive load (cf. Hvelplund 2017: 63).

It can therefore be assumed that there will be observable differences in the users’ behaviour when reading their own original text messages, the respective MT output and also in Catalan.

### **3 Research design**

#### **3.1 Participants and task**

For this study, 25 students with no proficiency in Catalan were recruited. Of those 25 participants, four had to be excluded due to insufficient data quality. Of the remaining cohort, 20 were students at the Leipzig University and one was

#### *4 Dialogue-oriented evaluation of MS Skype Translator for Catalan-German*

a student at the Leipzig University of Applied Sciences (HTWK). As the call for participation was sent to almost all departments of these two universities, the participants vary in terms of programs they are enrolled in.

Three Catalan native speakers – two female and one male, aged 26, 24 and 26, respectively – were recruited as text chat counterparts for this study. All three came from different cities in the Catalan Countries: Valencia, Girona and Barcelona. All three were proficient in German since they took part in an exchange program during their studies and/or lived in Germany for a while.

The task the participants had to fulfil was split into three steps. First, they were asked to answer a questionnaire on their communication behaviour and their foreign language proficiencies. Second came the text chat session with a Catalan native speaker via Skype, having the Skype Translator activated. This part was captured by an eye tracking system. In order to get comparable data, the participants were given an introductory instruction: To have a central theme the participants could chat about, they were told to imagine they were about to spend a year abroad in Catalonia trying to get some information in advance on where to live and how to start there. Therefore, they were contacting the Catalan native speaker. On the one hand, this task allowed the participants to text-chat freely in a naturalistic manner, according to their individual communication behaviour. On the other hand, this constraining task was intended to produce comparable linguistic data, which can be analysed in corpus studies.

Last, to get an impression of the participant's individual experience during the Skype session, they had to fill out another questionnaire afterwards, concerning the output quality of the Skype Translator.

The introductory questionnaire provides additional data regarding the composition of the cohort. The students participants mean age was 23.7 (SD = 4.0, range = 20-32 years). When it comes to (foreign) language proficiency with the Common European Framework of Reference for Languages (CEFRL) as criterion, all of them indicated German as their first language with respect to use in ordinary and work life. 17 participants had English as foreign language. As for Romance languages, French and Spanish were reported nine times each, Italian and Portuguese one time each. Possible influences of Romance language proficiencies on the participants' behaviour have to be taken into consideration in a full-range analysis, but will not be discussed in this article.

Taking a look at the user behaviour regarding Skype, 17 participants reported using the software, but 13 of them only less than one time per month. At the same time, with regards to the duration per session, four participants used Skype no longer than 15 minutes, five no longer than 30 minutes, four up to one hour and four even beyond one hour.

The next part of the questionnaire was devoted to the use of alternative software, which includes all of the Skype's functions or just some of them, such as voice chat, then going into a detailed inquiry on alternatives for the individual Skype functions voice chat, video chat and text chat. Of 17 participants using alternatives, 16 used WhatsApp for voice chats, 15 for video chats and 16 again for text chats. Some participants stated that they were using other alternatives such as Telegram or Discord, too. Only three of them declared Skype as their preferred and most used software for video chats. As for voice or text chat, Skype was mentioned zero times as preferred and most used software. Instead, WhatsApp was indicated to be used most of the time. Last, the questionnaire took into account the participants' experience of living abroad. 13 of them have reported some experience living abroad with a mean of 30.53 months (SD = 36.36, range = 1–108 months).

### **3.2 Data collection**

The *Eye Link Portable Duo* eye tracking system was used to conduct the study. The sessions were recorded in the *head-free-to-move* setup at a sampling rate of 1000Hz and bi-ocular tracing. The overall setup included an eye tracking camera on a tripod, which was placed directly between the screen and the keyboard – 60–70cm from the participants' heads –, a display computer with Skype and the screen captioning software packages installed, and a host computer to handle the eye tracking system. The software in use also allowed capturing messages (buttons pressed etc.). The core element of this study was the latest version of Skype on that date (8.x), which already presented the Skype Translator as a built-in system element. The only requirement was to start a new conversation and add the Skype Translator service by clicking on the respective button in the user's profile one wanted to chat with. The service displayed messages in a two column structure: original messages of the user appear right-aligned, the MT output of the user, and the counterpart's incoming messages and the respective MT output appear left-aligned (see Figure 4.1). R version 3.4.3 (R Core Team 2019) and RStudio were used to analyse the collected eye tracking data.

### **3.3 Data preparation**

There are two kinds of analysable data that come from this study. On the one hand, there is the bilingual, authentic linguistic material produced by the participants, the Catalan native speakers and the machine translation of Skype which can be subdivided into four categories: the German and the Catalan original and



#### 4 Dialogue-oriented evaluation of MS Skype Translator for Catalan-German

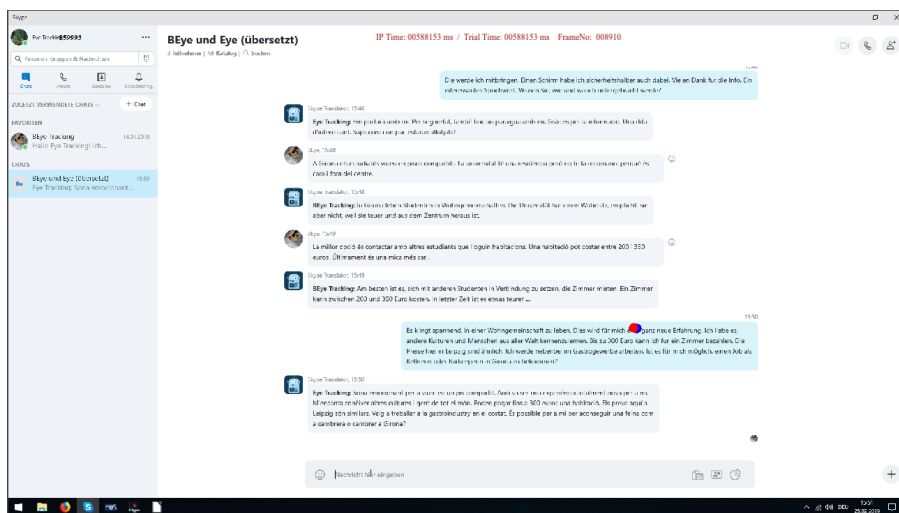


Figure 4.1: Example of text boxes in skype. Left-aligned (grey): incoming messages and all MT output. Right-aligned (light blue): original messages of the participant.

the machine translated output, respectively. This kind will be spared for further research and publications.

On the other hand, there are the screen captions of the eye tracking sessions. These had to be annotated with dynamic areas of interest as the single text elements in Skype move when a new message is displayed on screen. To allow for a detailed analysis of those four linguistic categories mentioned above, every text box of each session is marked by its own consecutively numbered area of interest (see Figure 4.1). Following the language codes proposed by ISO-639-2<sup>2</sup>, the following abbreviations were used to label those areas of interest: GerO: German original, GerMT: machine translation into German, CatO: Catalan original, and CatMT: machine translation into Catalan. The entry mask was labelled “Entry”. Moreover, these five categories allowed for a detailed analysis of the eye tracking data, as it was thus possible to create subsets sorted by participants, by label, by participant and label or other indicators.

The aforementioned 21 eye tracking sessions resulted in the video material of a total duration of 375 minutes, or 18 minutes on average per trial. Taking the interest area count as a measure, the mean of German text messages is 21

<sup>2</sup>See <https://www.bib-bvb.de/web/kkb-online/rda-sprachencode-nach-iso-639>, last accessed on 2020-08-31.

(SD = 9.60, range = 6–48), the mean of machine translated messages into Catalan 20 (SD = 9.79, range = 6–48, the mean of Catalan text messages 27 (SD = 10.85, range = 11–49) and the mean of machine translated messages into German 26 (SD = 10.66, range = 11–49). A diverging number of original and MT messages can be observed which is explained by the Skype Translator’s MT output that was for no obvious reason automatically merged into one text box even if two original messages were written.

## 4 Results

The following observations are based on the categories of interest area labels mentioned above (see Section 3.3). Table 4.1 shows the fixation count per area of interest label on aggregate which includes all fixations that fall into the dynamic interest areas of all the participants as described above. The participants are looking more often at the machine translation output than at the original messages regardless of the language. One prominent observation is that both the Catalan and German originals receive less than two thirds the amount of fixations of their respective machine translation equivalent (German original: 2469 to MT into German: 7037 and Catalan original: 3046 to MT into Catalan: 4553).

Figure 4.2 depicts the mean fixation duration per interest area category. The overall mean fixation duration of all fixations falling into the AOI is 314.05ms (SD = 157.94). The high standard deviation can be attributed to the fact that the participants were mainly resting their eyes at the entry mask when waiting for a response. This is also represented by the highest mean of all categories (348.43ms). The mean of fixations on the German MT output is slightly higher (320ms) than on the original message in German (305ms). The MT output from German into Catalan is fixated longer on average than the original, incoming messages in Catalan (344.93ms to 332.87ms).

### 4.1 Towards an analysis

As this is just a preliminary look into the collected data of the overall project, at the moment, it is undoubtedly not possible to give a full-range evaluation of the participants’ perception on using the Skype Translator to communicate with a counterpart whose mother tongue they do not speak. Nonetheless, a first look into the data set shows that the participants obviously take into account the machine translation output into Catalan despite not being proficient in that language. The differences between fixation count of the original messages and their

#### 4 Dialogue-oriented evaluation of MS Skype Translator for Catalan-German

Table 4.1: Fixation count per interest area tag

AOI tag	AOI total	Fix. count	mean	SD
GerO	433	2469	6.02	13.39
CatMT	429	4553	11.24	17.4
CatO	562	3046	5.61	10.20
GerMT	551	7037	13.22	16.33
Entry	21	3964	198.20	103.93

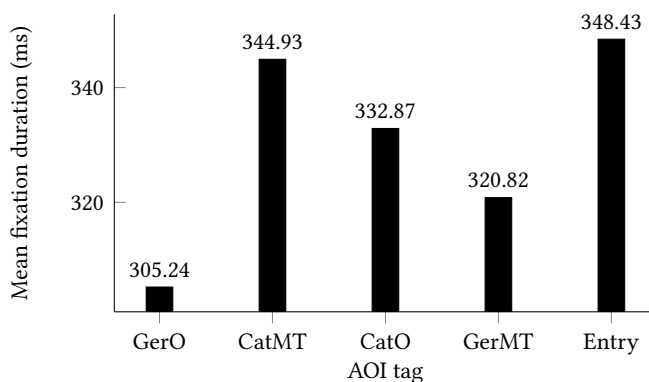


Figure 4.2: Mean fixation duration per AOI tag

machine translated counterparts may be taken as a hint for the participants either being at least curious about what their own message is translated to in Catalan or waiting for the Catalan counterpart to respond – therefore most services display a “...currently typing”-phrase somewhere near the entry mask. It is possible that by switching between the original and the MT, the participants check on the messages for their integrity or completeness. In addition, even though the fixation count of the Catalan original interest area is equally low to that of the German original compared to the machine translation ones, its fixation duration is higher than the German original and the machine translation into German. To be precise, there are fewer fixations but the ones last longer.

## 5 Conclusion

The research on computer-mediated communication has been a field of academic interest for a while. Therefore, looking at the bilingual conversations that are mediated by machine translation seems to be a crucial aspect for pointing out how such technology will change the way the users experience these forms of communication. The main task of analysing how conversation partners interact when none of them is proficient in the other's language requires thus special attention to their gaze behaviour, that is, to the way they pay attention to the MT output. Whether this means that they are – supported by the MT output – hypothesising about what they are reading or that their attention is simply drawn onto this as it is a new message on the screen has to be profoundly investigated. In this context, an analysis of the regressions that fall into or part from each AOI category seems to be a quite promising approach (cf. Eskenazi & Folk 2017). The collected data thus represents a starting point for evaluating linguistic, cognitive and technological aspects.

## Abbreviations

CatO	Catalan original	GerO	German original
CatMT	Machine translation into Catalan	GerMT	Machine translation into German

## Acknowledgements

I am grateful to our institute's student assistant Tim Feldmüller, who took care of preparing most of the collected research data.

## References

- Beißwenger, Michael. 2007. *Sprachhandlungscoordination in der Chat-Kommunikation* (Linguistik, Impulse & Tendenzen 26). Berlin: W. de Gruyter.
- Bowker, Lynne & Jairo Buitrago Ciro. 2019. *Machine translation and global research: Towards improved machine translation literacy in the scholarly community*. OCLC: on1075580986. Bingley, UK: Emerald Publishing.
- Eskenazi, Michael A. & Jocelyn R. Folk. 2017. Regressions during reading: The cost depends on the cause. *Psychonomic Bulletin & Review* 24(4). 1211–1216. DOI: 10.3758/s13423-016-1200-9. (2020-01-28).

- Everdell, Ian. 2014. Web content. In Jennifer Romano Bergstrom & Andrew Jonathan Schall (eds.), *Eye tracking in user experience design*, 163–186. Amsterdam & Boston: Elsevier. [https://ebookcentral.proquest.com/lib/leip/detail.action?docID=1651794#goto\\_toc](https://ebookcentral.proquest.com/lib/leip/detail.action?docID=1651794#goto_toc).
- Fišer, Darja & Michael Beißwenger (eds.). 2017. *Investigating computer-mediated communication: Corpus-based approaches to language in the digital world*. 1st edn. (Book series translation studies and applied Linguistics). Ljubljana: Ljubljana University Press. <https://e-knjige.ff.uni-lj.si/> (2017-10-27).
- Hvelplund, Kristian Tangsgaard. 2017. Four fundamental types of reading during translation. In Arnt Lykke Jakobsen & Bartolomé Mesa-Lao (eds.), *Translation in transition: Between cognition, computing and technology*, vol. 133, 56–78. Amsterdam: John Benjamins Publishing Company. DOI: 10.1075/btl.133.02hve. (2019-02-14).
- Kienle, Andrea, Michael Beißwenger, Linda Cedli, Torsten Holmer, Philipp Schlieker-Steens & Christian Schlösser. 2017. Eyetracking als Ressource zur Unterstützung des Interaktionsmanagements in synchroner Schriftkommunikation. In Michael Beißwenger (ed.), *Empirische Erforschung internet-basierter Kommunikation*, 143–174. Berlin, Boston: De Gruyter.
- R Core Team. 2019. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. <http://www.R-project.org>.
- Ramlow, Markus. 2009. *Die maschinelle Simulierbarkeit des Humanübersetzens: Evaluation von Mensch-Maschine-Interaktion und der Translatqualität der Technik* (TransÜD 27). OCLC: 553597343. Berlin: Frank & Timme.
- Storror, Angelika. 2001. Sprachliche Besonderheiten getippter Gespräche: Sprecherwechsel und sprachliches Zeigen in der Chat-Kommunikation. In Michael Beißwenger (ed.), *Chat-Kommunikation. Sprache, Interaktion, Sozialität & Identität in synchroner computervermittelter Kommunikation. Perspektiven auf ein interdisziplinäres Forschungsfeld*, 3–24. Stuttgart: ibidem.
- Verheijen, Lieke. 2017. WhatsApp with social media slang? Youth language use in Dutch written computer-mediated communication. In Darja Fišer & Michael Beißwenger (eds.), *Investigating computer-mediated communication: Corpus-based approaches to language in the digital world*, 1st edn. (Book series translation studies and applied Linguistics), 72–101. Ljubljana: Ljubljana University Press. <https://e-knjige.ff.uni-lj.si/> (2017-10-27).



# Chapter 5

## Investigating post-editing: A mixed-methods study with experienced and novice translators in the English-Greek language pair

Maria Stasimioti & Vilelmini Sosoni


Ionian University

In recent years, Post-Editing (PE) has been increasingly gaining ground, especially following the advent of neural machine translation (NMT) models. However, translators still approach PE with caution and skepticism and question its real benefits. This study investigates the perception of both experienced and novice translators vis-à-vis PE, it compares the technical, temporal and cognitive effort expended by experienced translators during the full PE of NMT output with the effort expended by novice translators, focusing on the English-Greek language pair and explores potential differences in the quality of the post-edited texts. The findings reveal a more negative stance of the experienced translators as opposed to novice translators vis-à-vis Machine Translation (MT) and a more pragmatic approach vis-à-vis PE. However, the novice translators' more positive attitude does not seem to positively affect the temporal and cognitive effort that they expend. Finally, experienced translators have a tendency to overcorrect the NMT output, thus carrying out more redundant edits.

### 1 Introduction

In recent years, the translation industry has seen a growth in the amount of content to be translated and has received pressure to increase productivity and speed at reduced costs. To respond to these challenges, it has turned to machine translation (MT) “which appears to be moving from the peripheries of the translation



Maria Stasimioti & Vilelmini Sosoni. 2021. Investigating post-editing: A mixed-methods study with experienced and novice translators in the English-Greek language pair. In Tra&Co Group (ed.), *Translation, interpreting, cognition: The way out of the box*, 79–104. Berlin: Language Science Press. DOI: 10.5281/zenodo.4545037 

field closer to the centre” (Koponen 2016: 131). The most common and widely expanding scenario – especially for certain language pairs and domains – involves the use of MT output to be then post-edited by professional translators (Koponen 2016). This practice is generally termed post-editing of machine translation (PEMT) or simply post-editing (PE) and is subcategorised into two types according to the required quality: full post-editing, which is expected to improve the final product to publishable quality, and light/rapid post-editing, which aims to correct the text for accuracy, but not style and fluency (Allen 2003).

PE has been increasingly gaining ground (O’Brien et al. 2014; O’Brien & Simard 2014; Lommel & DePalma 2016; Nunes Vieira et al. 2019), especially following the advent of neural machine translation (NMT) models which have been proven to consistently outperform statistical machine translation (SMT) models in shared tasks, as well as in various project outcomes (Bojar et al. 2016; Toral & Sánchez-Cartagena 2017; Castilho, Moorkens, Gaspari, Calixto, et al. 2017; Castilho, Moorkens, Gaspari, Sennrich, et al. 2017). In fact, NMT has been widely hailed as a significant development in the improvement of the quality of MT, especially at the level of fluency (Castilho, Moorkens, Gaspari, Calixto, et al. 2017; Castilho, Moorkens, Gaspari, Sennrich, et al. 2017), and the PE of NMT output has been found to be faster than translation from scratch (Jia et al. 2019).

However, translators still approach PE with caution and skepticism and question its real benefits (Gaspari et al. 2014; Koponen 2012; Moorkens et al. 2018). Their skepticism is directly related to the nature of PE which involves “working by correction rather than creation” (Wagner 1985: 2), to the belief that PE is slower than translating from scratch and to the view that MT is a threat to their profession (Moorkens et al. 2018: 58). Several studies were carried out aiming at identifying the extent to which attitudes to MT and PE affect PE effort. A positive attitude to MT has been found to be a factor in PE performance (De Almeida 2013; Mitchell 2015). Experienced translators have been found to exhibit rather negative attitudes to PE as opposed to novice translators (Moorkens & O’Brien 2015) and to be rather reluctant to take on PE jobs, while novice translators appear to be more positive towards MT and PE and more suited for PE jobs (García 2010; Yamada 2015). Previous research has shown that professional translators and novices generally exhibit different translation behaviour (Carl & Buch-Kromann 2010; Carl et al. 2010; Hvelplund 2011; Dragsted & Carl 2013; Moorkens & O’Brien 2015; Nitzke 2019; Schaeffer et al. 2019), while cognitive effort has been found to be greater for novice than for professional translators (Göpferich et al. 2011). Yet, there is still a lack of empirical studies about post-editing by different profiles (Mesa-Lao 2014). Under the light of the above, the



aim of this study is threefold: it seeks to investigate the perception of both experienced and novice translators vis-à-vis PE; it aims to compare the technical, temporal and cognitive effort (Krings 2001) expended by experienced translators with the effort expended by novice translators during the full PE of NMT output, focusing on the English-Greek language pair; finally, it aims to explore potential differences in the quality of the post-edited texts.

## 2 Related work

The comparison of the translation, revision and PE behaviour of student translators and professional translators has been the subject of several empirical studies.

Carl et al. (2010) compared the revision behaviour of 12 student translators with that of professional translators and found that professionals tend to have a longer revision phase after they have completed a first draft of the translation. Irrespective, though, of when the revision is made, students and professionals revised the same parts of the translations, presumably due to the fact that they face the same problems during translation.

Carl & Buch-Kromann (2010) analysed the user activity data (eye movements and keystrokes) of 12 students and 12 professional translators translating two small English texts into Danish. The human translations, as well as a machine translation produced by Google Translate, were evaluated and compared both automatically with BLEU and manually with human scores for fluency and accuracy. The results revealed that although both professionals and students produce equally accurate translations, professionals seem to be better at producing more fluent texts more quickly than students.

Dragsted & Carl (2013) asked 12 professional full-time translators with at least two years professional of experience in translation between Danish and English, and 12 MA students at the Copenhagen Business School, specialising in translation between Danish and English to translate three small English texts into Danish. Eye tracking and keylogging data were used in order to identify typical behaviour in translation students and professionals. The authors found differences between the two groups. In particular, contrary to the earlier findings by Jakobsen (2002), students carried out more initial planning than professionals, while professionals carried out more end revision than students. In addition, they found a tendency for professionals to prefer local orientation in the initial planning phase and during drafting and a more global perspective in the revision phase.

Moorkens & O'Brien (2015) carried out a study with nine expert translators (having on average 11.3 years of translation experience and 4 years of PE experience) and 35 undergraduate student translators, and studied the productivity/speed of the participants, the edit distance of their final product, and their attitude towards PE. The participants were asked to carry out two PE tasks from English into German. The professional group was much faster than the students, while the students tended to edit less of the MT output and they had a more positive attitude towards PE.

Nitzke (2019) asked twelve professional translators (university degree holders with at least some professional work experience) and twelve semi-professional translators (university students with limited professional work experience) to translate two texts from scratch, to bilingually post-edit two machine translated texts and to monolingually post-edit two machine translated texts. No significant difference between students and professionals was reported in terms of task duration and keystrokes (insertions and deletions).

Schaeffer et al. (2019) compared two different translation expertise levels, i.e. translation students and professional translators, and investigated the process of the revision of six English-German human pre-translated texts using eye-tracking and keystroke logging data. The results revealed that translation students correct significantly fewer errors and take longer to correct these errors as compared to professional translators.

The question of the role of translation expertise on PE has hardly been addressed. To the best of our knowledge, only two studies investigated the influence of translation expertise on PE (Moorkens & O'Brien 2015; Nitzke 2019), while only a handful of studies focused on the influence of the perception of MT and PT on PE (Moorkens & O'Brien 2015). In addition, no studies to date involved the role of translation expertise and Greek or focused on Greek translators' perception vis-à-vis MT and PE. We will therefore help fill this research gap by investigating PE carried out by experienced and novice translators in the English-Greek language pair.

### **3 Experimental setup**

As already pointed out, the study adopts a mixed-methods approach and triangulates findings from different methods. For the investigation of the participants' perception of PE, pre-assignment questionnaires are used. For the assessment of PE effort and following Koponen et al. (2019), the study uses both process-based approaches and product-based approaches. In particular, keystroke logging and

eye-tracking data are used to measure the temporal, technical and cognitive effort expended by translators during the PE of NMT output generated by the NMT system developed by Google (Google Translate NMT system). Moreover, the MT output is compared with the final post-edited text to determine the number and type of changes using the edit distance metric WER; this has been widely used as an indicator of MT quality and PE effort. Finally, an analysis of the number and the type of each edit is carried out to evaluate the quality of the post-edited texts.

The PE experiments were carried out in March 2018 at the HUBIC lab<sup>1</sup> (Raptis & Giagkou 2016) of the Athena research center<sup>2</sup> in Athens.

A detailed consent form was signed by all participants prior to the execution of the experiments, while all stored data were fully anonymised in accordance with Greek law 2472/97 (as amended by laws 3783/2009, 3917/2011 and 4070/2012).

### 3.1 The participants

Twenty translators – ten experienced translators and ten novice translators – participated in the experiments, in which their eye movements and typing activity were registered with the help of an eye-tracker and specialised software (see Section 3.2). In the present study, following Whyatt & Kościuczuk (2013) and Colina & Angelelli (2015), experienced translators are selected among those with more than five years of translation experience and novice translators among those with less than five years of translation experience. Experienced translators are considered to be the “products of long years of deliberate practice” who exhibit “consistently high levels of performance” (Shreve 2002: 151), while they employ some “cognitive routines” occurring “automatically” (Kaiser-Cooke 1994: 137). On the other hand, novice translators are considered to “lack the routine processes acquired by experience” (Palumbo 2009: 130) and thus face more problems during a translation task. It should also be noted that in the present study neither experience nor training in PE was a prerequisite for participating in the experiments.

Their selection followed a call for participation, which was sent to the members of the two biggest Greek associations of professional translators, i.e. the Panhellenic association of translators<sup>3</sup> (PEM) and the Panhellenic association of professional translation graduates of the Ionian University<sup>4</sup> (PEEMPIP), and was

---

<sup>1</sup><http://www.hubic-lab.eu/>

<sup>2</sup><https://www.athenarc.gr/en>

<sup>3</sup><http://www.pem.gr/el/>

<sup>4</sup><http://peempip.gr/el/>

shared on social media. Translators expressed their interest for participating in the study by filling in a Google form; they subsequently received an e-mail with details on the aim of the research and guidelines for the PE task along with some educational material on PE (see Section 3.3). In addition, they were asked to fill in a questionnaire before the actual experiment. The questionnaire consisted of 34 questions: 22 closed-ended (multiple choice) and 12 open-ended, all of which aimed at defining the profile of the participants and their perception of MT and PE.

The following graphs and tables provide information regarding the participants' profiles.

Table 5.1: Participants' gender, age distribution, education level and degree type

		Translators	
		Experienced	Novice
Gender	Male	0	1
	Female	10	9
Age distribution	20–30	3	8
	30–40	4	1
	40–50	3	1
Education level	Undergraduate	4	4
	Postgraduate	6	6
Degree type	Translation	7	8
	Language/linguistics	2	1
	Other	1	1

It should be noted that all participants had normal or corrected to normal vision. Five wore contact lenses and four wore glasses, yet the calibration with the eye-tracker was successful for all of them.

All experienced translators had many years of translation experience (see Table 5.2) compared to novice translators who had either 0 years of translation experience (3), 1 year of translation experience (5), 2 years of translation experience (1) or 3 years of translation experience (1).

5 Investigating post-editing: A mixed-methods study in English-Greek

Table 5.2: Participants' years of translation experience

Years of translation experience	Translators	
	Experienced	Novice
1–5	0	10
5–10	5	0
10–20	4	0
Over 20	1	0

Table 5.3: Participants' years of experience in PE

Years of experience in PE	Translators	
	Experienced	Novice
0 years	1	10
1 year	2	0
2 years	2	0
3 years	3	0
4 years	0	0
5 years	1	0
over 5 years	1	0

As far as experience in PE is concerned, none of the novice translators had experience in PE, compared to almost all of the experienced translators who had at least 1 year of experience in PE<sup>5</sup> (see Table 5.3).

As can be seen in Table 5.4, although novice translators had no experience in PE, half of them had received relevant training in the past. Half of the experienced translators had also received previous training in the past. More interestingly, as it emerges from Tables 5.5 and 5.6, all the novice translators would be interested in attending a PE course in the future, considering it to be either important or very important, while 7 experienced translators would be interested in attending a PE course in the future, considering it fairly important (5), important (4) or very important (1).

<sup>5</sup>One year of experience in PE corresponds to one year full-time equivalent (FTE) and it represents the number of working hours that one full-time worker completes during a year or during any other fixed time period.

Table 5.4: Participants' previous training in PE

Previous training in PE	Translators	
	Experienced	Novice
Yes	5	5
No	5	5

Table 5.5: Participants' interest in PE training

Interest in PE training	Translators	
	Experienced	Novice
Yes	7	10
No	3	0

Table 5.6: Participants' view on the importance of PE training

Importance of PE training	Translators	
	Experienced	Novice
Very important	1	5
Important	4	5
Fairly important	5	0
Slightly important	0	0
Not important	0	0

### 3.2 Description of the experiment

A Tobii TX-300 eye-tracker and the Translog-II software (Carl 2012) were used to register the participants' eye movements, keystrokes and time needed during the PE tasks they were asked to carry out. The participants were asked to work at the speed at which they would normally work in their everyday work as translators; therefore, no time constraint was imposed. However, they did not have Internet access and were not allowed to use online or offline translation aids as this could lead to a reduction in the amount of recorded eye-tracking data. In the case of offline translation aids (e.g. dictionaries), the participants might look away from the screen resulting in a reduced amount of analysable eye-tracking data; in the case of online translation aids (Internet), the eye-tracking data would partially reflect gaze activity that does not directly reflect source text (ST) processing, target text (TT) processing or parallel ST/TT processing (Hvelplund 2011: 86).

The experiment consisted of one session for each participant. Before the sessions, the participants were informed by email about the nature of the experiments, the task requirements and the general as well as task-specific guidelines they had to follow (see Section 3.3). The session started with a warm-up PE task to familiarise each participant with the procedure. After the warm-up task, the participants had to carry out full PE of the NMT output of the same two semi-specialised texts – in accordance with the detailed task-specific guidelines they received (see Section 3.3.2); the two texts were presented to them in the same order. During the experiment, the ST was displayed in the Translog-II software at the top half of the screen and the MT output at the bottom half, as suggested by previous studies (Hvelplund 2011; Mesa-Lao 2014; Carl et al. 2011; 2015; Schmaltz et al. 2015). Translators worked directly on the MT output. To facilitate eye-tracking measurements, texts were fully displayed to avoid any need for participants to scroll in either the ST or the TT window. For the purposes of this study, each ST and each TT was considered an Area of Interest.

The STs used in this study were short educational texts selected from OER Commons<sup>6</sup>, which is a public digital library of open educational resources. Three<sup>7</sup> excerpts of around 140 words each were selected from various courses on social change and the endocrine system and the titles of the courses were retained as context information for the participants. The texts were chosen with the following criteria in mind: they had to be semi-specialised and easy for participants to

---

<sup>6</sup><https://www.oercommons.org/>

<sup>7</sup>One text was used exclusively for the warm-up session and is not included in the ensuing analysis and discussion.

post-edit without access to external resources and they also had to be of comparable complexity. The texts chosen had comparable Lexile® scores (between 1300L and 1400L), i.e. they were suitable for 11th/12th graders. The Lexile Analyzer<sup>8</sup> was used as it relies on an algorithm to evaluate the reading demand – or text complexity – of books, articles and other materials. In particular, it measures the complexity of the text by breaking down the entire piece and studying its characteristics, such as sentence length and word frequency, which represent the syntactic and semantic challenges that the text presents to a reader. The outcome is the text complexity, expressed as a Lexile measure, along with information on the word count, mean sentence length and mean log frequency.

Table 5.7: Lexile® scores for the source texts used in the study

	Text 1–T1	Text 2–T2
Lexile® measure	1300L–1400L	1300L–1400L
Number of sentences	5	6
Mean sentence length	28.60	22.67
Word count	143	136
Characters without spaces	647	1761

The NMT-core engine used to produce the raw MT output was Google Translate (output obtained March 24, 2018). The MT output was evaluated using the bilingual evaluation understudy (BLEU) metric. Generally speaking, a score below 15 percent means that the engine is not performing optimally and PE is not recommended as it would require a lot of effort to finalise the translation and reach publishable quality, while a score above 50 percent is a very good score and means that significantly less PE is required to achieve publishable translation quality. The BLEU score was calculated for the two texts using the Tilde custom machine translation toolkit<sup>9</sup>. Both texts had a very good score. In particular, BLEU score for Text 1 was 51.33 and for Text 2 was 60.62. This means that PE could be used to achieve publishable translation quality in both cases.

### 3.3 PE task guidelines

In the PE task, the participants were asked to fully post-edit the raw output generated by the NMT system. Since previous training and experience in PE was

<sup>8</sup><https://lexile.com/>

<sup>9</sup><https://www.letsmt.eu/Bleu.aspx>



not a prerequisite for participating in the study, the participants received brief training in PE before executing the task. The training included a video and a presentation on PE. In addition, they received general as well as task-specific guidelines which they were instructed to follow with the aim of achieving consistency and in order to avoid interference with the eye-tracker connection. The task-specific guidelines, i.e. the guidelines for the full PE of the NMT output, were based on the comparative overview of full PE guidelines provided by Hu & Cadwell (2016) as these were proposed by TAUS (2016), O'Brien (2010), Flanagan & Paulsen Christensen (2014), Mesa-Lao (2013) and Densmer (2014).

### 3.3.1 General guidelines

- Your hair should not block your eyes.
- Do not wear mascara.
- Avoid touching your eyes (e.g. rubbing your eyes, removing and wearing eyeglasses, etc.).
- During the PE tasks, look exclusively at the computer screen in front of you.
- Try to keep your head as steady as possible.
- External resources (dictionaries, Internet, etc.) cannot be used

According to O'Brien (2009), the quality of the eye-tracking data may be affected by several factors, such as participants' optical aids, eye make-up, lighting conditions, noise, unfamiliarity, user's distance from the monitor, etc. In an effort to minimise the implications of some of these factors, the participants were given the above-mentioned general guidelines, while a controlled environment for the experiment was set up. In particular, a quiet room was selected, blackout blinds were used to reduce the amount of natural light, the same artificial light was used during all the experiments, and a fixed chair was used, so that the participants could not easily move about and increase or decrease the distance to the monitor (Hvelplund 2011: 103).

### 3.3.2 Task-specific guidelines

- Retain as much raw MT translation/output as possible.
- The message transferred should be accurate.
- Fix any omissions and/or additions (at the level of sentence, phrase or word).
- Correct mistranslations.

- Correct morphological errors.
- Correct misspellings and typos.
- Fix incorrect punctuation if it interferes with the message.
- Correct wrong terminology.
- Fix inconsistent use of terms.
- Do not introduce stylistic changes.

The training material was sent to the participants five days before the execution of the tasks. In an effort to ensure that they had actually studied the material and that there were no questions or doubts, the participants were interviewed prior to the execution of the tasks and were specifically asked about the training material and also about the guidelines they had received.

### **3.4 Process and product analysis**

As already pointed out, one of the aims of this paper is to compare the technical, temporal and cognitive effort (Krings 2001) expended by the experienced translators for the full PE of NMT output with the effort expended by the novice translators, focusing on the English-Greek language pair. Another aim is to investigate the quality of the post-edited texts and to explore potential differences in the work produced by the experienced and the novice translators.

According to Krings (2001), there are three categories of post-editing effort: (i) the temporal effort, which refers to the time taken to post-edit a sentence to a particular level of quality, (ii) the technical effort, which refers to keystroke and mouse activities such as deletions, insertions, and text re-ordering and (iii) the cognitive effort, which refers to the “type and extent of those cognitive processes that must be activated in order to remedy a given deficiency in a machine translation” (Krings 2001: 179). The cognitive effort is directly related to the temporal effort and the technical effort; however, these do not inform how PE occurs as a process, how it is distinguished from conventional translation, what demands it poses on post-editors, and what kind of acceptance it receives from them (Krings 2001: 61). Furthermore, temporal, technical and cognitive effort do not necessarily correlate, since some errors in the MT output may be easily identified, but may require many edits, while other errors may require a few keystrokes to be corrected, but involve considerable cognitive effort (Krings 2001; Koponen 2012; Koponen et al. 2019).

For our process-based analysis we used eye-tracking and keystroke logging data to measure the temporal, technical and cognitive effort. As far as the temporal effort is concerned, we measured the total time (in minutes) the participants

needed to post-edit the NMT output. It should be noted that the start time of the PE task was calculated from the moment we opened the project (i.e. when we pressed the “start logging” button); the task was considered completed when we pressed the “stop logging” button in the Translog-II. Technical effort is generally measured by the number of keystrokes, i.e. insertions and deletions. Finally, cognitive effort has been generally measured by calculating fixation count, fixation duration and gaze time (Sharmin et al. 2008; Carl et al. 2011; Mesa-Lao 2014; Jia et al. 2019; Doherty et al. 2010; Elming et al. 2014). In view of that, in the present study we measured the average fixation count, the mean fixation duration (in milliseconds), as well as the average total gaze time (in minutes), i.e. the sum of all fixation durations, to compare the cognitive effort expended by experienced translators and novice translators when post-editing the NMT output.

For our product-based analysis and similarly to previous studies (Carl & Buch-Kromann 2010; Moorkens & O’Brien 2015; Koponen et al. 2019; Koponen & Salmi 2017), we used an edit distance metric, i.e. word error rate (WER), to analyse the PE product and each participant’s final post-edited text was used as a reference text. WER is based on the Levenshtein<sup>10</sup> distance, and calculates the edits performed to measure the distance between the MT output and its post-edited version. However, these edits do not always reflect actual errors (Koponen et al. 2019). Previous studies (De Almeida 2013; Koponen & Salmi 2017; Koponen et al. 2019) have shown that post-editors either over-edit the MT output making preferential choices or they under-edit it leaving errors uncorrected, while sometimes they also introduce new errors. For that reason, following the calculation of WER, each edit operation was annotated manually by one annotator – a professional translator with 10 years of translation experience – with one of the following categories suggested by Koponen & Salmi (2017) and Koponen et al. (2019):

- unedited*: no change;
- form changed*: different morphological form;
- word changed*: different lemma;
- deleted*: word removed;
- inserted*: word added;
- order*: position of a word changed.

It should be noted that, following Koponen & Salmi (2017) and Koponen et al. (2019), in case a word had been affected by more than one edit type, it was annotated with both categories. For example, some words had both their morphological form and their position changed (form + order) or a different lemma was used and its position was also changed (word + order).

---

<sup>10</sup><http://www.levenshtein.net/>

Each word-level edit was then assessed for correctness of meaning (accuracy) and language (fluency) as well as for necessity, i.e. for establishing whether the edit was necessary to correct the meaning and/or the language or whether it was a preferential edit in terms of style or word choice. According to Koponen & Salmi (2017) and Koponen et al. (2019), each edit could be either correct or incorrect and either necessary or unnecessary. In some cases, no edit may be required in the MT output, while in some other cases post-editors may leave errors uncorrected or make preferential changes. Therefore, we decided in our study to cater for such cases by adding additional options. In particular, as far as correctness is concerned, we added the “correct no edit” option, which means that the post-editor was right to leave the MT output unedited given that there was no error in the MT output; we also added the “edit missing” option for cases when an error in the MT was not corrected by the post-editor; finally, we added the “redundant edit” option, which indicates a preferential change made by the post-editor. As far as necessity is concerned, we added the “necessary no edit” option, which means there was no error in the MT output and therefore no edit was required, and the “edit required” option, which means that there was an error in the MT which was not corrected by the post-editor. Summarising the above, the options for correctness and necessity are the following:

*Correctness*

- correct no edit
- correct edit
- incorrect edit
- edit missing
- redundant edit

*Necessity*

- necessary no edit
- necessary edit
- unnecessary edit
- edit required

## **4 Findings and discussion**

### **4.1 Perception analysis: Pre-assignment questionnaire**

As it emerges from the participants’ answers in the pre-assignment questionnaire in relation to their attitude towards MT and PE, novice translators appear to be more positive towards MT and PE compared to the experienced translators confirming, thus, the findings of previous studies (Moorkens & O’Brien 2015; García 2010; Yamada 2015).

#### **4.1.1 Perception of MT**

In the open-ended question regarding their view of MT, only one experienced translator characterised MT as “very positive”. Three experienced translators

gave negative responses characterising it as a “necessary evil” (P01), “still very inadequate” (P03) and “not very helpful” (P05). The rest of the experienced respondents were neutral and pointed to the deficiencies of MT or the improvements or conditions that are required for its efficient use by translators (i.e. the use of good quality data, domain specialisation and deep learning). For instance, respondent P06 wrote: “If the machine has been adequately trained, it’s OK. In any case, those machines should be continuously provided with new texts in order to be in a position to better ‘understand’ the structure of each language.” Respondents P09 and P10 referred to the MT quality for the English-Greek language pair, noting respectively: “MT is not very developed for Greek yet. It does, however, seem to be highly developed for other languages. Overall, I have not seen substantial results yet concerning the quality of the produced texts, but it is gradually getting better” and “It certainly needs improvement in the language combinations including Greek”. Finally, respondent P09 made a very interesting comment about MT and its impact on translators: “I don’t think it will ‘replace’ translators, but I do see it acting as a first stage in the localisation process, in the future, followed by PE and other quality assurance tasks”.

The novice translators, on the other hand, found MT “very helpful” (P01), “a vital part of the translation process” (P02), “extremely useful” (P04, P10), “necessary” (P06), and noted that “it helps translators save time and finish a project faster” (P08) and “saves money and time for clients” (P02). Only two novice translators expressed reservations saying: “MT is a tool that can be both useful and useless, depending on the extent it is used and the language knowledge, as algorithms see only words not meanings, and there, a human translator comes to the rescue” (P03) and “I find it quite challenging”. Finally, it is interesting that MT is not viewed as a threat by the novice translators, who appear to be more confident as regards the central role played by humans in the translation process. As respondent P06 mentioned, “MT can never replace human translation”.

#### 4.1.2 Perception of PE

In the open-ended question regarding their view of PE, the experienced translators appear to adopt a more pragmatic approach, accepting its necessity and its future dominance in the translation market. Six respondents highlighted its importance and/or its necessity saying: “It may be the solution to some problems of the industry” (P02), “It is necessary” (P03, P08) “It is very useful and time-saving” (P04) “it will be widely used in the future so familiarising myself with it will bring added value to my work” (P01) and “It is the future” (P10). Four experienced translators adopted a more reserved stance to refer to PE. Respondent P01 wrote: “Too much effort, too little time”, while respondent P09 noted “In my

personal experience, it makes my job harder, since MT is not very developed for Greek yet. However, it is now a hot trend in the localisation field and does seem to be gaining some ground. It could help with some aspects of the field, but does create problems in others.” Respondent P06 highlighted its usefulness in certain fields: “Useful only in standard fields, e.g. technical, automotive, maybe medical. Certainly not suitable for marketing, or other creative work” Finally, one experienced translator clearly expressed their preference for translation as opposed to PE saying: “I prefer translating from scratch” (P05).

As far as novice translators are concerned, nine out of the ten underscore the importance and crucial role of PE with the exception of one. They write: “It is important in order to achieve a legible and comprehensible text, whether to be published or not” (P02), “It is really necessary” (P01, P05, P09), “It is of primary importance”, “It is important as the first time you see a document, you are not always 100 percent focused and in that way you revise the text for making sure it is correct” (P03), “Post-editing contributes to the speed of the translation and controls the quality of the product” (P04), “Post editing is essential when it comes to delivering a satisfying translation. If this translation is a product of MT and not Human Translation. Post editing can also help improve MT in the way that AI systems ‘learn’ through their experience” (P07), “It could save time for the translator, as long as it is applied to technical texts” (P08). Finally, respondent P10 acknowledged the fact that it can be useful, albeit time-consuming: “it can be helpful but sometimes its processing requires more time than translation from scratch”.

## **4.2 Process analysis: Measuring PE effort**

### **4.2.1 Temporal effort**

As far as the temporal effort is concerned, we measured the average time (in minutes) experienced translators, on the one hand, and novice translators, on the other hand, needed to post-edit the two texts. As it emerges from Table 5.8, experienced translators needed less time ( $M = 7.17$ ,  $SD = 1.55$ ) to post-edit the NMT output compared to novice translators ( $M = 8.84$ ,  $SD = 2.98$ ). According to a two-tailed two-sample  $t$ -test, that difference in average task time between experienced and novice translators is statistically significant  $t(29) = -2.22$ ,  $p = 0.02$ . Our findings corroborate the findings of Carl & Buch-Kromann (2010), Moorkens & O’Brien (2015) and Schaeffer et al. (2019) who also reported professional translators to be faster than students. They come in contrast, though, with the findings by Nitzke (2019), given that in her study no significant difference was observed between professionals and students in terms of task duration.

Table 5.8: Temporal effort per group of participants: Mean and standard deviation values of the task duration (both texts averaged)

Participants	Task duration (in mins)	
	Mean	SD
Experienced translators	7.17	1.55
Novice translators	8.84	2.98

#### 4.2.2 Technical effort

As it emerges from Table 5.9, the experienced translators performed more keystrokes ( $M = 453, SD = 210$ ) compared to the novice translators ( $M = 318, SD = 179$ ). Similarly to the temporal effort, a statistically significant difference  $t(37) = 2.18, p = 0.02$  was reported. Nitzke (2019), who investigated also the technical effort, found no significant difference between professionals and students.

Table 5.9: Technical effort per group of participants: mean and standard deviation values for the total number of keystrokes, insertions and deletions (both texts averaged)

Participants	Total keystrokes		Insertions		Deletions	
	Mean	SD	Mean	SD	Mean	SD
Experienced translators	453	210	237	108	216	104
Novice translators	318	179	173	95	145	85

#### 4.2.3 Cognitive effort

As it emerges from Table 5.10, the novice translators triggered more fixations ( $M = 1208, SD = 374$ ) and longer gaze time ( $M = 6.85, SD = 2.14$ ) than the experienced translators ( $M = 1002, SD = 153$  and  $M = 5.80, SD = 1.11$ ). The differences in fixation count and total gaze time were both statistically significant ( $t(25) = -2.28, p = 0.02$  and  $t(29) = -1.95, p = 0.03$ ). No statistically significant difference was reported for mean fixation duration  $t(37) = 0.51, p = 0.30$ . Our findings corroborate the findings of Pavlović & Jensen (2009), who also found students to invest more cognitive effort into their translations than professionals.

Table 5.10: Cognitive effort per group of participants: Mean and standard deviation values of the fixation count, mean fixation duration and gaze time (both texts averaged)

Participants	Fixation count		Mean fixation duration (msec)		Total gaze time (mins)	
	Mean	SD	Mean	SD	Mean	SD
Experienced translators	1002	153	348.62	46.61	5.80	1.11
Novice translators	1208	374	341.63	39.15	6.85	2.14

### 4.3 Product analysis: Edits’ analysis

As it emerges from Table 5.11, the average WER score was lower for the novice ( $M = 0.19, SD = 0.07$ ) than the experienced translators ( $M = 0.28, SD = 0.10$ ), confirming the finding of Moorkens & O’Brien (2015) who found students to post-edit less the MT output. It should be noted that the difference in average WER score between novice and experienced translators was found to be statistically significant  $t(34) = 3.15, p = 0.002$ .

Table 5.11: Average WER for both texts per group

Participants	WER	
	Mean	SD
Experienced translators	0.28	0.10
Novice translators	0.19	0.07

Looking at the edit categories in Table 5.12, we observe that the number of unedited lexical items is higher in the case of the novice translators ( $M = 116, SD = 8$ ) compared to the experienced translators ( $M = 106, SD = 9$ ), with that difference being statistically significant ( $t(37) = -3.37, p = 0.0009$ ). The novice translators were also found to be more reluctant to change the word order ( $M = 4, SD = 4$ ) and the word form ( $M = 7, SD = 4$ ) in the text, as well as to rephrase the text ( $M = 8, SD = 6$ ) compared to the experienced translators ( $M = 6, SD = 5$  and  $M = 11, SD = 4$  and  $M = 13, SD = 8$  respectively). Unlike the difference in changing the word order ( $t(38) = 1.29, p = 0.10$ ), the differences



## 5 Investigating post-editing: A mixed-methods study in English-Greek

in changing a word form and in using a different lemma were statistically significant, ( $t(38) = 2.46, p = 0.009$  and  $t(34) = 2.21, p = 0.02$ ). Our findings seem to be in line with the observation made by Depraetere (2010) that during PE students follow the instructions given and do not rephrase the text if the meaning is clear, but do “not feel the urge to rewrite it” (ibid: 4), potentially leaving errors that should be corrected according to the instructions. Depraetere points out that this indicates a “striking difference in the mindset between translation trainees and professionals” (ibid: 6). In addition, as Yamada (2019) observed, a low error correction rate during PE the NMT output may be due the fact that NMT systems produce human-like errors, which make it more difficult for novice translators to post-edit.

Table 5.12: Average number and percentage of edits for both texts per edit category

	Translators	
	Experienced	Novice
unedited	106 (72.61%)	116 (80.10%)
form changed	11 (7.17%)	7 (5.13%)
word changed	13 (8.43%)	8 (5.23%)
deleted	5 (3.90%)	4 (2.55%)
inserted	6 (3.98%)	6 (4.05%)
order	6 (4.12%)	4 (2.94%)

Evaluating the correctness and necessity of each edit (Table 5.13), we noticed that the experienced translators perform more correct and necessary edits ( $M = 26, SD = 13$ ) compared to the novice translators ( $M = 17, SD = 6$ ) who tend to leave in their final post-edited text more errors that should be corrected (edits missing and required) ( $M = 18, SD = 5$ ), confirming, thus the findings by Depraetere (2010). In addition, the experienced translators seem to have a tendency to overcorrect the NMT output, thus carrying out more redundant edits ( $M = 8, SD = 5$ ) compared to novice translations ( $M = 5, SD = 7$ ). It should be noted that these differences were statistically significant  $t(28) = 3.13, p = 0.002, t(38) = -3.88, p = 0.0002$  and  $t(34) = 1.99, p = 0.03$  respectively.

Table 5.13: Average number and percentage of correctness and necessity for both texts per edit category

	Translators	
	Experienced	Novice
correct and necessary edit	26 (17.65%)	17 (11.09%)
correct and necessary no edit	95 (64.90%)	100 (68.08%)
incorrect and necessary edit	3 (2.18%)	5 (3.09%)
incorrect and unnecessary edit	3 (1.85%)	4 (2.64%)
edit missing and required	11 (7.72%)	17 (12.05%)
redundant and unnecessary edit	8 (5.69%)	4 (3.05%)

## 5 Conclusion and future work

The study confirms the findings of the previous studies on the more positive attitude of the novice translators vis-à-vis PE. Experienced translators exhibit a more negative stance vis-à-vis MT and adopt a more pragmatic approach to PE. However, this does not seem to affect the effort expended by the experienced translators when post-editing. In particular, the experienced translators expend less time and less cognitive effort during PE as opposed to the novice translators. On the other hand, the technical effort is found to be decreased in the case of the novice translators. This is due to the fact that they do not sufficiently rephrase the MT raw output as they are reluctant to change the word form, the word order and the syntax of the MT output and are not adequately critical of the content, thus leaving errors in the edited text. Finally, experienced translators have a tendency to overcorrect the NMT output, thus carrying out more redundant edits. These findings have several implications for the training of translators and their continuous professional development as they point to the need for a different approach when designing and delivering courses in PE. Experienced translators' training should aim at helping them appreciate the benefits of MT and PE and develop a more positive stance vis-à-vis MT and PE. In addition, their training should aim at helping them avoid the overcorrection of MT output. This can be achieved through extensive practical exercises, which focus on the identification of errors that require correction depending on the given PE guidelines, i.e. for full or light PE. Novice translators' training, on the other hand, should aim at helping them avoid the undercorrection of the MT output. Such training may include several practical exercises in error detection and correction.

Although the conclusions are clear and they point to several suggestions for the training of translators as outlined above, there are a number of limitations to this study. The main limitation involves the low ecological validity of the study, i.e. the fact that the experiments were carried out in an “artificial”, experimental situation rather than in a “natural”, real-world situation. In particular, the participants were asked to carry out the tasks at a research institute, the Hubic Lab in Athens, i.e. in an environment that differed from their usual work environment; they could not use any resources during the PE tasks, i.e. they could not use online or offline resources, such as dictionaries, termbases, parallel texts, etc., while they did not work in a translation memory environment. Finally, the error analysis was carried out by only one annotator and the sample size was small and consisted only of female participants. It is our intention in the future to address the present study’s limitations and carry out a more extensive research with a higher ecological validity and a more comprehensive and refined analysis of the edits performed by translators in order not only to understand the cognitive mechanisms behind their performance and the differences between the experienced and the novice translators, but also to improve work conditions and performance and thus enhance human-computer interaction.

## Acknowledgements

We would like to thank the HUBIC lab at the Athena research center in Athens for providing the Tobii X2-60 remote eye-tracker for the purposes of this study.

## References

- Allen, Jeffrey H. 2003. Post-editing. In Harold Sommer (ed.), *Computers and translation: A translator’s guide*, 297–317. Amsterdam: John Benjamins. DOI: 10.1075/btl.35.19all.
- Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéal, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor & Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the first conference on machine translation. Volume 2: Shared task papers*, 131–198. Berlin, Germany: Association for Computational Linguistics. DOI: 10.18653/v1/W16-2301.

- Carl, Michael. 2012. Translog-II: A program for recording user activity data for empirical reading and writing research. In *Proceedings of the eighth international conference on language resources and evaluation (LREC'12)*, 4108–4112. Istanbul, Turkey: European Language Resources Association (ELRA). [http://www.lrec-conf.org/proceedings/lrec2012/pdf/614\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2012/pdf/614_Paper.pdf).
- Carl, Michael & Matthias Buch-Kromann. 2010. Correlating translation product and translation process data of professional and student translators. In *Proceedings of EAMT*. Saint-Raphael, France. <http://www.lanaconsult.com/eamt2010/>.
- Carl, Michael, Barbara Dragsted, Jakob Elming, Daniel Hardt & Arnt Lykke Jakobsen. 2011. The process of post-editing: A pilot study. *Copenhagen Studies in Language* 41. 131–142.
- Carl, Michael, Silke Gutermuth & Silvia Hansen-Schirra. 2015. Post-Editing machine translation: Efficiency, Strategies, and Revision Processes in Professional Translation Settings. In Aline Ferreira & John W. Schwieter (eds.), *Psycholinguistic and Cognitive Inquiries into Translation and Interpreting*, 145–174. Netherlands: John Benjamins Publishing Company.
- Carl, Michael, Martin Kay & Kristian Tangsgaard Hvelplund Jensen. 2010. Long distance revisions in drafting and post-editing. In *Fremlagt på konferencen CICLing-2010*, Iași, Romania, March 21–27.
- Castilho, Sheila, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley & Andy Way. 2017. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics* 108(1). 109–120. DOI: 10.1515/pralin-2017-0013.
- Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Yota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Valerio, Miceli Barone & Maria Gialama. 2017. A comparative quality evaluation of PBSMT and NMT using professional translators. In *Proceedings of machine translation summit XVI*. Nagoya, Japan.
- Colina, Sonia & Claudia V. Angelelli. 2015. T and I pedagogy in dialogue with other disciplines. *Translation and Interpreting Studies* 10(1). 1–7. DOI: 10.1075/tis.10.1.01ang.
- De Almeida, Giselle. 2013. *Translating the post-editor: An investigation of post-editing changes and correlations with professional experience across two Romance languages*. Dublin City University. (Doctoral dissertation). [http://doras.dcu.ie/17732/1/THESIS\\_G\\_de\\_Almeida.pdf](http://doras.dcu.ie/17732/1/THESIS_G_de_Almeida.pdf).
- Densmer, Lee. 2014. *Light and full MT post-editing explained*. <http://info.moravia.com/blog/bid/353532/Light-and-Full-MT-Post-Editing-Explained>.

## 5 Investigating post-editing: A mixed-methods study in English-Greek

- Depraetere, Ilse. 2010. What counts as useful advice in a university post-editing training context? Report on a case study. In *EAMT 2010: Proceedings of the 14th annual conference of the European association for machine translation*. Saint-Raphaël, France.
- Doherty, Stephen, Sharon O'Brien & Michael Carl. 2010. Eye tracking as an automatic MT evaluation technique. *Machine Translation* 24. 1–13. DOI: 10.1007/s10590-010-9070-9.
- Dragsted, Barbara & Michael Carl. 2013. Towards a classification of translation styles based on eye-tracking and keylogging data. *Journal of Writing Research* 5(1). 133–158.
- Elming, Jakob, Laura Winther Balling & Michael Carl. 2014. Investigating user behaviour in post-editing and translation using the CASMACAT workbench. In Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard & Lucia Specia (eds.), *Post-editing of Machine Translation*, 147–169. United Kingdom: Cambridge Scholars Publishing.
- Flanagan, Marian & Tina Paulsen Christensen. 2014. Testing post-editing guidelines: How translation trainees interpret them and how to tailor them for translator training purposes. *The Interpreter and Translator Trainer* 8(2). 257–275. DOI: 10.1080/1750399X.2014.936111.
- García, Ignacio. 2010. Is machine translation ready yet? *Target* 22. 7–21. DOI: 10.1075/target.22.1.02gar.
- Gaspari, Federico, Antonio Toral, Sudip Kumar Naskar, Declan Groves & Andy Way. 2014. Perception vs reality: Measuring machine translation post-editing productivity. In *Proceedings of AMTA workshop on post-editing technology and practice*, 60–72. Vancouver.
- Göpferich, Susanne, Gerrit Bayer-Hohenwarter, Friederike Prassl & Stadlober Johanna. 2011. Exploring translation competence acquisition: Criteria of analysis put to the test. In Sharon O'Brien (ed.), *Cognitive Explorations of Translation*, 57–86. London: Continuum Studies in Translation.
- Hu, Ke & Patrick Cadwell. 2016. A comparative study of post-editing guidelines. In *Proceedings of the 19th annual conference of the European association for machine translation*, 346–353.
- Hvelplund, Kristian Tangsgaard. 2011. *Allocation of cognitive resources in translation: An eye-tracking and key-logging study* (PhD series 10.2011). Denmark: Samfundslitteratur.
- Jakobsen, Arnt Lykke. 2002. Translation drafting by professional translators and by translation students. In Gyde Hansen (ed.), *Empirical Translation Studies: Process and product*, 191–204. Copenhagen: Samfundslitteratur.

- Jia, Yanfang, Michael Carl & Xiangling Wang. 2019. How does the post-editing of neural machine translation compare with from-scratch translation? A product and process study. *The Journal of Specialised Translation* 31. 60–86.
- Kaiser-Cooke, Michèle. 1994. Translatorial expertise: A cross-cultural phenomenon from an interdisciplinary perspective. In Mary Snell-Hornby, Franz Pöchhacker & Klaus Kaindl (eds.), *Translation studies. An interdiscipline*, 135–139. Amsterdam, Philadelphia: John Benjamins Publishing Company.
- Koponen, Maarit. 2012. Comparing human perceptions of post-editing effort with post-editing operations. In *Proceedings of the seventh workshop on statistical machine translation (WMT '12)*, 181–190. Montréal, Canada: Association for Computational Linguistics. <https://www.aclweb.org/anthology/W12-3123>.
- Koponen, Maarit. 2016. Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. *The Journal of Specialised Translation* 25. 131–148.
- Koponen, Maarit & Leena Salmi. 2017. Post-editing quality: Analysing the correctness and necessity of post-editor corrections. *Linguistica Antverpiensia, New Series – Themes in Translation Studies* 16. 137–148.
- Koponen, Maarit, Leena Salmi & Markku Nikulin. 2019. A product and process analysis of post-editor corrections on neural, statistical and rule-based machine translation output. *Machine Translation* 33(1–2). 61–90. DOI: 10.1007/s10590-019-09228-7.
- Krings, Hans P. 2001. *Repairing texts: Empirical investigations of machine translation post-editing processes*. Vol. 5. Kent, OH: Kent State University Press.
- Lommel, Arle & Donald A. DePalma. 2016. *Europe's leading role in machine translation: How Europe is driving the shift to MT*. Tech. rep. Boston. <http://cracker-project.eu>.
- Mesa-Lao, Bartolomé. 2013. *Introduction to post-editing: The CasMaCat GUI*. [http://bridge.cbs.dk/projects/seecat/material/hand-out\\_post-editing\\_bmesa-lao.pdf](http://bridge.cbs.dk/projects/seecat/material/hand-out_post-editing_bmesa-lao.pdf).
- Mesa-Lao, Bartolomé. 2014. Gaze behaviour on source texts: An exploratory study comparing translation and post-editing. In Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard & Lucia Specia (eds.), *Post-editing of Machine Translation*, 219–245. United Kingdom: Cambridge Scholars Publishing.
- Mitchell, Linda G. 2015. *Community post-editing of machine-translated user-generated content*. Dublin City University. (Doctoral dissertation).
- Moorkens, Joss & Sharon O'Brien. 2015. Post-editing evaluations: Trade-offs between novice and professional participants. In İlknur Durgar El-Kahlout, Mehmed Özkan, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, Fred Hol-

## 5 Investigating post-editing: A mixed-methods study in English-Greek

- lowood & Andy Way (eds.), *Proceedings of the 18th annual conference of the European association for machine translation*, 75–81. Antalya, Turkey.
- Moorkens, Joss, Antonio Toral, Sheila Castilho & Andy Way. 2018. Translators' perceptions of literary post-editing using statistical and neural machine translation. *Translation Spaces* 7. 240–262. DOI: 10.1075/ts.18014.moo.
- Nitzke, Jean. 2019. *Problem solving activities in post-editing and translation from scratch: A multi-method study*. Berlin: Language Science Press.
- Nunes Vieira, Lucas, Elisa Alonso & Lindsay Bywood. 2019. Introduction: Post-editing in practice – process, product and networks. *The Journal of Specialised Translation* 31. 2–13.
- O'Brien, Sharon. 2009. Eye tracking in translation process research: Methodological challenges and solutions. In Inger M. Mees, Fabio Alves & Susanne Göpferich (eds.), *Methodology, technology and innovation in translation process research: A tribute to Arnt Lykke Jakobsen*, vol. 38 (Copenhagen studies in language), 251–266. Copenhagen: Samfundslitteratur.
- O'Brien, Sharon. 2010. *Introduction to post-editing: Who, what, how and where to next*. <http://amta2010.amtaweb.org/AMTA/papers/6-01-ObrienPostEdit.pdf>.
- O'Brien, Sharon & Michel Simard. 2014. Introduction to special issue on post-editing. *Machine Translation* 28(3–4). 159–164. DOI: 10.1007/s10590-014-9166-8.
- O'Brien, Sharon, Laura Winther Balling, Carl Michael, Michel Simard & Lucia Specia. 2014. *Post-editing of machine translation: Processes and applications*. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Palumbo, Giuseppe. 2009. *Key terms in translation studies*. London: Continuum.
- Pavlović, Nataša & Kristian T. H. Jensen. 2009. Eye tracking translation directionality. In Anthony Pym & Alexander Perekrestenko (eds.), *Translation Research Projects 2*, vol. 2, 93–109. Intercultural Studies Group.
- Raptis, Spyros & Maria Giagkou. 2016. From capturing to generating human behavior: Closing the interaction loop at the HUBIC lab. In *Proceedings of the 20th pan-Hellenic conference on informatics (PCI) with international participation*. Partas, Greece: ACM Digital Library, International Conference Proceedings Series. DOI: 10.1145/3003733.3003814.
- Schaeffer, Moritz, Jean Nitzke, Anke Tardel, Katharina Oster, Silke Gutermuth & Silvia Hansen-Schirra. 2019. Eye-tracking revision processes of translation students and professional translators. *Perspectives* 27(4). 589–603. DOI: 10.1080/0907676X.2019.1597138.
- Schmaltz, Marcia, Igor da Silva, Fabio Alves, Adriana Pagano, Derek Wong, Lidia Chao, Ana Leal, Paulo Quaresma & Caio Garcia. 2015. Translating and post-editing in the Chinese-Portuguese language pair: Insights from an exploratory

- study of Key logging and eye tracking. *Translation Spaces* 4. 145–169. DOI: 10.1075/ts.4.1.07sil.
- Sharmin, Selina, Oleg Spakov, Kari-Jouko Räihä & Arnt Lykke Jakobsen. 2008. Where on the screen do translation students look while translating, and for how long? In Susanne Göpferich, Arnt Lykke Jakobsen & Inger M. Mees (eds.), *Looking at Eyes: Eye-tracking studies of reading and translation processing*, vol. 36, 31–51. Samfundslitteratur.
- Shreve, Gregory M. 2002. Knowing translation: Cognitive and experiential aspects of translation expertise from the perspective of expertise studies. In Alessandra Riccard (ed.), *Knowing translation: Cognitive and experiential aspects of translation expertise from the perspective of expertise studies*, 150–171. Cambridge University Press.
- TAUS. 2016. *TAUS post-editing guidelines*. <https://www.taus.net/think-tank/articles/postedit-articles/taus-post-editing-guidelines>.
- Toral, Antonio & Víctor M. Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th conference of the European chapter of the Association for Computational Linguistics*, vol. Volume 1: Long papers, 1063–1073. Valencia, Spain: Association for Computational Linguistics.
- Wagner, Emma. 1985. Post-editing systran: A challenge for commission translators. *Terminologie et Traduction* 3. 1–7.
- Whyatt, Bogusława & Tomasz Kościuczuk. 2013. Translation into a non-native language: The double life of the native-speakership axiom. *mTm. A Translation Journal. Special Issue: Translation in an age of austerity* 5. 60–79.
- Yamada, Masaru. 2015. Can college students be post-editors? An investigation into employing language learners in machine translation plus post-editing settings. *Machine Translation* 29(1). 49–67. DOI: 10.1007/s10590-014-9167-7.
- Yamada, Masaru. 2019. The impact of Google neural machine translation on post-editing by student translators. *The Journal of Specialised Translation* 31. 87–106.



# Chapter 6

## The processing of website contents in native and non-native language

Jean Nitzke

University of Mainz & University of Adger

Eyetracking has been used widely to research the translation process in recent years. The reception of text in multimedia environments has also been studied with the help of eyetracking, where subtitles have been the focus of most studies. This paper presents a study which investigates the viewing behaviour and processing of materials' information (originals and translations) related to museum exhibitions by native and non-native speakers. The texts used in a museum context often address native and non-native readers to a similar extent. However, do we process information equally in our native and non-native language, assuming a very high language proficiency in the foreign language? The participants ( $n = 16$ ) read extracts of two Digitalorials® (one in German, one in English) provided on the website of the Schirn Art Gallery in Frankfurt. The participants' task was to prepare for a hypothetical test in the context of a cultural studies course (which are part of the translation degrees in Germersheim). The questions of the test were presented right after the participants had worked through the materials. The results show that the language in which the materials are presented has a statistically significant influence on the total fixation duration.

### 1 Introduction

Eyetracking has been used widely to research the translation process in recent years. The focus has been on general translation behaviour (e.g. Jakobsen & Jensen 2008, Dragsted & Carl 2013), the use and integration of tools (e.g. O'Brien 2006, Läubli et al. 2013), the use of machine translation (e.g. Doherty et al. 2010) and its post-editing effort (e.g. Moorkens et al. 2015, Nitzke 2019), or modelling



the translation process (e.g. Balling & Carl 2014, Carl & Schaeffer 2017). The reception of text in multimedia environments has also been studied with the help of eyetracking, e.g. to check the usability of websites (e.g. Nielsen & Pernice 2010). However, subtitles have been the focus of most translation related studies in a multimedia context so far. Orrego-Carmona (2015), for example, examined the reception of professional and non-professional subtitles (both in Spanish) for a US sitcom by 52 participants who had different levels of English proficiency. Amongst other results, the study showed that the kind of subtitles (professional vs. non-professional) did not influence the attention distribution, meaning that the participants spent an equal amount of time processing the subtitles or the image, irrespective of the kind of subtitle that was shown. However, the fixation duration was shorter overall on professional subtitles. In another study, Fox (2018) challenged the traditional positioning of subtitles, which are usually put at the lower part of the screen, and tested where they should ideally be placed so that participants do not spend too much time on finding and reading the subtitles instead of focusing on the images.

The perception of text and its translations in multimedia is, however, not restricted to films or even screens. In museums, for example, visitors use the information provided by written text, audio guides, or other digital aids to understand and contextualise the exhibits. Texts and audio information used in a museum context often address both native and non-native readers to a similar extent. However, do we process information equally in our native and non-native language, assuming very high language proficiency in the foreign language? The study at hand will investigate native and non-native language processing of materials that are related to exhibitions in an art gallery by analysing eyetracking data.

## **2 Setup of the experiment**

The study in this paper investigates the viewing behaviour and processing of information presented in museum materials by native and non-native speakers. I hypothesise that participants who are German native speakers read German materials faster and understand them better than English materials. This will be tested by analysing reading times and gaze behaviour of participants as well as the answers to comprehension questions.

Participants read extracts of two Digitorials®<sup>1</sup> (one in German, one in English) which are provided for different exhibitions on the website of the Schirn Art

---

<sup>1</sup><https://www.schirn.de/en/program/offerings/digitorial/>, last accessed 15th January 2019

Gallery in Frankfurt for free. The main purpose of the Digitalorials is to give extra information, either to prepare the visitors before their visit to Schirn, support them during the visit, or allow them to reflect on the exhibition after the visit. They are interactive websites which can be accessed via smartphone, tablet and desktop PC. However, I had to modify the material to balance the viewing experience with the effort of analysing the data, as the websites are too interactive to be processed by the website recording function in the analysis software (tobii studio) of the eyetracking system (tobii TX300). For example, some of the visual contents are in motion while the user is scrolling through the website. The created PDF files consist of screenshots of the Digitalorial contents and are still visually appealing. They have, however, lost their interactivity, which I considered acceptable for the research question in this study.

The participants ( $n = 16$ ) viewed excerpts of the Digitalorials for two exhibitions. The first dealt with the topic “wilderness” throughout the history of art, while the second introduced the Belgian painter René Magritte. Each participant had to read one excerpt in German and one in English. The participants were translation students who study English as their first or second foreign language. They had been learning English on average for 10.34 years ( $SD = 2.06$ ,  $\min = 8$ ,  $\max = 13$ ) so I could anticipate a high English language proficiency. They were all B.A. students at the FTSK in Gernersheim, University of Mainz, enrolled in a translation studies degree – on average in the 2nd semester ( $SD = 1.39$  semester). The participants’ task was to prepare for a hypothetical test for a course in cultural studies (included in the translation degrees in Gernersheim). The questions of the test followed right after the participants had worked through the materials. The questions were in German, which was the native language ( $n = 13$ ) or the first foreign language ( $n = 3$ )<sup>2</sup> for the participants. Therefore, the chance that they had problems understanding the question was kept to a minimum. None of them was an English native speaker. Afterwards, they had to assess how difficult they found the questions and how confident they were answering the questions.

### 3 Analysis and results

First, I will focus on the time spent reading the materials, because I hypothesized that it would take longer to process the non-native materials than the materials written in the participants’ native language (see e.g. Cop et al. 2015 for a reading study presenting similar results). The analysis shows that the reading time was not significantly different when materials were read in German (Wilderness:

---

<sup>2</sup>In any case, German was their dominant language compared to English

mean = 707s, SD = 154s and Magritte: mean = 769s, SD = 276s) or English (Wilderness: mean = 703s, SD = 88s and Magritte: mean = 902ms, SD = 292s) for both Digitalorials<sup>3</sup> (Wilderness:  $t(4.16) = 0.0406, p < 0.97$ ; Magritte:  $U = 9, p < 0.11$ ). Participants did not spend more time answering the questions when they read the texts in either language (Wilderness – German: mean = 300s, SD = 60s, English: mean = 335s, SD = 101s,  $t$ -test:  $t(8.91) = -0.73, p < 0.48$ ; Magritte – German: mean = 243s, SD = 45s, English: mean = 284ms, SD = 129s, Mann-Whitney- $U$ -test:  $U = 18, p < 0.73$ ).

Further, I hypothesized that the participants would answer more questions correctly when they read the materials in their native language. 11 out of 15<sup>4</sup> participants answered more questions correctly when they read the materials in German than in English (see Figure 6.1). On average, the participants answered 74.2% (SD = 12.3) of the questions correctly if they read the materials in German, while they answered only 63.4% (SD = 18.2) of the questions correctly when they were read in English. The difference is statistically significant (paired  $t$ -test;  $t = -2.67, p < 0.02$ ).

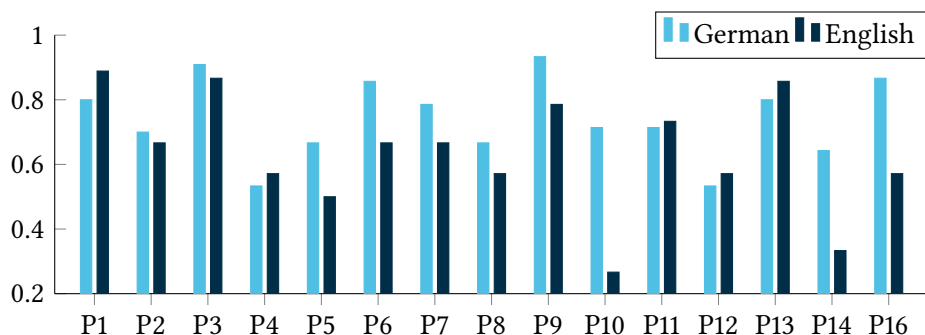


Figure 6.1: Proportion of correct answers per participant

P10 and P14 have very low scores for the English materials. They both read the Wilderness materials in English. However, participants’ answers were similarly often correct on average for the materials in German (Wilderness: mean = 0.73, SD = 0.24 and Magritte: mean = 0.75, SD = 0.24) and English (Wilderness: mean = 0.6, SD = 0.22 and Magritte: mean = 0.65, SD = 0.24). So, it can be ruled

<sup>3</sup>The data were tested for normal distribution with a Shapiro test for normal distribution. If they were normally distributed, a  $t$ -test was conducted. If not, a Mann-Whitney- $U$ -test was conducted.

<sup>4</sup>One of the participants had to be excluded from this calculation, because (s)he viewed both Digitalorials accidentally in German.

out that the text material was much more difficult in the foreign language than in the native language. Further, P10 and P14 claimed that they had learnt English for 10 and 14.5 years, respectively. As they answered the questions referring to the German materials quite well, it can be ruled out that the task itself was too difficult for them. However, reading the English Wilderness material was the first task they had to fulfill for both. Hence, they might have read the texts much less carefully than in the second task, although the reading times do not indicate this (P10 read both texts almost equally fast – 688s vs. 679s – and P14 was even faster with the Magritte text – 865s vs. 625s). However, they may have read more thoroughly or may have been more motivated than in the first tasks.

In the next step, I assessed the eyetracking data for the text parts that contained the answers to the questions. I hypothesised that the participants fixated on the English materials significantly longer and with a higher fixation count, because it would be cognitively more demanding to process the non-native language. For question 4 in the Wilderness dataset, for example, a linear regression showed that the total fixation duration was significantly higher when the participant answered the questions correctly ( $t(3.73) = -2.48, p < 0.03$ ), while the language in which the text was read had no influence ( $t(3.49) = 0.17, p < 0.87$ ), which contradicts my hypothesis. The same was true for the fixation count (correct answer:  $t(10.5) = -2.21, p < 0.05$ ; language:  $t(9.81) = 0.2, p < 0.85$ ).

All in all, I considered eight questions from both data sets. The questions were selected according to three criteria:

- whether all participants answered the respective question,
- whether there was a variety of correct and incorrect answers,
- whether the answer of the question could be bound to one text segment.<sup>5</sup>

The results for the single questions can be found in Table 6.1 for total fixation duration and Table 6.2 for fixation count, representing the values for correct vs. incorrect answers (*Corr*) and for the language of the materials (*Lang*).

For the single questions, only a few factors were significant, but no pattern is visible (see Tables 6.1–6.2). All 16 participants were considered for the initial evaluation to have more data points. However, when I combined the data of all eight

---

<sup>5</sup>Some questions, for example, referred to numerous text passages and/or pictures. Or they referred to an overall concept rather than a single text passage. Hence, AOIs cannot be drawn for certain text passages to answer the question.

Table 6.1: Results of linear regressions for single questions for Total Fixation Duration

	Corr		Lang	
	<i>t</i>	<i>p</i>	<i>t</i>	<i>p</i>
Q4_W	-2.48	0.03	0.17	0.86
Q6_W	1.36	0.2	1.24	0.24
Q9_W	-1.66	0.12	1.66	0.12
Q14_W	-0.14	0.89	-0.08	0.94
Q15_W	0.63	0.54	0.37	0.72
Q4_M	-1.85	0.09	2.56	0.02
Q6_M	-1.91	0.08	1.55	0.14
Q7_M	-0.49	0.63	1.12	0.28

Table 6.2: Results of linear regressions for single questions for Fixation Count

	Corr		Lang	
	<i>t</i>	<i>p</i>	<i>t</i>	<i>p</i>
Q4_W	-2.21	0.05	0.2	0.85
Q6_W	2.31	0.04	1.1	0.29
Q9_W	-1.44	0.18	1.99	0.07
Q14_W	-0.06	0.96	0.18	0.86
Q15_W	1.27	0.23	-0.23	0.82
Q4_M	-2.1	0.06	1.56	0.14
Q6_M	-1.2	0.25	0.21	0.84
Q7_M	-0.38	0.71	0.12	0.9

questions and assess the results of the linear regression, I excluded the three non-German-native participants as well as one data set with only data for reading behaviour in German (as already mentioned in the forth footnote). Let us first look at results for the correctness of the answer. The difference in the total fixation duration ( $t(2.81) = 0.3, p < 0.76$ ) as well as fixation count ( $t(5.88) = 0.43, p < 0.67$ ) on the AOIs are not significant in the model. The language, however, does have a significant influence on the total fixation duration ( $t(2.53) = 2.28, p < 0.03$ ) with the values significantly higher for English (mean = 20.41ms, SD = 14.75ms) than for German (mean = 14.72ms, SD = 9.14ms). The difference is not significant for fixation count ( $t(5.29) = 1, p < 0.32$ ). Accordingly, my hypotheses were only partly confirmed. However, the total gaze duration was significantly longer for the English texts than for the German, which implies more cognitive effort in processing the texts. As the gaze behaviour is not significantly different for the correct answers, the additional cognitive effort does not result in better answers, but in equally correct/incorrect answers.

## 4 Discussion and future research

The higher fixation duration on the English text was expected. The results suggest that although the participants need significantly more time to process the non-native text, they answered fewer questions correctly. This was only a preliminary study, but the results indicate that the participants process the contents

## 6 *The processing of website contents in native and non-native language*

faster when they read them in their native language and also answer more questions correctly. If the results can be confirmed in a larger study, it would indicate that it could be preferable to read and learn in the native rather than in the non-native language. In a follow-up study, more participants and a more controlled set-up will be needed.

Coming back to the museum situation, it is, obviously, helpful to have supporting text materials in a foreign language than none at all. However, these first results suggest that visitors can better process texts in their native languages. Hence, translations in several languages help in reaching more visitors, even if they have a high proficiency in one of the languages already offered. With the increasing worldwide digitalisation, it becomes easier and less expensive to present those texts, e.g. via apps and QR codes. Post-editing machine translation could also help to accelerate the translation process and decrease translation costs. However, to my knowledge, there are no studies yet on the quality of the output of current MT systems for museum texts.

A next step within this study will be to investigate if language has any influence on the perception of the picture elements in the materials. It would seem plausible, on one hand, that the participants would rather focus on the text in the non-native task, because the text is harder to process. On the other hand, more gazes on the pictures in the non-native task could also imply that the participants use the pictures as references to better process the text. Especially in an art gallery, the text is meant to supplement the art and not the other way around. Hence, the text should not distract from the art.

12 (or 16) participants is a rather small number to start with statistical investigations (O'Brien 2009). The study will be extended with more participants to make the results more reliable. Further, it would be very interesting to replicate the study with the interactive websites to see if this feature influences the reading behaviour.

Finally, we want to investigate the perception of text<sup>6</sup> in real-life museum environments in future research by measuring eye movements with eyetracking glasses. The aim will be to study how translated and non-native language influences the viewing behaviour. Mobile eye trackers have been used to explore the viewing behaviour on single exhibits (e.g. Walker et al. 2017, Tatler et al. 2016) and for the whole visitor experience (e.g. Eghbal-Azar et al. 2016). However, the techniques have not been applied to the ways in which the translation of printed material and audio-guide content might direct the visitor's gaze, and, subsequently, their response. Written and spoken texts are important features

---

<sup>6</sup>Maybe also considering the influence of audio guides on the reading and viewing behaviour.

of every exhibition, either as written explanations to the exhibits or as spoken audio-guides, if they are not the exhibits themselves, e.g. when important documents are on display or when speeches or videos are integrated with the exhibition. Often, these various kinds of texts are translated to make them accessible to a broader audience. Any number of translation strategies, including lexical choices, text size and acoustic features, might guide the visitor towards particular interpretations, influence their selection of points of interest and dwell time, and, ultimately, shape their cognitive and/or effective engagement with the museum, which might change depending on whether the text is written in the native or a non-native language of the visitor. However, the perception of neither texts in general nor translations has been studied in the museum context with the help of eye tracking, yet. Hence, this study would fill the gap.

## References

- Balling, Laura Winther & Michael Carl. 2014. Production time across languages and tasks: A large-scale analysis using the CRITT translation process database. In Aline Ferreira & John Schwieter (eds.), *Psycholinguistic and cognitive inquiries in translation and interpretation studies*, 239–268. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Carl, Michael & Moritz Schaeffer. 2017. Sketch of a noisy channel model for the translation process. In Silvia Hansen-Schirra, Oliver Czulo & Sascha Hofmann (eds.), *Empirical modelling of translation and interpreting*, 71–116. Berlin: Language Science Press. DOI: 10.5281/zenodo.1090954.
- Cop, Uschi, Denis Drieghe & Wouter Duyck. 2015. Eye movement patterns in natural reading: A comparison of monolingual and bilingual reading of a novel. *PloS one* 10(8). 1–38.
- Doherty, Stephen, Sharon O'Brien & Michael Carl. 2010. Eye tracking as an MT evaluation technique. *Machine Translation* 24(1). 1–13.
- Dragsted, Barbara & Michael Carl. 2013. Towards a classification of translation styles based on eye-tracking and keylogging data. *Journal of Writing Research* 5(1). 133–158.
- Eghbal-Azar, Kira, Martin Merkt, Julia Bahnmueller & Stephan Schwan. 2016. Use of digital guides in museum galleries: Determinants of information selection. *Computers in Human Behavior* 57. 133–142.
- Fox, Wendy. 2018. *Can integrated titles improve the viewing experience?: Investigating the impact of subtitling on the reception and enjoyment of film using eye tracking and questionnaire data*. Berlin: Language Science Press.



- Jakobsen, Arnt Lykke & Kristian T. H. Jensen. 2008. Eye movement behaviour across four different types of reading task. In Arnt Lykke Jakobsen, Susanne Göpferich & Inger Margrethe Mees (eds.), *Looking at eyes: Eye-tracking studies of reading and translation processing*, 103–124. Copenhagen: Samfundslitteratur.
- Läubli, Samuel, Mark Fishel, Gary Massey, Maureen Ehrensberger-Dow & Martin Volk. 2013. Assessing post-editing efficiency in a realistic translation environment. In *Proceedings of MT summit XIV workshop on post-editing technology and practice*, 83–91.
- Moorkens, Joss, Sharon O'Brien, Igor A. L. Da Silva, Norma B. De Lima Fonseca & Fabio Alves. 2015. Correlations of perceived post-editing effort with measurements of actual effort. *Machine Translation* 29(3). 267–284. DOI: 10.1007/s10590-015-9175-2.
- Nielsen, Jakob & Kara Pernice. 2010. *Eyetracking web usability*. Thousand Oaks: New Riders.
- Nitzke, Jean. 2019. *Problem solving activities in post-editing and translation from scratch: A multi-method study*. Berlin: Language Science Press.
- O'Brien, Sharon. 2006. Eye-tracking and translation memory matches. *Perspectives: Studies in translatology* 14(3). 185–205.
- O'Brien, Sharon. 2009. Eye tracking in translation process research: Methodological challenges and solutions. *Methodology, Technology and Innovation in Translation Process Research* 38. 251–266.
- Orrego-Carmona, David. 2015. *The reception of (non) professional subtitling*. Universitat Rovira i Virgili. (Doctoral dissertation).
- Tatler, Benjamin W., Ross G. Macdonald, Tara Hamling & Catherine Richardson. 2016. Looking at domestic textiles: An eye-tracking experiment analysing influences on viewing behaviour at owlpen manor. *Textile History* 47(1). 94–118.
- Walker, Francesco, Berno Bucker, Nicola C. Anderson, Daniel Schreij & Jan Theeuwes. 2017. Looking at paintings in the Vincent van Gogh museum: Eye movement patterns of children and adults. *PLoS One* 12(6). e0178912.



## Chapter 7

# Assessing indicators of cognitive effort in professional translators: A study on language dominance and directionality

Aline Ferreira<sup>a</sup>, Stefan Th. Gries<sup>a,b</sup> & John W. Schwieter<sup>c</sup>

<sup>a</sup>University of California, Santa Barbara <sup>b</sup>Justus Liebig University Giessen

<sup>c</sup>Wilfrid Laurier University

Recently in translation studies, important advances have been made with respect to directionality (i.e., whether translation is done into one's native or non-native language). What was once considered the “elephant in the room,” directionality now has a growing number of empirical studies that analyze factors which contribute differentially to translation. In this chapter, we review variables that have been previously identified as related to a higher or a lower degree of cognitive activity in direct and inverse translation (DT and IT, respectively). Against this backdrop, we present a study conducted among professional translators of English and Spanish who completed two translation tasks: one in which they translated a text from English into Spanish and another in which they translated another text from Spanish into English. We use behavioral and eye-tracking measures to analyze time, mouse events, keypresses, saccade index, and gaze index data. We also explore the effects of age and sex/gender. The results suggest that in terms of length, although translators spent longer in IT compared to DT, this difference was not statistically significant. However, there was a correlation between translation direction and fixation index such that participants showed a higher gaze event duration in IT. Age was correlated to fixation index (lower fixation index among older translators) and sex/gender was also related to fixation index (females presented lower values in IT in comparison to DT). Results also suggested a higher gaze point index in IT and a higher keypress index for English-dominant translators, and a higher gaze point index in DT for Spanish-dominant translators. Overall, our study suggests that although some of the variability in the results is likely due to individual differences, the observed patterns help us better understand differences between DT and IT.



## **1 Introduction**

Translating from one language into another and vice versa has been discussed from several perspectives (see Ferreira & Schwieter 2017 for a review). In the field of translation studies, researchers interested in directionality often analyze behavioral patterns at the discourse level (writing and reading mechanisms, questionnaires, think aloud protocols, teaching practices, etc.). In psycholinguistics, researchers commonly study how bilinguals perform translation tasks at the word level (word translation task, lexical decision, etc.). In this chapter, we draw on previous work in both psycholinguistics and translation studies to present a study which explores directionality among professional translators. Below we first discuss the theoretical and empirical foundations that help inform our work. We then present our study, hypotheses, participants, design, and procedures. Finally, we discuss the results and offer some implications for future work.

## **2 Theoretical and empirical foundations: Psycholinguistic studies on word-level translation**

In psycholinguistics and bilingualism, our understanding of how words are translated from one language to another has been greatly informed through developmental models. One such example is the revised hierarchical model (RHM; Kroll & Stewart 1994), a theoretical account of the bilingual mental lexicon explaining the differences between second (L2) to first (L1) language translation (i.e., direct translation; DT) and L1 to L2 translation (inverse translation; IT). In their study, the researchers compared native English-speaking students to Dutch-English bilinguals who first performed a picture naming task and then a word translation task, both of which were presented in either semantically-related or unrelated lists. The results showed that performance was slower in the categorized lists compared to the randomized lists. Importantly, the category interference effect in the picture naming task was eliminated when the task alternated with word naming, suggesting that in both picture naming and word translation tasks a conceptual representation is used to retrieve a lexical entry. “When conceptual activity is sufficiently great to activate a multiple set of corresponding lexical representations, interference is produced” (Kroll & Stewart 1994: 149). Category interference in word translation occurred only during IT, suggesting that both directions of translation engage different interlanguage connections. The RHM argues that there is a differential relationship between concepts and the L1 and L2 words mapped onto them that is sensitive to language dominance.

As proficiency increases in a language, the words in that language grow stronger connections with the concepts they represent. Further support for the model was reported in a study with unbalanced bilinguals who performed a word translation task, both in direct and inverse directions (De Groot et al. 1994). The predictor variables were imageability, context availability, definition accuracy, familiarity, frequency, length, and cognate status. The results showed that both directions were influenced by meaning variables, familiarity variables, and cognate status. Semantic variables played a more important role in DT compared to IT, supporting the RHM.

Ferreira & Schwieter (2014) compared L3-to-L1 and L1-to-L3 word translation by analyzing the semantic relatedness effects (translation facilitation or interference) that potentially arise in both directions. Specifically, the study investigated whether such effects are modulated when to-be-translated words are restricted to the same semantic category versus belonging to various categories. The results suggested that semantic relatedness effects manifest themselves in different ways depending on translation direction and semantic restrictedness of the to-be-translated items. The findings were in line with the predictions of the RHM with respect to less-proficient language learners whose weaker language's words are mediated through translation equivalents in the stronger language.

Klein et al. (1995) conducted a study to investigate word generation in English-French bilinguals, who performed three tasks: rhyme generation based on phonological cues, synonym generation requiring a semantic search, and word translation involving access to a semantic representation in the other language. Using positron emission tomography (PET), they investigated whether phonological and semantic word-generation activate similar regions and whether the same neural substrates underpin both languages. Results indicate that common neural substrates are involved in within- and across-language searches. Furthermore, the left inferior frontal region showed activation irrespective of whether the search was guided by phonological or semantic cues. In a follow-up study, La Heij et al. (1996) investigated whether word translation is based on word associations at a lexical level. They conducted four Stroop-like experiments in which a to-be-translated word was accompanied by a color or a picture. Results showed that effects were no larger in DT in comparison to IT as predicted by the RHM, but semantic context had a larger effect on DT compared to IT. The researchers argued that both DT and IT are largely conceptually mediated and concept activation is easier for L1 words than for L2 words.

Price et al. (1999) conducted a study with German-English adult bilinguals who were scanned while translating or reading words in German, English, or switching between German and English. Results showed that word translation

increased activity in the anterior cingulate and subcortical structures while decreasing activation in several other temporal and parietal language areas associated with the meaning of words. According to the authors, a possible explanation is that their participants were highly proficient bilinguals and therefore able to translate using the direct route, without semantic involvement.

Quaresima et al. (2002) carried out a study with English-Dutch students proficient in English who translated short sentences aloud from Dutch into English, from English into Dutch, and switching between English and Dutch. The study aimed at investigating the organization of language in the brain by using PET and functional magnetic resonance imaging (fMRI). Results showed that Broca's area is involved in the translation process. Furthermore, Broca's area activation is unaffected by the direction of the translation.

Duyck & Brysbaert (2004) conducted a study with Dutch-French bilinguals from birth and Dutch native speakers who started to learn French at school between 10 and 13 years of age to investigate whether translation of number words is semantically mediated or based on word associations at the lexical level. Results showed that, in both DT and IT, there is a semantic number magnitude effect as it takes longer to translate number words that represent large quantities than small quantities. According to the authors, at least for certain types of words, "the mappings between L2 words and their meaning are more important than the intralexical mappings between the L2 words and their L1 equivalents, already from the first stages of L2 acquisition" (p. 904).

### **3 Translation studies and higher-level translation**

As noted above, much work conducted in psycholinguistics has focused on word translation. In translation studies, researchers are often more interested in high-level translation processes (i.e., sentence and discourse level) and individual translator characteristics. For instance, Pokorn (2004) conducted a study to investigate to what extent native speakers of English could identify a native translator vs. a non-native translator and a single translator vs. a team of translators in a set of translations from Slovene into English. The results showed that native speakers of English were not always able to discriminate between native and non-native translators, or between single translators and a team of translators. Bartłomiejczyk (2004) employed a survey testing interpreting students' and professional interpreters' preferences for DT or IT. Results showed that professionals preferred to work into their mother language whereas students' preferences were not clear. Pavlović (2007) also conducted a questionnaire to inquire about

translators' and interpreters' preferences on language direction. While 20 participants reported that they preferred DT, 21 said that they preferred IT, and 20 said that they had no preference regarding directionality.

Pavlović & Jensen (2009) reported on an eye-tracking study with students and translators that investigated cognitive effort in processing source and target texts (ST and TT, respectively) and cognitive effort in DT and IT by analyzing gaze time, average fixation duration, total task length, and pupil dilation. Results showed that TT requires more cognitive effort than ST. For both groups, IT lasted longer than DT and pupil dilation values were higher in the IT than in the DT. Average fixation was higher in the group of professionals in the IT compared to the DT, while students presented a higher average fixation in the DT. Gaze time values were higher in the DT for both groups. Students presented higher average fixation durations in the DT compared to IT, but professionals presented higher values during IT. Professionals presented slightly higher pupil dilation values in the DT compared to the IT, whereas students presented a higher value in the IT compared to the DT. Although their findings were interesting, it is "premature to draw any definitive conclusions" (p. 108) given the very small sample size ( $N = 8$  professional translators vs.  $N = 8$  student translators).

More recently, Whyatt & Kościuczuk (2013) questioned whether "translation in the age of austerity is ready to abandon one of its major axioms, namely that professional translating should not be done into the translator's non-native language" (p. 60). These concerns were echoed by Ferreira's (2013) study which discussed directionality in translation and how assumptions have mostly been made based on the belief that DT is superior to IT. In Whyatt and Kosciuczuk's study, the researchers cross-examined the theoretical assumptions and recommendations about the translation job market and also the professional practice in the minor-major language combination. It is noteworthy to mention that in countries with languages of limited diffusion (e.g., Brazil, Hungary, Denmark, etc.), IT is carried out on a daily basis (Ferreira 2013; Pokorn 2004, 2005). Whyatt and Kosciuczuk conducted a survey among professional translators in Poland to understand the relationship between the assumption that translators should only work into their L1 and the reality – that translators (often) work into their L2 as well. The authors pointed out how existing translation competence models "do not place significant value on the requirement" (p. 60). In both study and practice, IT may be an uncomfortable situation. For instance, in the industry, translators are asked to carry out IT even though it is openly stigmatized. In academia, it is under-researched and, to some extent, still a taboo. Pavlović (2007) carried out a study to examine the situation of IT in Croatia. Questionnaires were completed by 193 respondents. As is the case in many places, the profession is not "very

well defined in Croatia and there are no translator training institutions as such” (p. 86). Pavlović inquired about 12 languages that are used by those professionals who completed the questionnaires, and explained that “by far, the largest group was that consisting of people who work with L1 Croatian and L2 English, without an L3” (p. 87). Over 50% reported translation/interpretation as a part-time job. As per their attitude regarding the difficulty of L2 translation/interpreting, most of them (61 individuals) reported that working into their L1 felt easier than working into their L2 (27 individuals). On a scale of 1 (strongly disagree) to 5 (strongly agree), the translators were also asked about Newmark’s (1988) statement that translating into the L1 is “the only way you can translate naturally, accurately, and with maximum effectiveness” (p. 3). The majority (30%) said that they “neither agree nor disagree,” while 42% either agreed or agreed strongly, and 20% disagreed or disagreed strongly. Interestingly, the participants were professionals who translate into their L2 on a daily basis, showing the contrast between what they believe and what they practice.

Ferreira et al. (2016) carried out an eye-tracking study in which professional translators (two English-Spanish and two Spanish-English bilinguals) translated one text from Spanish into English and another text from English into Spanish. To investigate the effects of directionality, the researchers analyzed total task length, fixation count, average fixation duration, and gaze time. The results showed that translators spent more time in IT compared to DT. They also presented higher fixation counts in IT, which also had a higher average fixation duration. Analyses were also conducted on dwell time on target text, source text, and internet browser areas. The findings showed that in both tasks, translators tended to present longer dwell time in the source text compared to the target text and the internet browser areas. Although Ferreira et al. had predicted that the dwell time would be higher for the internet browser during IT, results showed a higher dwell time in DT. They explained this finding by elaborating that translators are more critical of lexical decisions in their L1. The study provided some preliminary insights, albeit based on a very small sample, as to why there are differences between DT and IT processes.

While directionality continues to be under investigated, a few works have underscored the importance and common practice of IT in countries with languages of limited diffusion (Pavlović 2010; Ferreira 2013; Whyatt 2018; see Ferreira & Schwieter 2017 for a review). In the next sections, we discuss variables such as language dominance, experience, age, and sex/gender and why they should be studied when exploring directionality in translation at the discourse level.



## 4 Language dominance

Language dominance and experience in translation have been assessed to explain their possible cognitive effects during translation tasks at the word level and at the discourse level. Also, studies have been conducted to investigate possible differences among bilingual participants at different levels of proficiency when performing different linguistic (e.g., word translation tasks) and nonlinguistic tasks (e.g., memory tasks). García (2015) presents a review on psycholinguistic research on lexical translation equivalents. Spanning over 30 years of research, García identifies three stages in the development of the field: the foundational era, the take-off era, and the ongoing expansion era. In the first era, models of interlinguistic associations were presented. In the second era, empirical experiments aiming at assessing conceptual representations in DT were developed. Later, in the ongoing expansion era, the RHM was introduced, triggering several studies on whether word translation is modulated by directionality, L2 competence, and semantic relatedness (e.g., level of concreteness and cognate status). García explains that the impact of translation expertise on word translation and the exploration of the neural basis of translation had an important impact on studies on cognitive translatology. One such example was a study by García et al. (2014) which investigated how non-translators with different levels of L2 proficiency perform word reading and translation tasks. Participants had different levels of informal translation experience and also different levels of translation training. Results showed that word reading was faster than translating, and also unaffected by concreteness and cognate effects. In the word translation task, participants translated concrete and cognate words faster than abstract and non-cognate words. Bilingual isolated-word processing does not seem to be affected by translation training. However, previous studies have showed a causal relationship between L2 competence and directionality effects, and vice versa.

In another study, Guasch et al. (2008) tested beginning and intermediate Spanish-Catalan learners and highly proficient bilinguals to see whether L2 proficiency determines how lexical and semantic representations are functionally connected in bilingual memory. Form and semantic manipulations were analyzed in a word translation task with very close and close semantically-related word pairs and form-related pairs. Results showed that the influence of semantic relatedness depends on participant's level of proficiency. Furthermore, results also showed that form manipulation affects the performance of all groups. In a similar vein, Christoffels et al. (2003) conducted a study with untrained bilinguals to assess memory and lexical retrieval. Participants performed a reading span task in two languages and a verbal digit span task in their L1 to assess mem-

ory capacity. They also performed a picture naming and a word translation task to assess lexical retrieval time in both languages. Results showed a correlation between interpreting performance and word translation and picture naming latencies. Furthermore, digit and reading spans were associated with interpreting performance.

More recently, López & Vaid (2018) conducted a study with Spanish-English bilinguals who were divided into two groups (brokers vs. non-brokers) to measure conceptual divergence in bilinguals with informal translation experience. Participants provided exemplars for 10 categories, using the same or different language across sessions. Half of the items were tested in the same language twice and the other half were tested in different languages across test sessions which were separated by one week. The results showed a higher overlap in category exemplars when they performed the task in the same language across sessions than when they performed it in different languages across sessions. However, prior experience in informal translation did not affect results as there was no significant effect of group nor a group by condition interaction.

As Lörscher (1991a) states, it is sometimes assumed that “bilinguals take a specific approach to translation and/or are in possession of a special competence for translating” (p. 3). Lörscher questioned to what extent the two languages favor or hinder translation. In his project, mental representations of bilinguals’ two languages was the focus of the paper – an area of inquiry that is commonly studied in bilingualism (Paradis 1985, and see Kroll 2008 for a review). It might be the case that researchers in translation simply assume that translators possess such a high level of knowledge in both languages that such competence should not be taken into account when analyzing their translation processes: the way that they deal with translation problems and solutions would be independent of language competence. However, a lack of evidence with respect to bilingual competence in studies that investigate translation process could possibly lead to a misinterpretation of data. There is not enough information on participants’ language competence in previous work and accordingly, we might assume that they have very similar levels of language skills in both languages. As stated by Grosjean (2001), “...interpretation and translation entail that one has identical lexical knowledge in the two languages, something that most bilinguals do not have” (p. 11). Lörscher explains that translation competence is the result of a developmental process that is never final. He points to the fact that an innate predisposition is not controversial in translation theory. What is controversial is “the way translation competence develops from an individual’s innate predisposition” (p. 5). The author also describes the concept of a rudimentary ability to mediate (Lörscher 1991a,b) which assumes that “every individual who has a command

of two or more languages is also endowed with a rudimentary ability to mediate information between these languages” (Lörscher 2014: 6). Translation competence, in this sense, must be acquired and there is a consensus among scholars in translation studies (see also PACTE 2005 for a review). Lörscher states that the assumed rudimentary ability to mediate between languages results in performance products, or translations, even though they are imperfect or restricted. Therefore, this ability is a special case of at least two universal innate abilities of the human intellect: categorizing and comparing and differentiating similarities and dissimilarities.

The issue might be due to the fact that bilinguals present different levels of competence in each language. In other words, the notion of the “ideal” translator or interpreter is more a utopian belief than a reality, especially when the focus is on languages of limited diffusion. The “situation on the ground” in the translation market (at least in Brazil, Canada, and the United States) is that people with different profiles, educational levels, and L1s start to work as translators and interpreters and create their network to survive in an informal market. It is evident that more research is needed to scrutinize how language dominance might affect translation processes in both languages.

## 5 Age and sex/gender

Age-related differences in translators as well as sex/gender have not been broadly discussed in translation studies. As discussed by Torgrimson & Minson (2005), the term “gender is becoming more common in scientific publications to describe biological variation traditionally assigned to sex ... Sometimes “gender,” sometimes “sex,” and sometimes “gender/sex” is used to describe the recent advances that physiologists are making in recognizing the important implications of sex differences on all physiological systems” (p. 785). In the scope of this paper, we are taking into account the biological differences between “male” and “female” and ask whether this variable is related to DT and IT processes. Although age and sex/gender are not the main focus of this paper, it is worth looking at the results to see whether those variables should be taken into account in further studies in translation studies and importantly, in studies that investigate directionality. We report age and gender as control variables in our regression models.

## **6 Language experience**

As presented by Shreve (2006), “the cognitive resources that underlie expertise arise from the operation of pattern recognition, problem representation, “chunking,” schematization, and knowledge proceduralization processes on the contents of episodic memory over long periods of deliberate practice” (p. 27). An individual performing a translation, according to Shreve, brings multiple translation-relevant cognitive resources, referred to as translation competence. Because the use of those resources varies among individuals, they would perform differently when carrying out a translation task. Furthermore, not only the use of the cognitive resources may vary but also their linguistic competence, that is their “knowledge of their language, which is sharply distinguished from their performance” (Malmkjær 2009: 122; see also Chomsky 1965 for more details).

Göpferich (2009) presented a brief survey of how translation competence and its acquisition have been elaborated. In it, she presented a model of translation competence based on a longitudinal study with 12 students of translation over a period of three years. Relevant to our study is Göpferich’s view on the communicative competence in at least two languages – similar to PACTE’s (2005) definition of bilingual sub-competence, comprising lexical, grammatical, and pragmatic knowledge in both languages:

Pragmatic knowledge also includes knowledge about genre and situation-specific conventions in the respective cultures. Communicative competence in the source language is relevant primarily for source-text reception, whereas target-language competence determines the quality of the target text produced. Target-language receptive competence must not be neglected, however, because it is needed for monitoring processes in which source-language units and target-language units are compared for semantic equivalence, for example (p. 21).

In Göpferich’s (2009) study, good or very good grades in German and English A-level courses were required during the longitudinal study, which is a way to control for communicative competence in the languages, even though, as explained by the researcher, it is a more or less controlled variable. As in any experimental study with human subjects, idiosyncratic aspects cannot be discarded as participants have individual experiences such as studying/living abroad that shape their language competence and even intelligence and psychomotor skills. However, there were no further details on participants’ language skills, and the

focus was more on translation competence, and particularly on strategic competence, translation routine activation competence, and tools and research competence. This selection, according to Göpferich, represents the main “translation-specific competences in which translation competence differs from the competence of bilingual persons with no specific training in translation” (p. 29).

While language experience or competence are not common variables in studies on translation processes, translation experience has been explored in different contexts. Braga (2007) investigated the novice-expert translator continuum by using selected texts which belonged to different text types and presented a sequence of logic-semantic relations there were progressively complex (see also Halliday & Matthiessen 2014). Braga analyzed units associated with longer pauses during the tasks and also the time allocated to each stage of the translation process (orientation, draft, and revision; Jakobsen 2002, 2003), instances of meta-reflection in verbal protocols that could elicit the participants’ strategies. Braga also looked at the impact of the course on the students’ translation process and product. Results suggested an allocation of the students in different stages of the novice-expert continuum (intermediate and novice).

De Almeida & O’Brien (2010) investigated correlations between post-editing (PE) performance and previous translation experience. Some of the difficulties of measuring such variables derive from the fact that PE is a fairly new area of inquiry, with little training available and a great deal of variation from company to company. Furthermore, there are no internationally-adopted standardized quality metrics yet. In this case, not only do the participants’ profiles vary, but also their training on PE impacts the expected results, considering that the translators have to possess the ability to decide quickly whether the segment provided by the machine translation tool is useful or not. Three participants for French and three for Spanish were selected among the translators who took part in their project and who had different levels of professional experience (in years). Four of these participants had previous experience with PE and two had none. The findings suggested that the two most experienced translators, in number of years, for both languages were also the fastest post-editors and also made the highest number of what the authors considered essential changes. The two fastest post-editors also corrected nearly all errors. The findings also implied that experience may lead to a propensity to implement a higher number of stylistic changes.

Rothe-Neves (2002) analyzed the correlation between working memory, L2 dominance, experience with computers, and performance features of reading, writing, and translating. Participants were six undergraduate students and six professional translators. A copy task, a writing task, a reading task, and a translation task were carried out. In the copy task, participants copied 16 sentences

in the L1 and were asked to ensure that the sentences were typed correctly. As per the writing task, participants were asked to describe, in their L1, a detailed picture (depicting four males, three children and one adult, playing at the beach) that was shown to them. In the reading task, participants were asked to read the first page of the first chapter of *Pride and Prejudice* by Jane Austen. The software Leonline was used during the reading task, showing syntactical segments. Participants filled in a questionnaire about their language comprehension, interpretation, empathy, and literary knowledge. After the reading task, they were asked to translate the text from their L2 (English) into their L1 (Portuguese) by means of Writelog. After the translation tasks were finished, participants were asked about the translation difficulty level. Reading and translation time were both used as a measure of difficulty level as well. A battery for working memory measures was used (BAMT-UFMG, Wood et al. 2001). Participants were also asked about their L2 level in terms of formal instruction and any studying/living abroad experience. Dictionary searches were also registered and analyzed as a predictor of L2 experience. The copy task was used to predict their experience with computers. There were not significant differences based on sex/gender and participants were analyzed as part of the same group. Results showed that working memory was related to performance in both efficiency and quality. There was a relationship between working memory and reading. Professionals and students showed a different profile in all tasks. In terms of working memory, processing speed was more prominent among participants, in addition to activity coordination capacity. The author suggested that working memory is related to translation but it is not the main cognitive characteristic required for such a task. Translation, according to Rothe-Neves (2002), is a complex task that involves several skills that require further investigation. In this sense, it might be fruitful to combine less investigated variables (e.g., sex/gender and age) with variables that have been scrutinized more often in translation studies (e.g., time, keypress events, mouse events, etc.) to better understand how translators perform a task at hand and whether there is a change in their behavior depending on the direction of the translation.

## **7 The present study: Hypothesis**

Using eye-tracking technology and behavioral measures during IT and DT tasks, we compared two groups of translators. We assessed the potential effects of directionality along with language dominance and experience as translators and investigated variables indicative of cognitive effort, such as pupil dilation, saccades, mouse events, and keypress events. We hypothesize that IT will take longer than

DT and that mouse clicks, keypresses, fixation index, saccade index, gaze duration, and gaze index will be higher in IT.

## 8 Design and procedure

Twelve ( $N = 12$ ) English native-speaking translators and twenty ( $N = 20$ ) Spanish native-speaking translators were recruited from an agency in Los Angeles, California. The sample consisted of 22 females and 10 males and ranged in age from 25 to 81 years old ( $M = 48.43$ ,  $SD = 14.20$ ). All translators lived in the United States. Nineteen ( $N = 19$ ) participants were not born in the United States, coming from several different countries such as Spain, Argentina, Mexico, Peru, Paraguay, El Salvador, and Colombia. In the group of English-dominant, three participants were not born in the USA, arriving in the country at the ages of 10, 17, and 20 ( $M = 15.66$ ). In the Spanish-dominant group, the ages of immigration vary from 13 to 35 ( $M = 22.62$ ). Those participants have been living in the United States from 5 to 57 years ( $M = 20.82$ ). They were also asked about the percentage of time they normally spend translating into their L1 (i.e., DT). The English dominant group reported an average of 60.66%, and the Spanish-dominant group estimated an average of 60%. Finally, they were asked about the use of their L1 and L2 in their daily lives. The English dominant group reported using English 64% of the time in interactions at work or outside work (e.g., with family, friends, at stores, church, etc.) and using Spanish 30.08% of the time, and also using another language(s) 3.2% of the time. The Spanish-dominant group reported using Spanish 40.58% of the time, English 56.41, and another language(s) 3% of the time. They were hired from a prestigious translation agency in Los Angeles, California, which tested translators' written and oral skills before hiring them. Participants have been working as professional translators for at least five years.

Participants were tested individually and reported no discomfort from the procedure. They were allowed to take breaks whenever they felt necessary. They first signed the consent form and filled in a demographics and a language questionnaire. After being informed about the procedure and becoming familiarized with the keyboard and the internet browser, the participants translated one text from Spanish into English on a research-dedicated laptop. After finishing the first translation task, participants were asked to comment on their translation problems and solutions during the process. Translators then performed three psycholinguistic tasks that are part of a separate study and will be analyzed on another occasion. Even though participants had been told to take breaks whenever they felt necessary, a five-minute break was compulsory after finishing the psycholinguistic tasks. Only then did participants translate the text from English

into Spanish and the retrospective protocol was recorded. Translog<sup>1</sup> was used to register the keyboard movements and a screen-based remote eye-tracker device (Tobii Pro X2-30) was used to record eye movements for fixation- and gaze-based analyses. Both source texts (in Spanish and in English) were taken from Ferreira et al. (2016). Both texts are popular science texts on different topics, yet similar with respect to their length and structure, being similar in terms of “coherence” or more specifically, “relation” and “nuclearity” (e.g., how text spans relate to each other, and which text spans have a specific role relative to the other) according to rhetorical structure theory (Taboada & Mann 2006). The text in Spanish contained 189 words and was about the “electronic tongue,” a sensor device for artificial assessment of taste and flavor of coffee. The text in English contained 187 words and was about the behavior of crumpled sheets in which it explained how the size of the sheets changes in relation to the force they withstand. After translating each text, retrospective protocols were recorded by using the Replay function in Translog. After both texts were translated and the verbal protocols were recorded, participants filled in a questionnaire to assess their perception of the tasks (e.g., level of difficulty, satisfaction with their target texts, etc.).

Before translating the texts, participants were informed about the use of the keyboard (accents, capitalization, copy and paste commands, etc.) and were asked to type a sentence in Spanish to become familiarized with the Translog display, internet browser, and the keyboard. They were informed that they were allowed to use the internet browser whenever they needed. Mouse events, saccade index, and gaze index were calculated in each task as they were considered indicators of cognitive effort as well. Data were filtered and exported to an Excel spreadsheet. Taking into consideration that participants’ profiles would vary amongst participants, data related to their L1 and L2 skills, dominance, education, studying and living in another country where the L2 is spoken were considered.

Data were analyzed with the open source statistical programming language R (R Core Team 2019). Specifically, we ran linear mixed effects models on each on the seven dependent variables (mouse, keypress, fixation index, saccade index, gaze duration, gaze index). For these analyses, each dependent variable was first Box-Cox transformed in order to address skewness. Each analysis involved a backwards model selection process that started with an initial model that allowed the critical predictor TASK (IT vs. DT) to interact with the variables Age, Gender, and Dominance, with varying intercept adjustments per participant; in each analysis, the initial model was simplified implementing Occam’s razor via successive likelihood ratio deletion tests and checks for multicollinearity every step of the way (using variance inflation factors). In the next section, we report the results of each final regression model by providing its  $R^2$  marginal- and  $R^2$

---

<sup>1</sup><http://www.translog.dk/>



conditional-values (quantifying the amount of explained variance without and with random effects) as well as  $p$ -values for the final predictors in the model; finally, we visualize the effects of the predictors in each final model using effects plots for predicted means/regression lines with confidence intervals/bands (Fox & Weisberg 2019).

## 9 Results

In our test of the hypothesis that IT takes longer than DT, there was no significant correlation between TASK and total time: the variable TASK was eliminated last during the model selection process ( $p = 0.1229$ ), leading to a model with no significant predictors for total time. One possible explanation is that both conditions may not have differed in their required time because they do not involve differential demands for that group as both groups work into both directions on a regular basis as professional translators. Obviously, our small sample size will also have affected the results.

Table 7.1: Time, mouse events, keypress events, fixation index, saccade index, gaze duration, and gaze index in direct (DT) and indirect (IT) translation. FI: Fixation index; SI: Saccade index; GD: Gaze duration; GI: Gaze index.

	Time	Mouse	Keypress	FI	SI	GD	GI
DT							
Total	48493	3148	68912	95164	129192	6602532786	1312959
Mean	1515.41	98.38	2153.5	2973.88	4037.25	206329149.60	45274.45
SD	615.28	65.85	1422.84	2136.84	2970.99	749586689.30	31216.13
IT							
Total	55436	4421	77679	103653	146286	11014531567	1443841
Mean	1732.38	138.16	2427.47	3239.16	4571.44	344204111.50	48128.03
SD	619.51	103.96	836.87	1502.78	2224.44	804214370.90	19857.09

Table 7.1 shows time, mouse events, keypress events, fixation index, saccade index, gaze duration, and gaze index in direct (DT) and indirect (IT) translation (all values before Box-Cox transformations). For the results regarding fixation index, the final model was significant and explained a decent amount of variability (LR = 25.226,  $df = 4$ ,  $p < 0.001$ ,  $R^2m = 0.229$ ,  $R^2c = 0.662$ ). These  $R^2$ s are due to (i) a significant effect of Age (see Figure 7.1, in which the point characters d and i represent the task (direct vs. inverse)): the older the subject, the lower the fixation index; also, there was a significant interaction between Task and Gender

(see Figure 7.2: females have lower values of fixation index in DT compared to IT; for males it is the opposite).

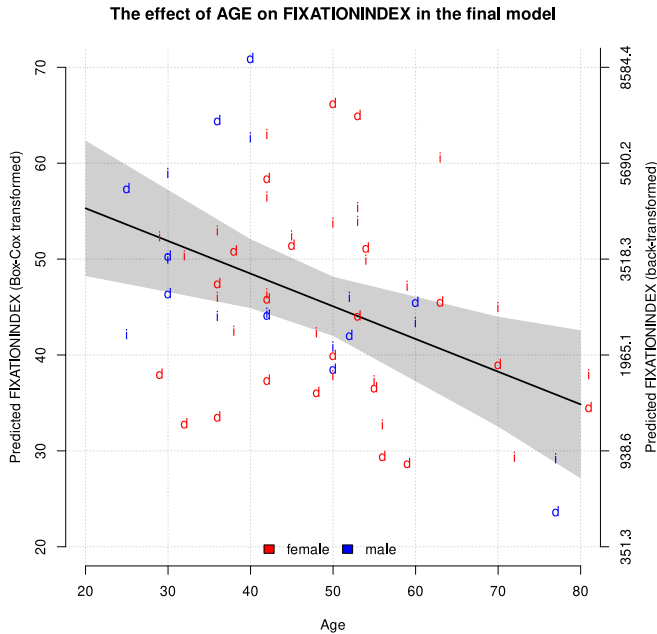


Figure 7.1: Age and number of fixations

For gaze event duration, the final model was significant (LR = 16.321,  $df = 1$ ,  $p < 0.001$ ), but the correlation of TASK was only weak and most of the variability accounted for was due to individual variation ( $R^2m = 0.053$ ,  $R^2c = 0.078$ ). Figure 7.3 shows that gaze duration was higher in IT compared to DT.

Regarding gaze point index, the final model was significant (LR = 19.431,  $df = 3$ ,  $p < 0.001$ ) but the correlation was again not strong ( $R^2m = 0.07$ ,  $R^2c = 0.207$ ). Task and Dominance are interacting significantly such that English-dominant speakers had higher values of gaze point index during IT than DT, but Spanish-dominant speakers had higher values of gaze point index in DT compared to IT.

In terms of keypresses, the final model was significant (LR = 12.355,  $df = 3$ ,  $p < 0.001$ ), with a significant correlation ( $R^2m = 0.112$ ,  $R^2c = 0.506$ ) due to the interaction between Task and Dominance, English-dominant speakers had more keypresses in the IT compared to DT, but Spanish-dominant speakers had the same keypress values regardless of the direction (see Figure 7.5).

As per mouse events, the overall model did in fact not do significantly better than a null model, but three main effects showed significant results, leading to a

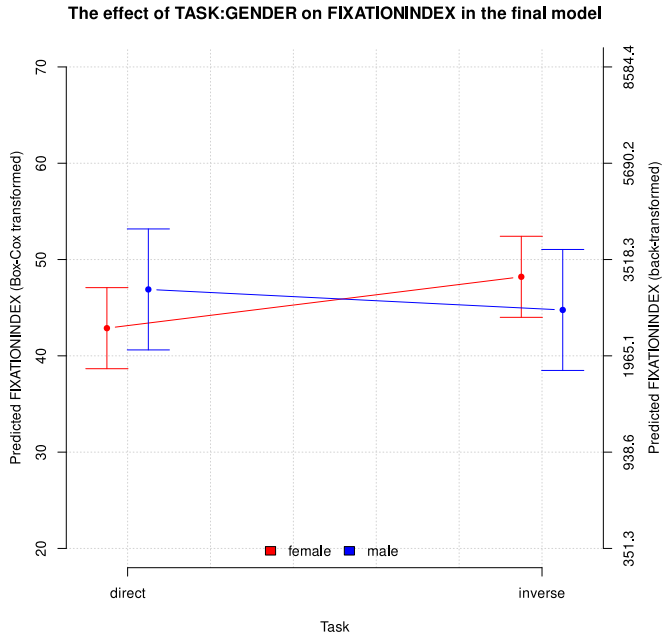


Figure 7.2: Task/gender and number of fixations

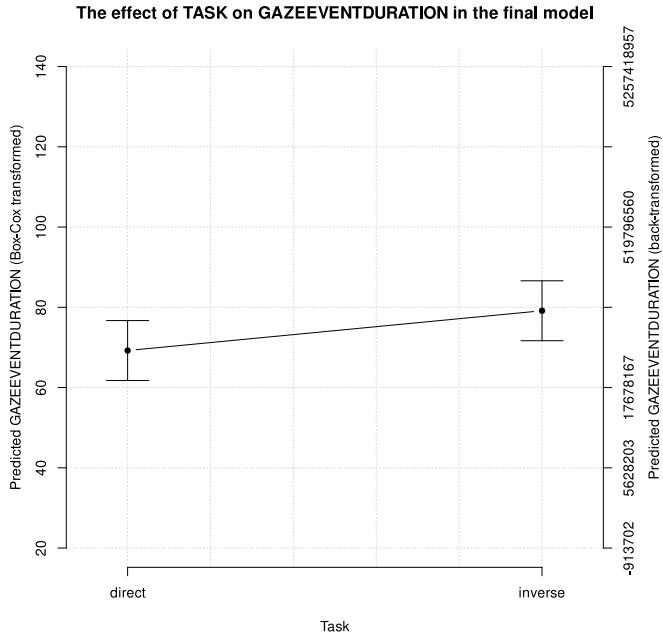


Figure 7.3: Task and gaze duration

decent correlation: ( $R^2m = 0.281, R^2c = 0.615$ ). In Figure 7.6, a significant effect for age can be seen ( $p < 0.01$ ): the older the subject, the lower the number of mouse events.

As shown in Figure 7.7, (in which the point characters d and i represent the task (direct vs. inverse) mouse events were also significantly lower for Spanish-dominant participants ( $p < 0.04$ ).

Mouse movements also were dependent on the task: there were more mouse movements during IT compared to DT ( $p < 0.02$ ) (see Figure 7.8).

As per saccade index, the final model was significant (LR = 23.229,  $df = 4, p < 0.001$ ) with a decent correlation ( $R^2m = 0.218, R^2c = 0.6$ ). The main effect for age was significant ( $p < 0.01$ ): the older the translator, the lower is the number of saccades (Figure 7.9, in which the point characters d and i represent the task (direct vs. inverse)).

There was also an interaction between Task and Sex/Gender ( $p < 0.05$ ): females had lower values of saccade index during DT than in IT. The opposite pattern was found for males, who had slightly higher values of saccade index in DT than in IT (Figure 7.10).

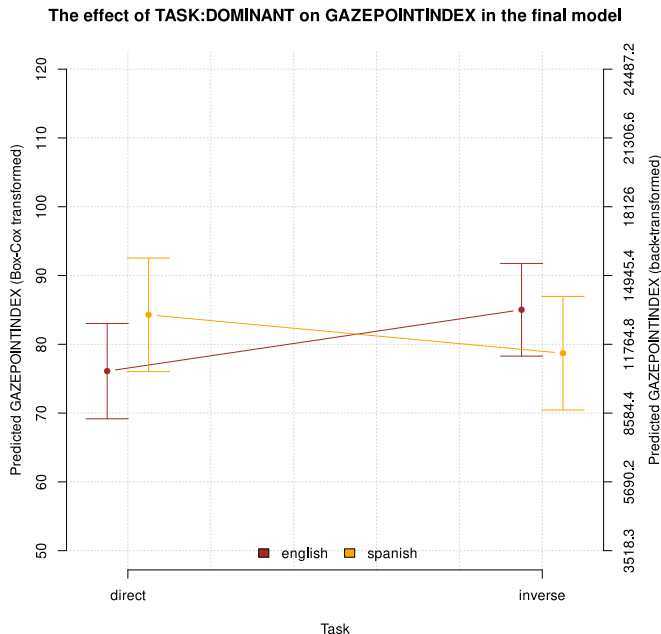


Figure 7.4: Task/dominance and gaze point index

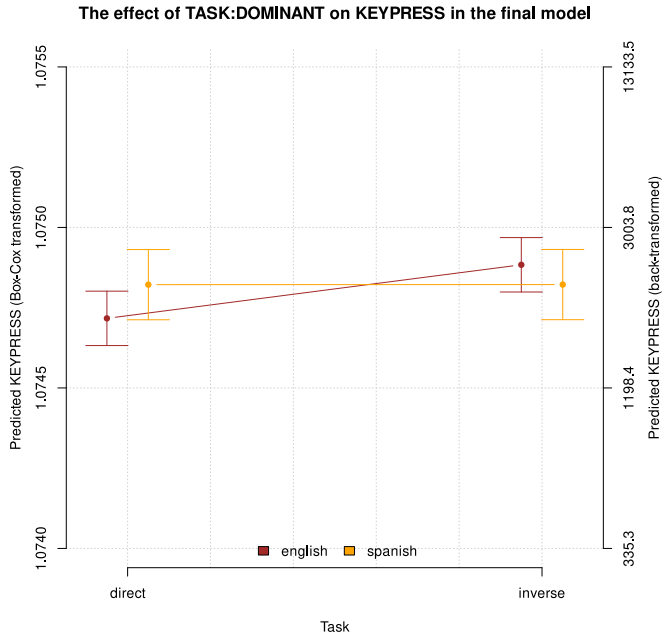


Figure 7.5: Task/dominance and keypress events

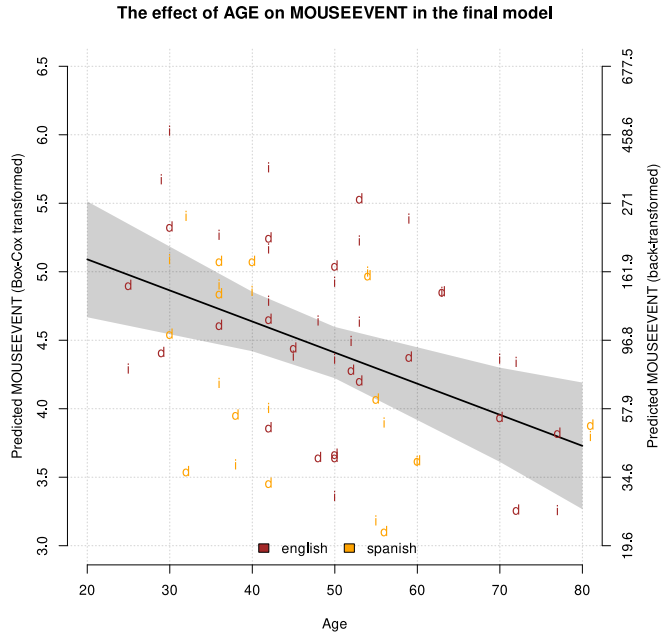


Figure 7.6: Age and mouse events

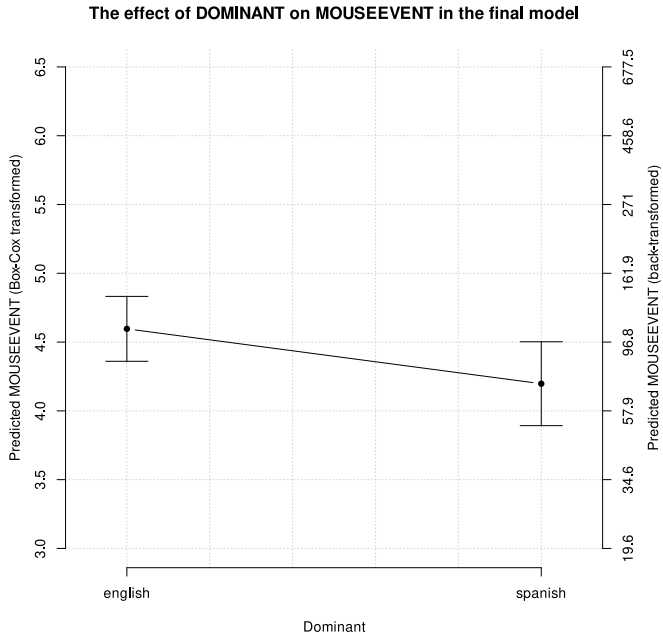


Figure 7.7: Mouse events and dominance

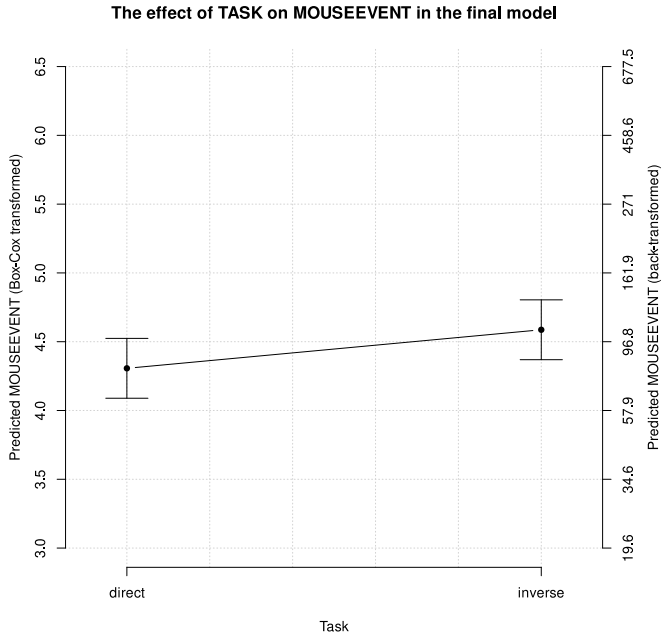


Figure 7.8: Task and mouse events

7 Assessing indicators of cognitive effort in professional translators

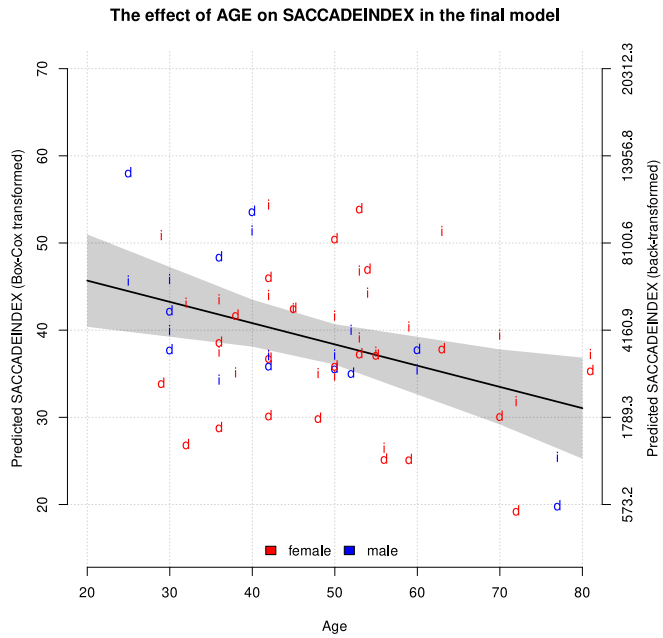


Figure 7.9: Age and saccade index

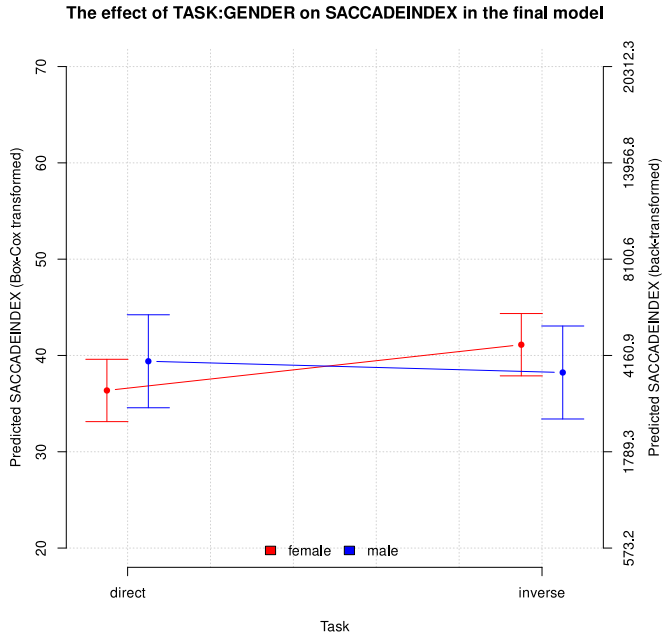


Figure 7.10: Task/gender and saccade index

## 10 Discussion

In this study, we have provided some preliminary insight on the effects of directionality. Based on Ferreira et al.'s (2016) analysis that took into account the total task length, fixation count, average fixation duration, and gaze time to investigate the effects of translation direction, we hypothesized that time is an indicator of cognitive effort and IT takes longer than DT. However, linear modeling for TASK (DT vs. IT) revealed no significant differences.

Carl et al. (2016) explained that a “translator’s actions, keypresses, gaze dwell times, and mouse events, are manifestations of the translator’s cognitive states as the translation is produced.” Assuming that IT is more demanding, translators were expected to not only spend longer time in the IT but to also present higher mouse index, fixation index, gaze event duration, gaze point index, keypresses, and saccade index. With respect to fixation index, we found a correlation but the *t*-test showed no significant differences. Therefore, the variability is probably due to translators’ differences as we stated in a footnote above. Results also showed that age is correlated to fixation index (the older the subject, the lower the fixation index). This could possibly be due to a more risky reading strategy in which they are more likely than younger adults to infer the identities of upcoming words using prior context and only partial word information (Rayner et al. 2009; Rayner et al. 2013; Wang et al. 2018).

With respect to sex/gender, our results showed that females had lower values of fixation index in IT compared to DT, whereas males presented higher values of fixation index during DT than in IT. Shen & Itti (2012) investigated whether gender-based differences also existed in visual attention during a related listening task and found that men and women orient attention differently during conversational listening. Women more often exhibited “distracted saccades,” looking at the background scene element, whereas men consistently selected regions which expressed more variation in dynamic features. Further analyses will be conducted in order to analyze how those differences in DT and IT manifest in our cohort.

As per gaze event duration, results showed that translators presented a higher gaze event duration in the IT, which suggests that gaze duration is an indicator of higher cognitive effort in the IT. On the other hand, English-dominant translators showed a higher gaze point index in IT, whereas Spanish-dominant translators presented a higher gaze point index in DT, suggesting that both groups present a higher number of gaze points when translating into Spanish. This could be due to the Spanish written system which arguably has an orthographic system that may be more demanding than that of English. Further analysis on the source text,



target text, and internet browser areas of activation could shed some light to this hypothesis.

English-dominant translators presented higher values of keypress in IT, whereas Spanish-dominant translators showed the same amount regardless of the direction of translation. Again, it is possible that translating into Spanish may be more demanding due to its orthography. Spanish-dominant translators are more familiar with its morphosyntax and translating into Spanish might not be significantly more effortful than into English. Similarly for keypress index, Spanish-dominant translators showed a lower mouse index in both tasks, whereas mouse index is higher in the IT. Results also showed that older participants tended to use the mouse less than younger translators. Smith et al.'s (1999) results showed that older participants had more difficulty performing mouse tasks in comparison to their younger peers. Differences in performance attributable to age were found in more complex tasks, and age-related changes in psychomotor abilities were related to age differences in performance. Age was also related to saccade index: the older the subject, the lower is saccade index. Peltsch et al. (2011) showed that saccadic ability decreased with age, providing insight into deficits due to normal brain changes in aging. It is likely that this is also the case in our study.

Our results also complement recent dialogues in the L2 acquisition literature. Larsen-Freeman (2018) explains that “in this era of rapid change and turmoil, there are both perils and opportunities afforded by globalization” (p. 55). This suggests that researchers in the field adopt an ecological perspective to elaborate complexity guided by the relationship between variables and individual differences. Even in social approaches to language development, cognitive aspects cannot be ignored and therefore, the socio-cognitive process would provide a contribution to language development. According to Larsen-Freeman,

Socio-culturalists see social relationships as mediating learners' cognitive development...unlike cognitive approaches, sociocognitive approaches favor patterns over rules as the object of learning, and like some of the social approaches before them, sociocognitive approaches blur the boundary between language use and its acquisition (p. 58).

Larsen-Freeman (2018) further highlights the relevance of the sociopolitical context and the nature of the limitations that shape any particular acquisition context. “Language learning does not occur in an ideological vacuum but rather is affected in a serious way by prevailing beliefs held by others, including the general public” (p. 59). IT has been misconceived to be an almost impossible task –

only being possible in cases of perfect bilingualism, without any solid empirical evidence. The fact is that people translate into the non-native language and, from what we have experienced since the first empirical studies on directionality, there is no evidence that the practice of IT will diminish moving forward simply because of beliefs or statements.

While there are certainly “increasing complexities of language use in a global society” (Kibler & Valdes 2016: 110), we should also reflect on how translation situates itself in society. In translation studies, for instance, language differences should be taken into account when designing studies and interpreting results. Defining one’s L1 and L2 is not an easy task: home country language, family language, and primary language of use, all of which is a common practice in fields such as developmental psychology and L2 acquisition, might have an impact on how we perceive translator’s performance. In a similar way, we should also attempt to analyze less examined variables that could possibly account for variances among participants.

## **11 Conclusion**

Although previous studies (Ferreira et al. 2016; 2018) suggested that language direction might have an impact on time spent in DT and IT tasks, the present study demonstrates that individual differences between subjects (including ones related to language experience), but also across subject groups such as dominance, modulate these effects. These individual differences are crucial to examine when analyzing translators’ performance and drawing conclusions, and these effects require multifactorial mixed-effects modeling of a kind that is not yet widespread in translation studies. Our sample is formed by a heterogeneous group in which professional translators’ age and experience vary, which might impact our analyses. However, eye-tracking data showed that variables other than time can be used to measure effort in translation. Not only can studying directionality help us to better understand patterns in translation, but language typology might also illuminate things. Furthermore, age and sex/gender – used here as controlled variables – should also be further analyzed to test whether any difference can be due to the factors. As this is an ongoing project, more data will be collected to create more homogeneous subgroups.

## References

- Bartłomiejczyk, Magdalena. 2004. Simultaneous interpreting A-B vs. B-A from the interpreters' standpoint. In Gyde Hansen, Kirsten Malmkjær & Daniel Gile (eds.), *Claims, changes and challenges in translation studies*, 239–249. Amsterdam: John Benjamins. <https://www.jbe-platform.com/content/books/9789027295552-btl.50.20bar>.
- Braga, Camila. 2007. *Indagando o perfil de tradutores em formação: Um estudo de caso*. Federal University of Minas Gerais. (MA thesis).
- Carl, Michael, Srinivas Bangalore & Moritz J. Schaeffer. 2016. Computational linguistics and translation studies. In Yves Gambier & Luc van Doorslaer (eds.), *Border Crossings: Translation studies and other disciplines*, 225–244. Series Title: BLT 126. Amsterdam & Philadelphia: John Benjamins. DOI: 10.1075/btl.126.11car.
- Chomsky, Noam. 1965. *Aspects of the theory of syntax* (MIT Research Laboratory of Electronics : special technical report). Cambridge: M.I.T. Press.
- Christoffels, Ingrid K., Annette M.B. De Groot & Lourens J. Waldorp. 2003. Basic skills in a complex task: A graphical model relating memory and lexical retrieval to simultaneous interpreting. *Bilingualism: Language and Cognition* 6(3). 201–211.
- De Almeida, Giselle & Sharon O'Brien. 2010. Analysing post-editing performance: Correlations with years of translation experience. In *Proceedings of the 14th annual conference of the European association for machine translation*, 26–28.
- De Groot, Annette M.B., Lucia Dannenburg & Janet G. Van Hell. 1994. Forward and backward word translation by bilinguals. *Journal of Memory and Language* 33(5). 600–629.
- Duyck, Wouter & Marc Brysbaert. 2004. Forward and backward number translation requires conceptual mediation in both balanced and unbalanced bilinguals. *Journal of Experimental Psychology: Human Perception and Performance* 30(5). 889–906.
- Ferreira, Aline. 2013. *Direcionalidade em tradução: O papel da subcompetência bilíngue em tarefas de tradução L1 e L2*. Federal University of Minas Gerais. (Doctoral dissertation).
- Ferreira, Aline, Alexandra Gottardo & John W. Schwieter. 2018. Decision-making processes in direct and inverse translation through retrospective protocols. *Translation, Cognition & Behavior* 1(1). 98–118.

- Ferreira, Aline & John W. Schwieter. 2014. Underlying processes of L1 and L3 word translation: Exploring the semantic relatedness effect. In John W. Schwieter & Aline Ferreira (eds.), *The development of translation competence: Theories and methodologies from psycholinguistics and cognitive science*, 87–106. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Ferreira, Aline & John W. Schwieter. 2017. Directionality in translation. In Aline Ferreira & John W. Schwieter (eds.), *The handbook of translation and cognition*, chap. 5, 90–105. Hoboken: John Wiley & Sons, Ltd.
- Ferreira, Aline, John W. Schwieter, Alexandra Gottardo & Jefferey Jones. 2016. Cognitive effort in direct and inverse translation performance: Insight from eye-tracking technology. *Cadernos de Tradução* 36(3). 60–80.
- Fox, John & Sanford Weisberg. 2019. *An R companion to applied regression*. Thousand Oaks, CA: Sage.
- García, Adolfo M. 2015. Psycholinguistic explorations of lexical translation equivalents. *Translation Spaces* 4(1). 9–28.
- García, Adolfo M., Agustín Ibáñez, David Huepe, Alexander L. Houck, Maëva Michon, Carlos G. Lezama, Sumeer Chadha & Álvaro Rivera-Rei. 2014. Word reading and translation in bilinguals: The impact of formal and informal translation expertise. *Frontiers in Psychology* 5. 1302.
- Göpferich, Susanne. 2009. Towards a model of translation competence and its acquisition: The longitudinal study TransComp. *Behind the mind: Methods, models and results in translation process research* 4(4). 11–37.
- Grosjean, François. 2001. The bilingual's language modes. In Janet Nicol (ed.), *One mind, two languages: Bilingual language processing*, 1–22. Malden, MA: Wiley-Blackwell.
- Guasch, Marc, Rosa Sánchez-Casas, Pilar Ferré & José E. García-Albea. 2008. Translation performance of beginning, intermediate and proficient Spanish-Catalan bilinguals: Effects of form and semantic relations. *The Mental Lexicon* 3(3). 289–308.
- Halliday, Michael & Christian M.I.M. Matthiessen. 2014. *An introduction to functional grammar*. London: Routledge.
- Jakobsen, Arnt Lykke. 2002. Translation drafting by professional translators and by translation students. *Copenhagen Studies in Language* 27. 191–204.
- Jakobsen, Arnt Lykke. 2003. Effects of think aloud on translation speed, revision and segmentation. In Fabio Alves (ed.), *Triangulating translation: Perspectives in process oriented research* (Benjamins Translation Library 45), 69–95. Amsterdam & Philadelphia: John Benjamins Publishing Company.

- Kibler, Amanda K. & Guadalupe Valdes. 2016. Conceptualizing language learners: Socioinstitutional mechanisms and their consequences. *The Modern Language Journal* 100(S1). 96–116.
- Klein, Denise, Brenda Milner, Robert J. Zatorre, Ernst Meyer & Alan C. Evans. 1995. The neural substrates underlying word generation: A bilingual functional-imaging study. In vol. 92, 2899–2903. National Acad Sciences.
- Kroll, Judith F. 2008. Juggling two languages in one mind. *Psychological Science Agenda, American Psychological Association* 22(1). <https://www.apa.org/science/about/psa/2008/01/kroll>.
- Kroll, Judith F. & Erika Stewart. 1994. Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language* 33(2). 149–174.
- La Heij, Wido, Andre Hooglander, Robert Kerling & Esther Van Der Velden. 1996. Nonverbal context effects in forward and backward word translation: Evidence for concept mediation. *Journal of Memory and Language* 35(5). 648–665.
- Larsen-Freeman, Diane. 2018. Looking ahead: Future directions in, and future research into, second language acquisition. *Foreign Language Annals* 51(1). 55–72.
- López, Belem G. & Jyotsna Vaid. 2018. Fácil or A piece of cake: Does variability in bilingual language brokering experience affect idiom comprehension? *Bilingualism: Language and cognition* 21(2). 340–354.
- Lörscher, Wolfgang. 1991a. *Translation performance, translation process, and translation strategies: A psycholinguistic investigation*. Vol. 4. Tübingen: Gunter Narr.
- Lörscher, Wolfgang. 1991b. Thinking-aloud as a method for collecting data on translation processes. In *Empirical research in translation and intercultural studies*, 67–77. Tübingen: Gunter Narr.
- Lörscher, Wolfgang. 2014. Bilingualism and translation competence: A research project and its projected results. In M. Piotrowska & S. Tyupa (eds.), *Challenges in translation pedagogy*, 1–10. inTRAlinea Special Issue.
- Malmkjær, Kirsten. 2009. What is translation competence? *Revue Française de Linguistique Appliquée* 14(1). 121–134.
- Newmark, Peter. 1988. *A textbook of translation*. Vol. 66. New York: Prentice hall New York.
- PACTE. 2005. Investigating translation competence: Conceptual and methodological issues. *Meta: Journal des Traducteurs* 50(2). 609–619.
- Paradis, Michel. 1985. On the representation of two languages in one brain. *Language Sciences* 7(1). 1–39.

- Pavlović, Nataša. 2010. What were they thinking?! Students' decision making in L1 and L2 translation processes. *Journal of Language and Communication Studies* 44. 63–87.
- Pavlović, Nataša & Kristian T. H. Jensen. 2009. Eye tracking translation directionality. In Anthony Pym & Alexander Perekrstenko (eds.), *Translation research projects* (2), chap. 2, 93–109. Tarragona: Intercultural Studies Group.
- Pavlović, Nataša. 2007. *Directionality in collaborative translation processes*. Universitat Rovira i Virgili. (Doctoral dissertation).
- Peltsch, A., A. Hemraj, A. Garcia & D.P. Munoz. 2011. Age-related trends in saccade characteristics among the elderly. *Neurobiology of Aging* 32(4). 669–679.
- Pokorn, Nike K. 2004. Challenging the myth of native speaker competence in translation theory. In Gyde Hansen, Kirsten Malmkjær & Daniel Gile (eds.), *Claims, changes and challenges in translation studies*, 113–124. Amsterdam: John Benjamins Publishing Company. DOI: 10.1075/btl.50.10pok.
- Pokorn, Nike K. 2005. *Challenging the traditional axioms: Translation into a non-mother tongue*. Vol. 62. Amsterdam: John Benjamins Publishing.
- Price, Cathy J., David W. Green & Roswitha Von Studnitz. 1999. A functional imaging study of translation and language switching. *Brain* 122(12). 2221–2235.
- Quaresima, Valentina, Marco Ferrari, Marco C. P. van der Sluijs, Jan Menssen & Willy N. J.M. Colier. 2002. Lateral frontal cortex oxygenation changes during translation and language switching revealed by non-invasive near-infrared multi-point measurements. *Brain research bulletin* 59(3). 235–243.
- R Core Team. 2019. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. <http://www.R-project.org/>.
- Rayner, Keith, Monica S. Castelhana & Jinmian Yang. 2009. Eye movements and the perceptual span in older and younger readers. *Psychology and aging* 24(3). 755.
- Rayner, Keith, Jinmian Yang, Susanne Schuett & Timothy J. Slattery. 2013. Eye movements of older and younger readers when reading unspaced text. *Experimental Psychology* 60. 354–361.
- Rothe-Neves, Rui. 2002. *Características cognitivas e desempenho em tradução: Investigação em tempo real*. Federal University of Minas Gerais. (Doctoral dissertation).
- Shen, John & Laurent Itti. 2012. Top-down influences on visual attention during listening are modulated by observer sex. *Vision Research* 65. 62–76.
- Shreve, Gregory M. 2006. The deliberate practice: Translation and expertise. *Journal of Translation Studies* 9(1). 27–42.

- Smith, Michael W., Joseph Sharit & Sara J. Czaja. 1999. Aging, motor control, and the performance of computer mouse tasks. *Human Factors* 41(3). 389–396.
- Taboada, Maite & William C. Mann. 2006. Rhetorical structure theory: Looking back and moving ahead. *Discourse Studies* 8(3). 423–459.
- Torgrimson, Britta N. & Christopher T. Minson. 2005. Sex and gender: What is the difference? *Journal of Applied Physiology* 3. 785–787. DOI: 10.1152/jappphysiol.00376.2005.
- Wang, Jingxin, Lin Li, Sha Li, Fang Xie, Min Chang, Kevin B. Paterson, Sarah J. White & Victoria A. McGowan. 2018. Adult age differences in eye movements during reading: The evidence from Chinese. *The Journals of Gerontology: Series B* 73(4). 584–593.
- Whyatt, Bogusława. 2018. Old habits die hard: Towards understanding L2 translation. *Między Oryginałem a Przekładem* 41. 89–112. DOI: 10.12797/MOaP.24.2018.41.05.
- Whyatt, Bogusława & Tomasz Kościuczuk. 2013. Translation into a non-native language: The double life of the native-speakership axiom. *mTm Translation Journal* 5. 60–79.
- Wood, Guilherme Maia de Oliveira, Maria Raquel Santos Carvalho, Rui Rothe-Neves & Vitor Geraldi Haase. 2001. Validação da bateria de avaliação da memória de trabalho (BAMT-UFGM). *Psicologia: Reflexão e Crítica* 14(2). 325–341.





# Chapter 8

## Attention distribution and monitoring during intralingual subtitling

Anke Tardel<sup>a</sup>, Silvia Hansen-Schirra<sup>a</sup>, Moritz Schaeffer<sup>a</sup>,  
Silke Gutermuth<sup>a</sup>, Volker Denkel<sup>b</sup> & Miriam Hagmann-  
Schlatterbeck<sup>b</sup>


<sup>a</sup>University of Mainz <sup>b</sup>ZDF Digital

This paper presents results from a small pilot study carried out within the EU-funded Compass project. With eye tracking, subtitlers' distribution of attention and monitoring during subtitling with the commercial subtitling software FAB Subtitler was investigated and analysed. During three intralingual subtitling tasks for excerpts of German documentaries we recorded data on gaze activity of eight subtitlers. An annotation of the created subtitles based on a comparison of the subtitles from the recording and the corrected subtitles after post-experiment proofreading allows us to link product with process data. We found that during subtitling, attention is shifted back and forth between monitoring the content of the evolving subtitles and the timing and segmentation thereof. Results show that subtitle reading times were longer on subtitles that were corrected during post-experiment proofreading. In addition, we found a significant interaction with subtitle ID in that incorrect subtitles had longer reading times than correct subtitles in the beginning of the process but for subtitles created later in the session this difference was no longer significant. This suggests that later in a subtitling session participants' monitoring capacities are impacted possibly by fatigue or cognitive overload. In this paper, we will elaborate on the methodology and procedure and suggest interpretations and possible implications.

### 1 Introduction

Subtitling as a process, particularly intralingual subtitling for the deaf and hard of hearing (SDH), has yet to be researched more in depth with empirical methods.



Anke Tardel, Silvia Hansen-Schirra, Moritz Schaeffer, Silke Gutermuth, Volker Denkel & Miriam Hagmann-Schlatterbeck. 2021. Attention distribution and monitoring during intralingual subtitling. In Tra&Co Group (ed.), *Translation, interpreting, cognition: The way out of the box*, 145–162. Berlin: Language Science Press. DOI: 10.5281/zenodo.4545043 

In times of increasing workloads for subtitlers as well as ever-changing working environments, it becomes even more important to better understand the processes involved so we can react and adapt tools and practices accordingly. What good is a subtitling tool if it speeds up subtitling but increases cognitive load on the subtitler and negatively influences final subtitle quality?

Established methods from translation process research such as eye tracking and keylogging seem promising to be applied to subtitling as well (Orrego-Carmona et al. 2018). These kinds of methods allow us to empirically study subtitlers' behaviour such as where they look, what and how they type, but they also help us better understand the process on a cognitive level. Measures such as total and average fixation duration as well as fixation count are established measures to interpret cognitive load (Buettner 2013). In cases where these measures can be linked to poor target text quality, e.g. subtitles that do not comply with given standards, they inform us about failure in cognitive control mechanisms such as monitoring (Schaeffer et al. 2019).

In this pilot eye tracking study that was carried out within the scope of the EU-funded Compass project, we recorded subtitlers during the production of intralingual SDH for three excerpts of German documentaries to gain insights on how they interact with the subtitling software as well as implications for final subtitle quality.

## 2 The intralingual subtitling process

In subtitling for TV and film we differentiate between two main kinds of subtitling: intralingual (language of film audio matches that of the subtitles) and interlingual subtitling (film language is translated into target languages; Cintas 2003). What they both have in common is the translation of dialogue in audiovisual (AV) content into written content in a one- to two-line subtitle format. Regarding the target audience, we differentiate traditional subtitles from SDH, which are typically intralingual and, in addition to the dialogue, include description of sounds and speaker identification. Though translation studies have so far mainly focused on interlingual subtitles, due to the involved language transfer, we propose that intralingual subtitling also presents a form of translation similar to translation in easy language (Hansen-Schirra & Maaß 2019). In order to understand subtitling processes, it seems promising to first look into intralingual subtitling, as both intralingual and interlingual subtitling are subject to similar time-space constraints and the audiovisual content needs to be transcribed into condensed written subtitles. To our knowledge there are no studies looking into intralingual subtitling

with eye tracking. Findings from studying intralingual subtitling and comparing them to interlingual subtitling might help in understanding how subtitling tools need to be adapted to the two forms of subtitling.

In this study, we focus on intralingual subtitling as the topic of SDH moves more and more into focus, especially after many countries across the world have passed accessibility laws (e.g. BGG<sup>1</sup> in Germany in 2002 or the EU Audiovisual Media Services Directive<sup>2</sup> in 2010) and regulations on the proportion of public AV content that needs to be made accessible to all target groups via subtitles, sign language or audio description. Through the growing availability of AV content online, we see an increasing demand for subtitles. Just like the demand for translation surpasses the availability of qualified translators, this problem is no different for subtitling. This is why the industry is constantly trying to find ways to optimise current processes by introducing assisting technology and tools with a wide range of functionalities and features. The question is to what extent these are beneficial to the subtitling process.

Intralingual subtitling is a complex and cognitively demanding task consisting of several subprocesses; both audio and visual content need to be processed and transcribed (spoken to written language). Similar to regular translation tasks, there is no one solution on how a translation or in this case subtitle has to look like. Even in intralingual subtitling there are several correct subtitle renditions of the same utterance possible. These written utterances need to follow certain standards and style guides that assure the quality and readability of subtitles for particular target groups (deaf and hard of hearing, children, language learners, etc.). In addition, subtitles need to be synced to the timing of the audio and moving images (shot changes, banners, etc.) as closely as possible (cf. *contract of illusion*, Pedersen 2017), while at the same time bearing in mind that subtitlers are limited by the maximum reading speed of the target audience, as that often does not match the speech rate. The reading speed controls the minimum and maximum display times of subtitles depending on the number of characters. Especially fast-paced dialogue makes it inevitable to condense the written renditions to fit the limited one to two lines and maximum number of characters per line and subtitle. There is usually not one correct way to subtitle: for example, whether an utterance is rendered in two separate subtitles or one two-line subtitle is up to the subtitler.

---

<sup>1</sup>Gesetz zur Gleichstellung von Menschen mit Behinderungen (Behindertengleichstellungsgesetz), <http://www.gesetze-im-internet.de/bgg/>

<sup>2</sup>Directive 2010/13/EU, <http://data.europa.eu/eli/dir/2010/13/2018-12-18>

### 3 Monitoring in subtitling software

Based on the complex nature of subtitling described in the section before, subtitling tools feature a variety of functionalities to support the subtitler. These functionalities range from a basic subtitle editor and video player to audio-wave display, visualisation of the ratio between number of characters and minimal reading time to error codes and subtitle overviews. While these features are meant to assist subtitlers in their work, they also mean that subtitlers' attention needs to shift away from the audio-visual content and the subtitles themselves. Monitoring in the subtitling task therefore goes beyond monitoring what is being typed. In addition, subtitlers need to review the subtitles they create and monitor whether the solutions they create match the expectations, i.e. style guide. For the purpose of this study, we adopt Kitchener's second level of cognitive processing. While the first level refers to the cognitive tasks involved, which in the case of subtitling include, e.g., listening, watching, and reading, Kitchener's second-level concept of monitoring as a meta-cognitive activity is defined by "processes which are invoked to monitor cognitive progress" (Kitchener 1983: 225). This model of monitoring "account[s] for complex monitoring when individuals are faced with ill-structured problems" (Kitchener 1983: 222). In the case of our study, errors in subtitles, i.e. subtitles that need to be adapted to comply with a given style guide, can be regarded as such ill-structured problems triggering monitoring processes while working on them.

Innovative subtitling software attempts to assist subtitlers in solving these ill-structured problems to minimise the cognitive load, e.g., in helping the subtitler apply subtitling strategies to comply with the style guide. Common features of these subtitling tools include a video player and subtitle editor as well as an overview of all the subtitles in a file. Usually, in and out times as well as the sequential subtitle ID and subtitle duration are displayed on screen as well. Many tools also visualise the audio in waveform to easily navigate in the video and support subtitle spotting, i.e. setting the in and out timestamps of subtitles synchronously to when speakers start or end their dialogue. Commercial tools can display an additional "time bar" that indicates the proper subtitle display time per number of characters, and error codes are displayed on-screen when a subtitle is too long or short, etc. The list of visual features a subtitler is faced with during the task of watching the video and reading the subtitle text as it is produced is long. Features are used successfully and monitoring can be assumed to have worked well when the produced subtitles no longer contain style guide-related errors. If this is not the case, the visual features are either not used properly or the distribution of attention on so many different areas leads to the result that something

is overlooked, i.e. monitoring fails. In SDH, there is a thin line between rendering and describing everything and risking to lose the target audience and censoring or patronising the target audience by condensing information or changing relevant information. So there is always the possibility that even if the style guide is being followed, the target audience might not like it. Identifying indicators for when monitoring in subtitling fails can help adapt subtitling tools and develop strategies.

Based on the many aspects subtitlers need to monitor during subtitling, we propose the following research questions:

*RQ 1.* How is attention distributed on the user interface in the subtitling tool during intralingual subtitling?

*RQ 2.* How is attention distribution and monitoring of one's own subtitling related to the final subtitle quality?

## 4 Data and method

The study applies two methodologies, keylogging and eye tracking, but in this paper only the eye tracking and product data are analysed. We applied these methods to the subtitling process, similar to the study done by Orrego-Carmona et al. (2018), who observed interlingual subtitling processes of students and professional subtitlers working with and without transcript. In our study, however, we observed the intralingual subtitling process and recorded behavioural and product data with these methods. Demographic data and information on participants' use of subtitling tools were recorded with a follow-up questionnaire.

### 4.1 Participants

The participants ( $N = 8$ , all female) in this pilot study were experienced subtitlers (mean = 3.5 years) at a German broadcasting company with German as their native language. Four of the participants are regular employees who have been working as subtitlers at the company for an average of 5 years ( $SD = 2.5$ ) and most of their work (80%) consists of proofreading subtitle files. The other four participants are experienced students with 2.3 years' ( $SD = 1.3$ ) professional experience in intralingual subtitling. All eight participants are familiar with FAB Subtitler Standard as subtitling tool and work with it on a regular basis. Though the majority of the work is for productions of German TV broadcasters with their own set of subtitling rules, all subtitlers were somewhat familiar with the Netflix

timed text style guide (Netflix 2018) as many productions, especially documentaries, are subtitled for Netflix. We chose the Netflix timed text style guide general requirements (ibid.) and the supplement for German as it is internationally used and provides a rather strict set of rules.

## 4.2 Procedure

Participants were recorded creating German SDH for three video snippets from German documentaries using the stand-alone subtitling software FAB Subtitler (Standard Edition). The initial aim was to observe subtitlers while using a subtitling tool they were using on a regular basis and to study the distribution of attention. For this purpose, we divided the tool into areas of interest (AOI) that represent a specific feature or functionality in the tool. Important AOIs include:

- the subtitle editor (current subtitle, CS)
- the video player
- audio track with subtitle in and out indicators overlaid
- reading speed control bar, i.e. time bar (characters/subtitle duration)
- subtitle navigation on the right-hand side

All AOIs were labelled the same in all recordings except for the AOI of the current subtitle (e.g. CS46) which matched the ID of the subtitle that participants currently worked on. All numbered CS AOIs were grouped as CS in order to compare this AOI to the other AOIs in the overall process. An overview of all AOIs is given in Figure 8.1.

The tool was configured for the Netflix timed text style guide and participants were able to use the browser for external research and checking the style guide. The subtitling brief included the instruction to create German SDH according to the style guide. Subtitling was done from scratch, i.e. aside from the style guide and web browser participants had no additional reference material, no assistance via a script or automatic detection of shot changes. All eight participants subtitled all three videos in a randomised order. Before the first recording, subtitlers performed a copying task to record typing speeds. Individual subtitling sessions lasted about an hour per video. The videos were controlled for their length (five minutes) and topic. They were taken from the German documentary series *TerraX* covering topics such as anthropology, archaeology, history and architecture.

## 8 Attention distribution and monitoring during intralingual subtitling



Figure 8.1: Screenshot of FAB Subtitler user interface with labelled AOIs

Participants performed the tasks during their regular working time in their regular work environment on a laptop equipped with an eye tracking device. Eye movements were recorded with an SMI RED mobile eye tracker (250 Hz) and screen recording in SMI Experiment Center. Participants were seated at approximately 60 cm from the eye tracking device where their eyes were calibrated (5-point calibration) before the beginning of each recording session and then every 15 minutes during the recording to avoid drift. After the third recording, participants filled out a questionnaire to collect demographics and data on their usage of the tool. Based on feedback from the subtitlers who participated in this study, we learned that it is common practice to distribute the subtitling work for an hour-long video or episode among three to four subtitlers, each subtitling a section of fifteen minutes. Splitting the work between subtitlers is a practice to meet the tight deadlines as the time spent on a one-hour video can be reduced and coherence is ensured in the overall proofreading of all parts by one proofreader. The experiment design of three five-minute video excerpts therefore seemed realistic enough. On average, intralingual subtitling of a five-minute documentary without any further assistance by a transcript or automatic detection of shot changes takes about an hour. This is already rather long for an eye tracking study and had to be accommodated for with repeated calibrations.

Half a year after the initial recording, the four subtitlers who usually perform the proofs were given the subtitle files for blind proofreading. The six-month

time-lag was necessary so the subtitlers would not remember having subtitled the excerpts before. The proofread subtitle files were then compared to the subtitles created during the experiment and annotated for timing errors, linguistic errors and style-guide-related errors. The annotation is based on what proofreaders corrected and the timing configurations regulated by the Netflix style guide. There was no process data recorded on the proofreading and it is just one proofreading per target text, i.e. subtitled excerpt. This adds to the ecological validity of this study as this is how quality assurance is done in the broadcasting company where we recorded the subjects.

Subtitling quality is another complex and to this point often debated concept which will not be further discussed in this paper. We therefore only looked at the broad distinction between correct and incorrect subtitles, leaving error types and weighting of errors aside for now. Whether a subtitle was correct or incorrect is based on the proofread versions. No edits meant a subtitle was correct, whereas changes made to the subtitle indicated an incorrect subtitle. There was no particular pattern regarding how errors were distributed within and across the three videos. An overview on how errors were distributed can be found in Figure 8.2.

### 4.3 Measures

During the subtitling task, both product and process data were recorded for all sessions, i.e. three sessions per participant. Regarding the process data, in this analysis, we focus on the eye tracking data. The two gaze measures of interest in this analysis are *average fixation duration* on an AOI, e.g. the current subtitle and *total fixation duration* which indicates the sum of all fixation durations on an AOI. In the case of the subtitle AOI the total fixation duration is the total reading time (TRT). Longer average fixation duration are taken to indicate higher cognitive effort, i.e. longer processing.

For the product data, the final subtitle files were analysed and annotated with the following parameters:

*ID*: the sequential number of a subtitle, used to align process data from the AOIs with the product data from the subtitle files. As the subtitling process is chronological, lower ID indicates the subtitle appeared earlier in the video and was therefore created early in the process (per recording).

*CharCount*: the number of characters in a subtitle, irrespective of the number of lines in a subtitle.



## 8 Attention distribution and monitoring during intralingual subtitling

*CharDiff*: the difference between the maximum number of characters in a subtitle and number of characters. The maximum number of characters is determined by the Netflix style guide configuration set in FAB.

*Timing*: if the *CharDiff* in a subtitle is negative by more than three characters, i.e. the subtitle contains more than three characters too many for the set display time of the subtitle, the subtitle was annotated with *Err*, i.e. the timing is incorrect. Subtitles with an excessive display time were not penalised.

*Text*: subtitles were annotated with *Err* whenever a subtitle was edited during post-experiment proofreading. In this variable, the type of edit was not distinguished, and unedited subtitles were annotated with *Cor*.

*Proof*: indicates whether a subtitle is correct (*Cor*) or contains an error (*Err*), i.e. the subtitle contains an error either in *Timing* or *Text*.

*ErrorType*: Indicates the nature of the error in *Proof*. We differentiated between errors of *Ling* (language-related, e.g. grammar, punctuation or terminology), *Style* (style guide-related, e.g. segmentation, numbers and units, labelling) and *Timing*.

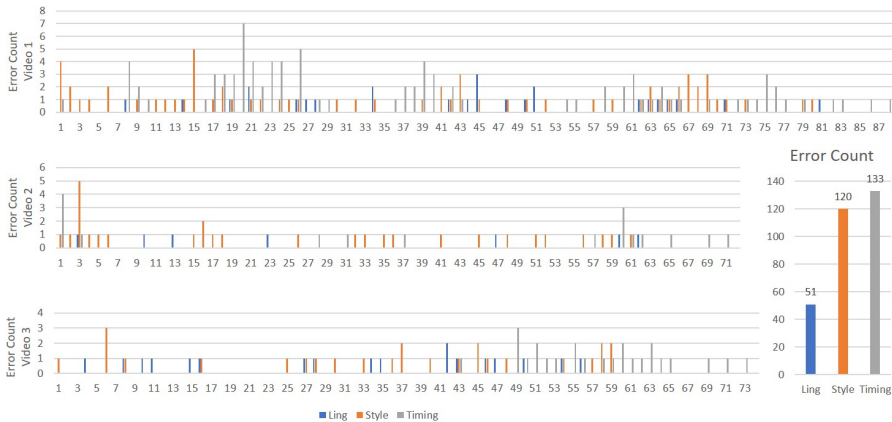


Figure 8.2: Distribution of errors per subtitle ID and video across participants. Errors are divided into the three error classifications Ling, Style and Timing.

The diagram in Figure 8.2 shows how errors were distributed per video and subtitle ID. Here, we see that there was no pattern whether errors occurred early or later in a session and there was also no pattern for the different error types. Video 1, however, seems to contain more errors in total than the other two videos.

## 4.4 Data analysis

For statistical analysis of the data we used R, version 3.6.0 (R Core Team 2019). Linear mixed effects models (LMEs) were fitted with the packages `languageR` (Baayen 2008), `lme4` (Bates et al. 2015), and `lmerTest` (Kuznetsova et al. 2017) was used to calculate significance values. To support the interpretation of the models, the effects were visualised in effects plots using the `effects` package (Fox & Hong 2009). Contrasts within a model were calculated with the package `emmeans` (Lenth et al. 2019). The LMEs were all fitted with one of the process-related measures as dependent variable and participant as random variable. Dependent variables were not log-transformed.

## 5 Results

### 5.1 Attention distribution

As described in Section 2, the task of intralingual subtitling is rather complex and involves various subprocesses, such as watching the video (images and shot changes) while listening to and understanding the audio (dialogue and sounds) and keeping an eye on the timing as well as spatial limitations of the subtitle that is created. Modern subtitling software contains a number of features to support subtitlers in this complex process, among them a video player, subtitle editor, audio track, etc. For a screenshot of the user interface of FAB Subtitler refer to Figure 8.1. We assume that, during the subtitling, the subtitlers' attention is divided between the various windows and functions onscreen but also the audio. Thus, our first research question was: how is attention distributed on the subtitling tool during intralingual subtitling?

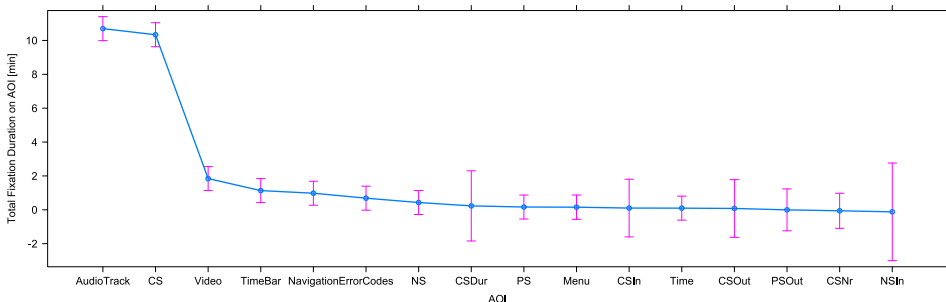


Figure 8.3: Effect of AOI on total fixation duration (in minutes)

Figure 8.3 visualises significant effects of the AOIs *audio track* and *current subtitle* (CS) on *total fixation duration* in contrast to all other AOIs in FAB Subtitled. With AOI *video* as reference, the effects for *audio track* ( $\beta = 8.9$ , SE = 0.4, df = 238,  $t = 21.4$ ,  $p < 0.0001$ ) and CS ( $\beta = 8.5$ , SE = 0.4, df = 238,  $t = 20.5$ ,  $p < 0.0001$ ) were both highly significant and positive.

In a next step, we had a look at the average fixation duration on the different AOIs. Both longer total reading time (TRT) and longer average fixation duration are indicative of increased cognitive effort while average fixation durations are in a certain sense an earlier measure than TRT. Here, we found the highest average fixation duration for the AOI *audio track*, the one AOI that did not have a significantly lower total fixation duration than AOI CS, which is the subtitling area where the current subtitle is being typed and monitored. Adjustments in the timing are done by listening to the *audio track* and at the same time fixating the AOIs *time bar* and *error codes*.

An effects plot for the second LME is shown in Figure 8.4 with the AOIs ordered for their average fixation duration. Here, we clearly see that CS lies in the centre of the plot. If we take AOI CS as reference and look at the contrasts for *average fixation duration* with the other AOIs, we find significant positive effects for *audio track* ( $\beta = 382$ , SE = 40, df = 238,  $t = 9.6$ ,  $p < 0.0001$ ) and *error codes* ( $\beta = 196$ , SE = 40, df = 238,  $t = 4.9$ ,  $p < 0.0002$ ) and the previous subtitle out time (*PSOut*;  $\beta = 330$ , SE = 64, df = 241,  $t = 5.2$ ,  $p < 0.0001$ ), i.e. average fixation durations were significantly longer for these AOIs than on CS. Only marginally significantly shorter average fixation durations were found on the current subtitle ID (*CSNr*;  $\beta = -177$ , SE = 54, df = 239,  $t = -3.3$ ,  $p < 0.092$ ) and on the next subtitle (*NS*;  $\beta = -132$ , SE = 40, df = 238,  $t = -3.3$ ,  $p < 0.079$ ). The AOI next subtitle only contains content when the subtitler has already created the preceding subtitle and went back to check the preceding subtitle. This indicates these areas might be processed faster than the current subtitle and seem to require less attention as also indicated by the first LME in Figure 8.3.

## 5.2 Monitoring and cognitive load

The second part of the analysis was concerned with the product of subtitling in relation to the process data. Subtitles underwent post-experiment proofreading and were annotated if they contained some kind of error (Linguistic, Style or Timing). In this first analysis, we did not differentiate between the nature of the error but simply whether a subtitle was correct or incorrect (*Err*).

A plot for the first LME is shown in Figure 8.5. Here, we found a significant positive effect for the total reading time *total fixation duration* for subtitles that still

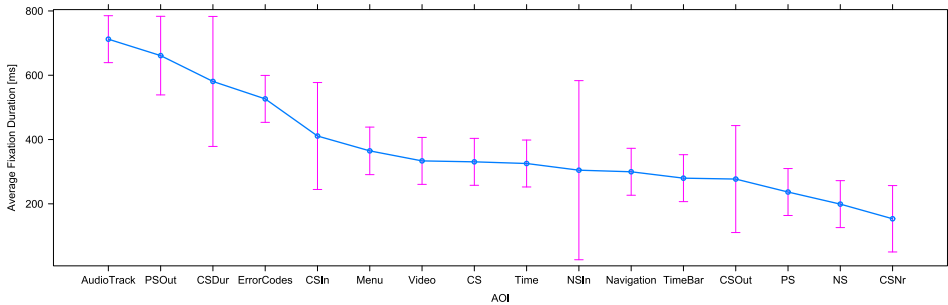


Figure 8.4: Effect of AOI on average fixation duration (in milliseconds)

contained an error at the end of the session, i.e., subtitles that were edited during proofreading. Subtitles which were corrected during proofreading (6 months after the experiment) were fixated longer during the subtitling session than subtitles which had not been corrected during proofreading ( $\beta = 0.9$ ,  $SE = 0.28$ ,  $df = 1680$ ,  $t = 3.3$ ,  $p < 0.001$ ), indicating that participants worked longer, or rather read these subtitles longer, yet failed to produce a correct subtitle. This can be seen as an indicator that subtitles that resulted in an error required more attention than those subtitles that successfully followed the linguistic rules and the subtitle style guide. *Character count* was included in the model as a control variable and had a highly significant effect ( $\beta = 0.13$ ,  $SE = 0.006$ ,  $df = 1680$ ,  $t = 20.7$ ,  $p < 2 \times 10^{-16}$ ).

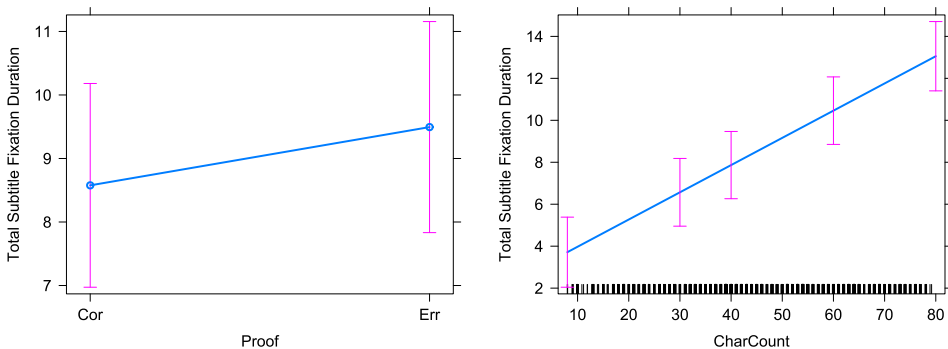


Figure 8.5: Effect of proof and character count on the total subtitle fixation duration (in seconds)

We tested whether the sequential numbering of subtitles, which roughly corresponds to the time when the subtitle was produced (ID), would have an effect on the reading time. The plot in Figure 8.6 shows the positive effect of ID (numbered

subtitles) on the reading time of the respective subtitle. This means that during subtitling, participants spent significantly more time reading subtitles they created later in the session ( $\beta = 0.014$ ,  $SE = 0.06$ ,  $df = 1664$ ,  $t = 2.6$ ,  $p < 0.01$ ). This also makes sense, as later in the session participants have concentrated and processed already quite a few subtitles and it can be assumed that cognitive load increases with time (if no break was taken).

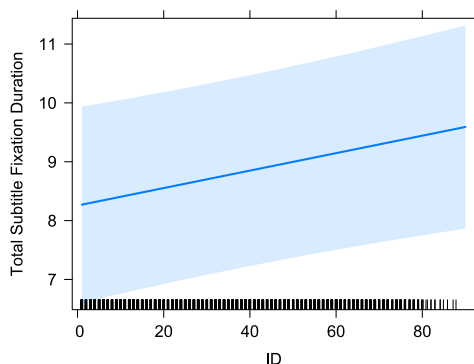


Figure 8.6: Effect of ID, i.e. whether a subtitle was created early in the session or later on Total Subtitle Fixation Duration (in seconds)

After finding these two effects, we were curious whether the ID and the post-experiment annotation for the subtitles containing an error (timing, linguistic or style guide-related error) would interact. Indeed, as can be observed in Figure 8.7, these two interact significantly ( $\beta = -0.05$ ,  $SE = 0.01$ ,  $df = 1661$ ,  $t = -4$ ,  $p < 0.001$ ) in that the reading times for subtitles later in the session did not differ significantly for correct and incorrect subtitles. Additional analyses show that for subtitles annotated as correct, the effect of ID on the reading time was only marginally significant and positive ( $\beta = 0.03$ ,  $SE = 0.01$ ,  $df = 1662$ ,  $t = 4$ ,  $p < 0.001$ ). For subtitles that contained some kind of error the effect of ID on reading time was significant and negative ( $\beta = -0.03$ ,  $SE = 0.01$ ,  $df = 1661$ ,  $t = -2$ ,  $p < 0.05$ ). While we do find a significant difference in reading times for correct and incorrect subtitles early in the session, this difference disappears around two thirds into a session.

In the next section, results will be interpreted and discussed regarding cognitive processing and monitoring during subtitling.

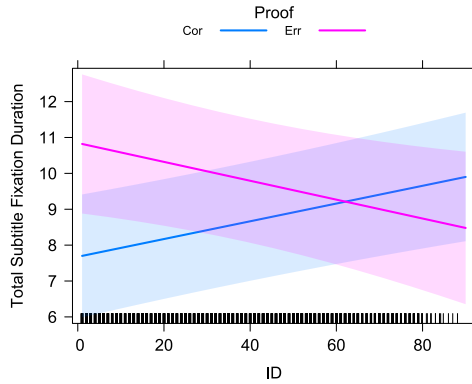


Figure 8.7: Interaction effect of subtitle ID and post-experiment subtitle correction (proof) on total subtitle fixation duration (in seconds)

## 6 Discussion

In cognitive load theory, gaze counts and gaze duration are regarded as established measures to quantify mental effort and cognitive load, which are, in turn, correlated with working memory capacities (Buettner 2013). The methodological foundation for this measurement is corroborated by the eye-mind hypothesis, which assumes that information which is fixated with the eyes is immediately cognitively processed (Just & Carpenter 1980). Based on this assumption, parameters like length, number or direction of the fixations, as well as reading time allow conclusions about cognitive load, on the one hand, but also on monitoring processes during translation, on the other (Carl & Dragsted 2012; Schaeffer et al. 2019). In the following, we will interpret our results against this background.

Figure 8.6 shows, for instance, that the total fixation duration increases the later a subtitle occurs in the whole subtitling session. This can be interpreted as an indicator of increasing cognitive load, since the later subtitles require more visual attention to be processed compared to the early ones. This result is not surprising since it may in turn be interpreted as increased cognitive load as a consequence of fatigue.

As mentioned above, for the purpose of this study, we adopted the concept of monitoring by Kitchener (1983) and defined subtitles with errors as ill-structured problems that trigger monitoring processes while reading them. Keeping this definition in mind, Figure 8.6 can be interpreted as an indicator for ongoing monitoring processes that add up during the session. In this special case, the subtitlers' total fixation duration is positively affected by the ID, i.e. later in the session the

total fixation durations were longer, when the subtitle they produced still contained an error (see Figure 8.5). However, this effect is not triggered by correction processes as the respective subtitles have not been self-corrected by the subtitlers or, if they have, still resulted in an error. More interestingly, the effect seems to be triggered while creating the particular subtitle. Therefore, we interpret this effect as increased cognitive load due to monitoring.

Figure 8.7 shows another interesting phenomenon: the monitoring effect just described seems to be weakened throughout the reading during subtitling. As soon as about two thirds of the subtitles are processed, the total fixation duration for correct subtitles is even longer. We interpret this result as follows: monitoring of subtitles that did not follow the style guide, hence contained some kind of error, can only take place in the first half of the subtitling session when necessary cognitive resources are still available. If these resources suffer from fatigue, the monitoring processes for incorrect subtitles do not differ from those of correct ones, which are in general characterised by shorter total fixation duration. This could indicate cognitive overload since incorrect subtitles do not attract as much visual attention anymore.

To sum up, the effects discussed here make monitoring visible although no successful revision processes have taken place. This enables us not only to visualise but also to quantify possibly unconscious monitoring processes as well as the break-even point for cognitive overload.

## 7 Conclusion and limitations

As discussed in the section above, with this small-scale pilot study we were able to obtain a first idea of when and where monitoring processes in intralingual subtitling might take place and we presented a methodology of how these processes can be studied when linked to final target text quality. In this analysis, we looked at two measures: total and average fixation duration per subtitle or AOI. We find that further into the session, monitoring becomes less efficient and participants generally read subtitles longer irrespective of the number of characters they contain. While earlier in the session subtitlers take longer for subtitles that result in an error, this difference is no longer significant towards the end of the session. This suggests that, due to increased cognitive load, subtitlers' monitoring processes become less or not successful at all.

Still, this study was able to shed light on monitoring processes during subtitling. Ipsen & Dam (2016) also report on errors detected but not corrected. They, in contrast, rely on video recordings and interviews, which reveal conscious and

explicitly visible processes. In our study, unconscious language control mechanisms become for the first time visible and measurable. By combining eye tracking with further data from keylogging and retrospective interviews with replay we might be able to cover both conscious as well as unconscious processes and deliver explanations for unsuccessful revision processes in future studies.

Based on these findings, the aim of a subtitling tool should be to improve monitoring for subtitlers or make these language control mechanisms more efficient. In this respect, the usability of the tool has a direct impact on the cognitive ergonomic conditions during professional translation (Ehrensberger-Dow & Massey 2014). Subtitling software, just like computer-aided tools for professional translation, are developed to assist the tasks in terms of an ergonomic workflow. However, the usability of these tools has so far not been empirically tested. The features included in the tools seem necessary and helpful but right now our results suggest that this assisting technology has room for improvement, e.g. error messages regarding incorrect timing are not easily detected – especially later in the process – and errors regarding segmentation or linguistic problems are overlooked. The methodology and study design presented in this paper could be useful in comparing subtitling software that claims to be more ergonomic (e.g. the Compass tool, Hansen-Schirra et al. 2019).

The results presented here are somewhat limited due to the small sample size as well as the nature of the experiment as participants subtitled only short excerpts from the documentaries. Subtitling sessions usually take longer than an hour, but already finding significant results in these shorter sessions suggests that these effects might hold true also for longer sessions. If we consider subtitlers' authentic work spaces and the complex workflows of replaying and stopping the videos to type and spot the subtitles according to the complex style guides, it makes sense to further investigate these processes. The methodology applied in the experiment was conducted with rather high ecological validity such that the results hold true beyond the lab environment. We hope the successful application of this methodology and the resulting insights encourage further research in this direction with other subtitling tools, languages, style guides or practices.

## Acknowledgements

The Compass project received funding from the European Union within the call *Crowdsourcing subtitling to increase the circulation of European works* (CNECT 2017/3135124) from 2018–2019.



## References

- Baayen, R. Harald. 2008. *Package languageR: Analyzing linguistic data: A practical introduction to statistics*. Version 1.5.0. Cambridge University Press. <http://www.cran.r-project.org/web/packages/languageR/>.
- Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. DOI: 10.18637/jss.v067.i01.
- Buettner, Ricardo. 2013. Cognitive workload of humans using artificial intelligence systems: Towards objective measurement applying eye-tracking technology. In *Lecture notes in artificial intelligence (LNAI)*, 37–48. Springer.
- Carl, Michael & Barbara Dragsted. 2012. Inside the monitor model: Processes of default and challenged translation production. *Translation: Corpora, computation, cognition* 2. 127–145.
- Cintas, Jorge Díaz. 2003. Audiovisual translation in the third millennium. In Gunilla M. Anderman & Margaret Rogers (eds.), *Translation today: Trends and perspectives*, 192–204. Clevedon ; Buffalo, N.Y: Multilingual Matters.
- Ehrensberger-Dow, Maureen & Gary Massey. 2014. Cognitive ergonomic issues in professional translation. In John W. Schwieter & Aline Ferreira (eds.), *The development of translation competence: Theories and methodologies from psycholinguistics and cognitive science*, 58–86. Cambridge: Cambridge Scholars Publishing.
- Fox, John & Jangman Hong. 2009. Effect displays in R for multinomial and proportional-odds logit models: Extensions to the effects package. *Journal of Statistical Software* 32(1). 1–24. <http://www.jstatsoft.org/v32/i01/>.
- Hansen-Schirra, Silvia & Christiane Maaß. 2019. *Translation proper: Kommunikationsbarrieren überwinden*. DOI: 10.25528/015.
- Hansen-Schirra, Silvia, Anke Tardel, Silke Gutermuth, Moritz Schaeffer, Volker Denkel & Miriam Hagmann-Schlatterbeck. 2019. Computer-aided subtitling: Split attention and cognitive effort. In *9th AIETI international conference (AIETI9)*.
- Ipsen, Helene & Helle V. Dam. 2016. Translation revision: Correlating revision procedure and error detection. *Hermes* 55. 143–156.
- Just, Marcel A. & Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological review* 87(4). 329.
- Kitchener, Karen Strohm. 1983. Cognition, metacognition, and epistemic cognition. *Human development* 26(4). 222–232.

- Kuznetsova, Alexandra, Per B. Brockhoff & Rune H.B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82(13). 1–26. DOI: 10.18637/jss.v082.i13.
- Lenth, Russell, Henrik Singmann, Jonathon Love, Paul Buerkner & Maxime Hervé. 2019. *Emmeans: Estimated marginal means, aka least-squares means*. <https://CRAN.R-project.org/package=emmeans>.
- Netflix. 2018. *Timed text style guide: General requirements*. <https://partnerhelp.netflixstudios.com/hc/en-us/articles/215758617-Timed-Text-Style-Guide-General-Requirements> (11 March, 2020).
- Orrego-Carmona, David, Łukasz Dutka & Agnieszka Szarkowska. 2018. Using translation process research to explore the creation of subtitles: An eye-tracking study comparing professional and trainee subtitlers. *The Journal of Specialised Translation* 30. 150–180.
- Pedersen, Jan. 2017. The FAR model: Assessing quality in interlingual subtitling. *Journal of Specialised Translation* 28. 210–229.
- R Core Team. 2019. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. <http://www.R-project.org/>.
- Schaeffer, Moritz, Sandra Halverson & Silvia Hansen-Schirra. 2019. “Monitoring” in translation. The role of visual feedback. *Translation, cognition and behavior* 2(1). 1–33.

# Chapter 9

## Eye tracking study of reading for translation and English-Russian sight translation

Elena Kokanova, Maya Lyutyanskaya & Anna Cherkasova  
Northern (Arctic) Federal University

This paper presents the results of an eye tracking study which compares reading for translation and English-Russian sight translation. The participants of this study included both students and professional interpreters who were asked to read and sight translate two texts from their B language (English) into their A language (Russian). The study revealed significant differences in oculomotor activity during reading and sight translating within a group of students and within a group of professionals. This can be explained by the difference in the efficiency of reading for translation, translation strategy and general translation skills.

### 1 Introduction

The study of oculomotor activity during the reading process has been abundant and focused on various aspects (eye movement characteristics, eye movement control, perceptual span, etc.). Eye tracking studies have mostly focused on readability and processing effort for the given text type and thus on empirical research in neurophysiology (Jakobsen & Jensen 2008; Schnitzer & Kowler 2006; Clifton et al. 2016). Eye tracking has proved to be a powerful tool in scientific research and has recently been used in applied linguistics and translation studies (Hansen-Schirra & Grucza 2016). It allows identifying the objects of attention with high spatial accuracy and temporal precision. Participants try to fixate their gaze on highly informative elements but each person can choose a different strategy for investigating a stimulus and can change it when presented the same



stimulus for the second time. This explains numerous findings in fields such as translation memory, reading for translation, distribution of cognitive effort during translation, etc. (Hvelplund 2014). Sight translation is a form of transposing a written text in the source language into an oral text in the target language. The concept of sight translation is understood differently by researchers. One of the disputed issues concerns the status of this form of translation, whether it is considered as a separate form of interpreting or as a training exercise for other forms of interpreting. Most of the current research supports the idea that the key characteristic features of sight translation include the following:

- time pressure (caused by limited time for text comprehension, minimum time for finding the translation decisions, high speed of speaking);
- strict self-control (as self-corrections are not allowed) (Chmiel & Mazur 2013; Kokanova 2016; Thawabteh 2015).

In cognitive terms, sight translation is a complex set of brain operations including processing visual input in one language, creating the oral message in another language and control of the translation process at the same time. The actual application of sight translation takes place in a number of professional settings and, despite this fact, seems to be rarely taught as a separate form of interpreting.

## **2 Research design**

The objective of the present research is to collect and compare statistical data on oculomotor activity during reading for translation and English-Russian sight translation by a group of students and a group of professional interpreters. The hypothesis of the study is that within each group of participants there will be differences between experimental tasks of reading for translation and sight translation, which will allow us to see if professional interpreters demonstrate some kind of translation strategy affecting the result.

### **2.1 Participants**

The study was conducted at Northern (Arctic) Federal University, Arkhangelsk, Russia. The first group of participants included eighteen bachelor and master students (average age: 21) with one year of sight translation training. The group of participants included students with B2/C1 level of the English language. Command of English was tested before the experiment (<https://cambridgeenglish.org>).

The second group consisted of ten participants (average age: 35). All of the participants are professional interpreters working in various fields in Arkhangelsk; the average work experience is 12 years. All participants denied having suffered any brain injuries, neurological conditions, or eyesight pathology and took part in the study voluntarily.

## 2.2 Procedure and equipment

Gaze behaviour of the participants was recorded on the basis of saccades and fixations in the infrared radiation spectrum. For the recording of eye tracking data, the system iView XTM RED (SMI, Germany) for non-contact measurement was used. The collected data were analyzed using BeGaze software. The frequency of the system was 500 Hz; the viewing distance was 55–60 cm from the screen. The experiment was conducted in accordance with ethical standards represented in the Declaration of Helsinki (DoH) and European Community directives (8/609 EC).

The participants were asked to read for two minutes and sight translate two texts from their B language (English) into their A language (Russian). Time for translation was not limited. The participants' translations were recorded for further linguistic analysis and the participants were informed about this.

The texts included abbreviations, position titles, references to historic and cultural events and phenomena such as direct speech, epithets, and metaphors. The dependent variables included measures assumed to indicate cognitive load of lexical units, such as fixation count and saccade count (Kokanova et al. 2018).

## 2.3 Coh-Metrix analysis of the source texts

Both texts were analyzed by the computer tool Coh-Metrix (Graesser et al. 2004) using a number of parameters. The first parameter concerned the overall readability of the texts, i.e. their difficulty level. The output of the Flesch reading ease formula is a number from 0 to 100, with a higher score indicating easier reading. Text 1 was assessed as fairly difficult to read in accord with the Flesch reading ease formula and was given a score of 51.597. The score for Text 2 was 72.022.

Syntactically, Text 1 was simpler than Text 2 (syntactic simplicity 73.57% and 53.98%, respectively). This component reflects how low the number of words are and how simple the syntactic structure is, which is less challenging to process. At the opposite end of the continuum are texts that contain sentences with more words and use complex syntactic structures.

Text 1 contained mostly factual information presented by such language units as abbreviations, position titles, references to historic and cultural events and phenomena. This is confirmed by the concreteness level (92.36%). Texts containing content words that are concrete, meaningful, and evoke mental images are generally considered to be easier to process and understand. Text 2 was more descriptive and contained such elements as metaphors, epithets, abstract words, so that the concreteness level was lower (53.98%). Abstract words represent concepts that are considered difficult to visualize. Texts that contain more abstract words are more challenging to understand.

Text 1 was characterised as having a higher connectivity level (71.23%), the component which reflects the degree to which the text contains explicit adversative, additive, and comparative connectives to express relations in the text. This component reflects the number of logical relations in the text that are explicitly conveyed. This score is likely to be related to the reader's deeper understanding of the relations in the text. The connectivity level of Text 2 was very low (14.23%).

### **3 Data analysis and results**

The statistical analysis of the parameters under research was carried out using SPSS version 22.0. Data processing included a comprehensive analysis of the normal distribution, and since a number of parameters did not match the Gaussian distribution, the Mann-Whitney  $U$  test was used to compare the samples. To describe the data, the median ( $Me$ ) and the first and third quartiles ( $Q1$ ;  $Q3$ ) were taken. Differences were considered statistically significant when the probability of erroneous acceptance of the null hypothesis of the absence of differences between samples was  $p < 0.05$ .

We assessed the eye movement parameters for each group of participants separately. Mostly we wanted to see if there were any significant difference between reading for translation and sight translation.

The results for students are presented in Table 9.1. The data revealed no real difference in fixation and saccade count between Text 1 reading and translation. Total fixation duration was lower during the translation process. Average saccade velocity and average saccade amplitude increased while translating whereas frequency of fixations decreased.

It was observed for Text 2 that fixation count and saccade count were substantially down during the translation task, compared to Text 1. Total fixation duration and fixation frequency are also on the decline. Average saccade velocity and average saccade amplitude did not show significant changes.

9 Eye tracking study of reading for translation and ENG/RUS sight translation

Table 9.1: Eye tracking measurements for reading and sight translation in the group of students

Metric	Reading Text 1			Translation Text 1			<i>p</i>
	Me	Q1	Q3	Me	Q1	Q3	
Fixation count	395.0	380.3	419.3	360.0	262.0	506.8	0.141
Saccade count	399.0	351.5	407.5	383.0	271.5	550.5	0.776
Total fixation duration, sec.	96.9	93.3	98.6	80.1	66.2	103.0	0.016
Av. saccade velocity, degree/sec	82.4	72.4	89.5	96.9	94.1	114.8	0.002
Fixation frequency, fix/sec	3.3	3.17	3.5	2.6	2.3	3.3	0.008
Av. saccade amplitude, degree	3.5	3.25	4.0	4.2	3.25	4.0	0.005
Metric	Reading Text 2			Translation Text 2			<i>p</i>
	Me	Q1	Q3	Me	Q1	Q3	
Fixation count	400.5	361.0	416.8	309.0	200.3	391.5	0.014
Saccade count	381.0	338.0	428.5	289.0	221.0	413.0	0.018
Total fixation duration, sec	94.6	89.5	101.9	73.9	49.8	94.8	0.011
Av. saccade velocity, degree/sec	88.6	71.6	98.1	97.5	78.8	110.5	0.064
Fixation frequency, fix/sec	3.4	3.0	3.5	2.9	2.6	3.3	0.002
Av. saccade amplitude, degree	3.8	3.3	4	4.4	3.5	5.0	0.247

The statistical analysis of eye tracking parameters in the group of professionals showed some differences between the experimental tasks (Table 9.2).

Fixation count and saccade count for sight translation task were lower than for reading task both in Text 1 and Text 2. However, the saccade count the difference between reading and translation tasks was bigger for Text 2. Total fixation duration went down during translation task compared to reading for translation.

The total fixation duration during reading for translation in the group of students were about 80% and in the group of professionals were about 70%. This parameter goes down in the translation task for both groups, although in the group of professionals this difference is bigger.

It should be noted that eye tracking data revealed meaningful differences in fixation and saccade count between reading and sight translation only in Text 2 in the group of students. From the noticeable decrease of fixation count (from 400.5 to 309.0) and saccade count (from 381.0 to 289.0) in Text 2 it can be assumed that there are some factors making translation of the second text easier for students. This may have been an effect of the warming-up period. Also, after finishing Text 1 the participants were more adapted to the stressful situation and, as both texts have the same subject matter, the general context could become a supporting factor.

As for the group of professional interpreters, fixation count and saccade count decreased in the translation tasks for both texts. There were certain stability in oculomotor behaviour of professional interpreters when performing experimental tasks. As fixations are the period of time when the eyes remain fairly still and new information is acquired from the visual array, and saccades search for new meaningful areas of fixation (Rayner 2009), supposedly, this shows that professional interpreters demonstrated some strategy in analyzing the context and searching for translation equivalents while reading the text.

This leads to the assumption that professional interpreters do the quicker search for key support words in the source text during sight translation. There is a clear-cut difference between the translation time of Text 1 and Text 2. In the group of students the average translation time for Text 1 was 2 min 12 sec and for Text 2 it was 1 min 47 sec. In the group of professionals the average translation time for Text 1 was 1 min 39 sec, and for Text 2 it was 1 min 16 sec.

In the student group the frequency of fixations during translation was lower than during the reading task. Translation of Text 1 shows an increase in the average saccade amplitude and velocity. Translation of Text 2 indicates a decrease in the fixation and saccade count. Supposedly, this shows the quicker search for key support words in the source text in a stressful situation like sight translation.



9 Eye tracking study of reading for translation and ENG/RUS sight translation

Table 9.2: Eye tracking measurements for reading and sight translation in the group of professionals

Metric	Reading Text 1			Translation Text 1			<i>p</i>
	Me	Q1	Q3	Me	Q1	Q3	
Fixation count	397.0	354.5	423.5	267.5	217.8	326.6	0.010
Saccade count	383.0	352.6	414.3	322.0	203.0	364.3	0.034
Total fixation duration, sec	93.8	85.0	97.7	61.3	41.7	80.1	0.010
Av. saccade velocity, degree/sec	82.7	68.8	97.8	100.0	81.6	102.4	0.174
Fixation frequency, count/sec	3.5	3.3	3.8	2.9	2.6	3.1	0.053
Av. saccade amplitude, degree	4.4	4.1	5.3	4.5	4.2	5.1	0.306
Metric	Reading Text 2			Translation Text 2			<i>p</i>
	Me	Q1	Q3	Me	Q1	Q3	
Fixation count	374.0	355.0	401.6	249.0	154.0	306.6	0.001
Saccade count	387.5	361.0	439.5	232.5	186.0	321.0	0.001
Total fixation duration, sec	83.8	78.7	96.4	57.8	32.7	74.3	0.005
Av. saccade velocity, degree/sec	96.4	86.4	98.3	114.2	103.7	128.2	0.131
Fixation frequency, count/sec	3.15	3.0	3.4	2.7	2.3	3.3	0.160
Av. saccade amplitude, degree	4.9	3.9	5.9	5.2	4.3	6.7	0.570

## 4 Conclusion and further research

The eye tracking data seem to support the hypothesis of the present study as professional participants did demonstrate significantly lower fixation and saccade count between reading and translation tasks. The meaningful difference in fixation and saccade counts between reading and translation tasks in the students' group was observed only in Text 2. The research has shown that English-Russian sight translation can cause difficulties for students because of the low level of silent reading skills. Oculomotor behaviour of professional interpreters is more stable. They seem to reduce their search activity in the form of fixation and saccade count during the sight translation task.

Prospects for further research can include a longitude eye tracking study of reading, reading for translation and English-Russian sight translation from beginners to semi-professionals based on a more thorough selection of texts using special computer tools for text parameters analysis and introduction of reading for translation training into the interpreting course. The results of further research can be used to work out recommendation for students on how to use the reading time more efficiently, how not to miss key elements in the text, how to overcome garden-path sentences and so on.

An interdisciplinary approach in translation studies can shed more light on translation as a decision making process and provide teachers with more tools for improving students' professional skills.

## References

- Chmiel, Agnieszka & Iwona Mazur. 2013. Eye tracking sight translation performed by trainee interpreters. In Catherine Way, Sonia Vandepitte, Reine Meylaerts & Magdalena Bartłomiejczyk (eds.), *Tracks and treks in translation studies: Selected papers from the EST congress, Leuven 2010*, 189–205. Amsterdam: John Benjamins Publishing Company. DOI: 10.1075/btl.108.10chm.
- Clifton, Charles, Fernanda Ferreira, John M. Henderson, Albrecht W. Inhoff, Simon P. Liversedge, Erik D. Reichle & Elizabeth R. Schotter. 2016. Eye movements in reading and information processing: Keith Rayner's 40 year legacy. *Journal of Memory and Language* 86. 1–19. DOI: 10.1016/j.jml.2015.07.004. (2020-09-01).
- Graesser, Arthur C., Danielle S. McNamara, Max M. Louwerse & Zhiqiang Cai. 2004. Coh-Matrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers* 36(2). 193–202. DOI: 10.3758/BF03195564. (2020-09-01).

9 *Eye tracking study of reading for translation and ENG/RUS sight translation*

- Hansen-Schirra, Silvia & Sambor Grucza. 2016. Eyetracking and Applied Linguistics. In Silvia Hansen-Schirra & Sambor Grucza (eds.), *Eyetracking and applied linguistics (translation and multilingual natural language processing 2)*, chap. Eyetracking and Applied Linguistics, 1–3. Berlin: Langsci Press. DOI: 10.17169/LANGSCI.B108.232.
- Hvelplund, Kristian Tangsgaard. 2014. Eye tracking and the translation process: Reflections on the analysis and interpretation of eye-tracking data. *MonTi: Monografías de Traducción e Interpretación* 1. 201–223. DOI: 10.6035/MonTI.2014.ne1.6.
- Jakobsen, Arnt Lykke & Kristian T. H. Jensen. 2008. Eye movement behaviour across four different types of reading task. *Copenhagen Studies in Language* 36. 103–124.
- Kokanova, Elena. 2016. *Practicum po perevodu s lista [Site translation study guide (Russian-English)]*. Arkhangelsk: Northern (Arctic) Federal University.
- Kokanova, Elena, Maya Lyutyanskaya & Anna Cherkasova. 2018. Eye tracking study of reading and sight translation. *SHS Web of Conferences* 50. 1–5. DOI: 10.1051/shsconf/20185001080.
- Rayner, Keith. 2009. Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology* 62(8). 1457–506. DOI: 10.1080/17470210902816461. (2010-09-17).
- Schnitzer, Brian S. & Eileen Kowler. 2006. Eye movements during multiple readings of the same text. *Vision Research* 46. 1611–1632.
- Thawabteh, Mohammad Ahmad. 2015. Difficulties of sight translation: Training translators to sight translate. *Current Trends in Translation Teaching and Learning E 2*. 171–195.



# Chapter 10

## Emotion and the social embeddedness of translation in the workplace

Hanna Risku<sup>a</sup> & Barbara Meinx<sup>b</sup>

<sup>a</sup>University of Vienna <sup>b</sup>Oesterreichische Nationalbank

This paper examines the emotional aspects of the translation process while taking into account the social embeddedness of translators in their working environments. It explores how it feels to be a translator, looks at what makes translators thrive or despair and examines how they cope emotionally with their work circumstances. The empirical setting for this qualitative workplace study is the translation department of an Austrian public sector institution, where authentic work situations were investigated using ethnographic research methods, i.e., participatory observation and semi-structured expert interviews. The results indicate that the translators in this department experience a wide variety of emotions ranging from satisfaction, pride, relief and enjoyment to stress, disappointment and frustration. These are inextricably linked with the networks, actors and environments involved in the translation processes and occasionally lead to the use of coping strategies.

### 1 Introduction

Translation process research has increasingly begun to emphasise the role of affective and attitudinal factors in translation (Laukkanen 1996). Psychological aspects like translator personality (Hubscher-Davidson 2009) and ergonomic aspects like translator well-being (Ehrensberger-Dow & Jääskeläinen 2018) are attracting increasing attention as explanatory factors of translation activities and processes. At the same time, the role of emotions in translation has become a legitimate object in translation process research.

This paper focuses on how it feels to be a translator, what makes translators thrive or despair and how they deal emotionally with their work circumstances.



Accordingly, it examines both the emotional aspects of the translation process and the social embeddedness of translators in their real working environments, concentrating thereby on three research questions: (1) Which emotions do translators experience during a translation process? (2) Which factors trigger emotions in translators and how do these relate to the social setting in which they work? (3) Do translators apply strategies to deal with emotions?

Before describing and discussing our research results, we will first provide an outline of prior empirical research on emotional aspects of translation as well as a brief discussion of the topic of emotions from the psychological, cognitive and situated perspectives.

## **2 Emotions as an object of translation process research**

Emotion variables have already been mentioned in various studies of the translation process (e.g., Kußmaul 1991; Tirkkonen-Condit & Laukkanen 1996; Jääskeläinen 1996; Davou 2007; Rojo & Ramos 2014; 2016; Hubscher-Davidson 2009; 2013; Lehr 2014). They have been described as an area of relevance to translation process research in which a number of aspects have yet to be explored.

Kußmaul (1991) provides a detailed description of the influence of positive emotions on the translation process. In a 1991 study examining the processes involved in the development of creative solutions to translation problems, he asked two translators to work as a team to translate a text from English into German and to explain and discuss their methods with each other while they worked. The resulting dialogue protocols led Kußmaul to conclude that the creative solutions which emerge during the translation process are linked to situations in which the translators experience positive emotions. This finding is reiterated in a 2007 study by Davou.

Rojo & Ramos (2014) show that negative emotions can have adverse effects on the translation process. They used a reaction time experiment to examine the influence of words and expressions which contradict the translator's own ideological stance. The results corroborate their assumption that words and expressions that are incompatible with a translator's ideology have an adverse effect on the decision process and lead to more time being required to find an adequate translation solution. By contrast, words and expressions that are compatible with the translator's ideology facilitate the decision process, allowing a suitable translation to be found more quickly.

Lehr's (2014) results indicate that the emotions triggered in translators through feedback on completed translations influence subsequent translation processes

in different ways depending on their valence (positive or negative). Feedback that prompts positive emotions enhances idiomatic expression and stylistic appropriateness in subsequent translations, while feedback that prompts negative emotions enhances coherence and correctness of terminology.

Rojo & Ramos repeated Lehr's study in 2016 and reached similar conclusions. In addition to replicating her methodology, they used Block & Kremen's ego-resiliency scale to explore the influence of personality traits on the translation process. Their results show that positive and negative emotions triggered by feedback on performance lead to different processing styles, increasing either creativity or accuracy in translation. They also suggest that personality traits play a role – albeit not a statistically significant one – in guiding translational behaviour.

### 3 Emotions from the psychological, cognitive and situated perspectives

From a psychological perspective, emotional and affective phenomena are positioned on a spectrum ranging from acute, physiological changes (e.g. intense fear) through to more stable personality traits (e.g. irritability; Davou 2007: 40). According to Ekkekakis (2012: 322), *emotions* “are elicited by something, are reactions to something, and are generally about something”. Conversely, *core affects* (such as pleasure, tiredness or tension) are consciously accessible states that are experienced on an ongoing basis, albeit with varying levels of intensity. Unlike emotions, they do not relate to specific situations (including people, events or things, whether past, present, future, real or imagined). Similarly, it is not always easy to identify the causes and stimuli of *moods*, which differ from emotions in that they typically last longer and tend to be diffuse and global rather than specific (ibid.).

Along with the recent reorientation in cognitive science towards a more situated, embodied and distributed understanding, emotions have entered the field as central elements of cognition. It is now increasingly acknowledged that emotion and cognition are indeed inseparable. Emotions play a critical role in all – also high-level – cognition and decision-making (Damasio 1994; 1999): they drive cognition, make it meaningful and steer our attention and motivations. From this point of view, emotions are allocated a primary role in cognition, and it is emotion that “enslaves” the brain and moves the body. Nevertheless, emotions have not undergone such a revolutionary redefinition as cognitive phenomena in cognitive translation studies.

Embodied cognition approaches underline the notion that cognition consists of or is enacted through interaction with social and physical environments. According to Stephan et al. (2014: 67), this also applies to emotions, since neither our modes of thinking nor the way we feel occur in isolation. This calls for an approach which takes into account that cognitive and emotional processes are “possibly (1) linked to our physical state (i.e. are ‘embodied’), and (2) dependent on our environment (i.e. are ‘embedded’), and thus might therefore (3) go beyond the limits of our body (i.e. be ‘extended’), and (4) only develop in the interaction with our environment (i.e. be ‘enacted’)” (Wilutzky et al. 2011: 285; translated by the authors).

Griffiths & Scarantino (2009: 438) support such an approach and demonstrate that the adoption of a situated perspective goes hand in hand with a reorientation in research into emotions:

By shifting theoretical focus from the intrapsychic to the interpersonal, from the unbidden to the strategic, from the short-lived to the long-lived, from the context-independent to the context-dependent, from the static to the dynamic, the situated perspective points the attention of the research community to aspects of emotions that have been unduly neglected and that may hold the key to understanding the nature and function of a large class of emotions. (Griffiths & Scarantino 2009: 448f)

They thus reject the cognitivist understanding of emotions as merely evaluative judgments of internal cognitive representations. This definition might not make sense if cognition or intelligence is not about internally representing the environment and manipulating these representations but rather about (inter)acting in the environment. They also maintain that the situated view of emotions identifies them as “acts of relationship reconfiguration” in a social context and as “forms of skillful engagement with the world [...] scaffolded by [...] and dynamically coupled to an environment” (Griffiths & Scarantino 2009: 438).

This “transactional” and situated view of emotions as temporal processes of continuous exchange with the (social and material) environment thus constitutes an “affective parallel” to the situated view of cognition (Griffiths & Scarantino 2009: 438). Situated approaches therefore view emotions as material and social interactions that are best studied from a process and network perspective.

## **4 Research design and case study setting**

This is precisely the perspective adopted for our study, which aims to illustrate the relevance of the social environment and processes for the emotions trig-



gered in translators at the workplace. To achieve this, we examined authentic translation processes using qualitative ethnographic methods (participatory observation and semi-structured expert interviews). While participatory observation in the translation workplace provides insights into the emotional content of translation processes and translators' interactions with their environment, semi-structured interviews serve to reconstruct the translation processes in their entirety from the (emotional) perspective of the translators. The empirical setting for our case study is the translation department (working languages: German and English) of an Austrian public sector institution, where we conducted observation sessions and seven interviews over a period of 17 working days. Five of the seven interviews were conducted with the department's inhouse translators (T1–T5), one with a retired colleague (T6) who now works for the department on a freelance basis and one with the head of the department.

Participatory observation is a standard field research method in which the researchers do not passively observe the object of study but instead actively participate in the situation in which it is embedded. The researchers thus interact directly with those being studied and collect data by participating in their everyday situation (Mayring 2002: 80). In our participatory observation of the aforementioned translation department, our prerogative was therefore to be open to all manner of different aspects and to take copious field notes that recorded these as precisely as possible. Since this faced us with the problem that not all phenomena in a situation can always be noted down immediately, we expanded our field notes during breaks in and directly after the individual observation periods. Moreover, to enable the researchers to reconstruct the broader context and experience behind the observed actions, participatory observation – as was the case in our study – frequently relies on questions being posed during the observation sessions (Mayring 2002: 82; Flick 2002: 295).

Expert interviews are generally conducted as so-called guided interviews. Accordingly, they are based on a list of open questions (the interview guideline) that is prepared in advance; they are also defined by the specific choice and status of the interviewees (the experts) (Gläser & Laudel 2009: 111; Helfferich 2014: 559). When preparing the guideline, researchers are recommended to keep it as open as possible and as structured as necessary (Helfferich 2014: 566), an approach we likewise adopted when preparing the interview guideline for our own interviews. Our interview guideline (see Appendix A) thus incorporated open requests to talk about the spectrum of activities involved in the translators' own work and any frequently occurring processes therein. It also included questions on the social network and any artefacts (tools) required to fulfil the translators' tasks. Finally, the topic of emotions experienced during translation processes was explicitly

raised. We thereby chose not to use predefined lists of emotions – a common approach in such interviews (Scherer 2005: 712) – since we wanted our experts to talk freely and in their own words about emotional events. We also felt that their work setting could trigger complex emotions that would not necessarily fit into predefined emotion categories. Furthermore, since some of our questions were already covered in the topics raised by the interviewees, we did not always need to ask all the questions in our guideline or keep to the predefined order of questions. In fact, we felt it was more important to give the interviewees space to say what they wanted to say and address the topics that were important to them – regardless of the order in which they were raised.

Both the audio recordings of the interviews and the field notes taken during the observations were then transcribed using an adapted version of the conventions described in Selting et al. (2011). This was followed by a software-assisted (MAXQDA) qualitative analysis of the interview and observation protocols in line with the qualitative content analysis method proposed by Gläser & Laudel (2009). This method helps to extract the relevant information, i.e., it separates it from the original empirical material by applying a set of categories that can be developed deductively from prior theory, inductively by a data-driven procedure, or – as in our case – using a combination of both these strategies (Gläser & Laudel 2013: 21f).

To extract the relevant information from our data, we used six very general deductive categories derived from the *dynamic network model of translatorial cognition and action* by Risku et al. (2013), namely “cognition”, “action”, “social network”, “artefacts”, “environments” and “time”. Within these categories, we constructed a number of emotional subcategories that emerged inductively from the data we had gathered and ranged from enjoyment and satisfaction to frustration and disappointment (see Table 10.1). These emotions were either clearly defined as such by the translators or inferred from the statements recorded in the interviews as well as the comments, actions, gestures and facial expressions noted during the observation sessions. In doing so, we used the aforementioned situated view of emotions as the relational and interactive qualities of what is said and done as a basis for identifying emotions. We then checked the extracted and categorised raw data segments for redundancies and contradictions and sorted them by aspects relevant to our analysis. This structured information base was then used to analyse and interpret the relevant data in line with our research questions (Gläser & Laudel 2009: 202f).

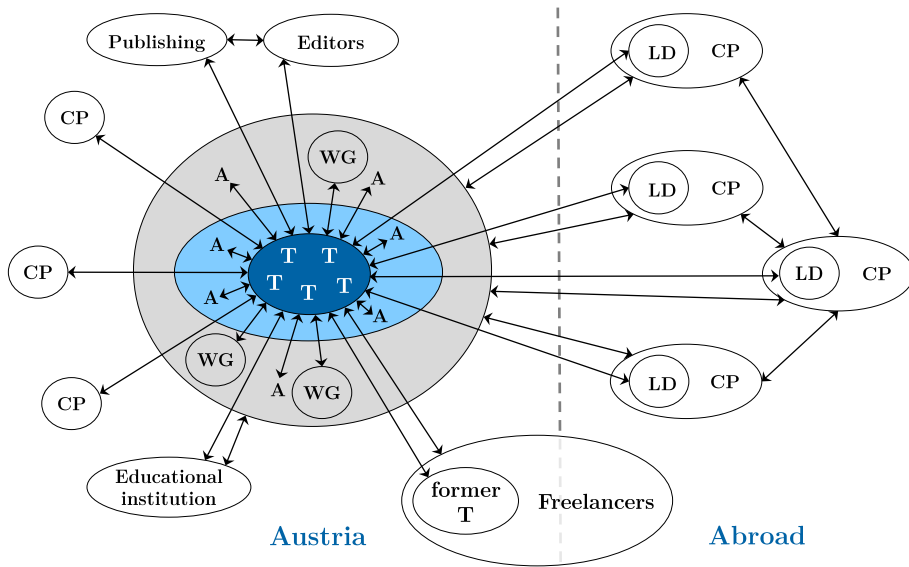
Table 10.1: Emotions related to the nodes in the social network

Nodes	Emotions
Translation team	Enjoyment Ownership Satisfaction
Source text authors (= clients)	Anger Aversion Caution/wariness Enjoyment Impotence Reluctance Stress
Management	Dissatisfaction Frustration Impotence Interest Pride
External translators	Enjoyment Relief
(Non-existent) readers	Disappointment

## 5 Research results

The results obtained from our interviews and observations indicate that the translators working in this public sector institution are members of a complex social network. As Figure 10.1 shows, the translation department (indicated in dark blue) is part of a larger department (light blue), which, in turn, is part of the institution-wide social network (grey).

As voiced in the interviews and noted in the observations, the translators experience a multitude of different emotions in relation to their social embeddedness in this network. These range from satisfaction, pride, relief and enjoyment to interest, caution and ownership (the feeling of having to stand up for themselves and take sides). They also include disappointment, aversion or reluctance as well as stress, a sense of impotence and even frustration and anger.



CP: Cooperation partner      WG: Working group  
 LD: Language department      A: Author

Figure 10.1: The translation department’s social network

The emotions detected in our analysis relate to the following nodes in the social network: the translation team, source text authors (as clients of the translation department), management, external translators and readers (see Table 10.1). Since discussing all the emotions detected would exceed the scope of this paper, we have restricted ourselves to one or two examples for each node.

### 5.1 Translation team: Satisfaction

The translators appear to be largely satisfied with the underlying structures in their team. For example, they hold an annual meeting to decide who will assume responsibility for which publications in the coming year and thus for the corresponding proofreading, translation and administrative tasks (Interview T3).

A regular meeting is also held every two weeks to enable the translators to keep each other updated. When talking about this meeting, T6 notes: “We were

a bit sceptical and worried at the beginning because it's very structured, yet now we really wouldn't want to be without it. It's a very important way for us to share information" (Interview T6).

T1 points out that inconsistencies in the past had necessitated this sharing of tasks, and that they had developed this approach themselves in a professional supervision project. She also explains that the department had begun expanding rapidly around two years after she joined and that "there had to be some reallocation of responsibilities and rotations etc. so that everyone was satisfied with the way the work was shared. [...] [The supervision project] clearly helped us a lot because all our structures are still based on what we agreed back then" (Interview T1).

## 5.2 Source text authors: Stress

Resources within the translation department are extremely limited. The team had previously consisted of six translators and had been able to handle its workload relatively easily, even at peak times. It has since been reduced to 4.5 full-time equivalents, which according to T5 "is okay but not always easy, especially when someone's away, e.g. ill, on a business trip or on holiday. [...] We sometimes have ridiculous amounts of holiday left over that never get used. It's just really busy" (Interview T5).

This heavy workload could also be clearly observed, especially when the translators had to meet very tight deadlines. Yet they appear to take it in their stride: "Why do [the clients] have to get so stressed out about it, like it's the first time we've ever had something come in at the last minute?" (Observation T2). Nonetheless, the translators' reactions to such tight deadlines do differ. According to T4, some of them take a "we'll manage it somehow" attitude, while others say: "They must be joking. If [the client] needs a long text like that [...] we'll need more time" (Interview T4).

Despite the differences of opinion regarding manageable workload, the translators do their best to handle such (large) texts within the team. "We split texts up: I do a couple of pages, [T5] does a couple of pages, and someone else does the quality assurance" (Interview T3). The department also asks its clients to let them know in advance about large texts so it can plan ahead and assign tasks accordingly (Interview T3). "But sometimes you do have to tell the client if it's not possible and try to educate them" (Interview T5).

### **5.3 Management: Frustration and impotence**

T6, whose position was not filled after she retired, laments the staff cutbacks in the department: “Savings start lower down the ladder. [...] The [authors] are sacred, and [the translation department] is a support function. So it’s an easy place to start cutting costs whilst still profiting from the fact that [the translation department] raises the quality of the work” (Interview T6).

T5 expresses frustration at the lack of appreciation for the department that is encountered in the organisation as a whole. Some managers do not even see the need for a translation department “because everyone can speak English anyway, and we can just do it ourselves” (Interview T5). This frustration is shared by T1 and T4, who explain that the department’s work is frequently taken for granted and that it only receives any attention if something has not gone well (Interview T4). “It’s the typical service problem. If you get any feedback at all, then only when someone isn’t happy with something” (Interview T1).

The translators also feel powerless in the face of economic or policy decisions made within the organisation such as the recent decision that authors should draft their texts in English and the translators should simply proofread them – a decision which has since been implemented for the majority of publications. “We often don’t understand the cost-cutting policy. [...] And then they think it’s a really good idea that – because we’ve got one less person – we should outsource texts or just not translate them” (Interview T2).

Yet T2 has also given great thought to how she can be actively involved in this changing work environment, provide added value and continue to enjoy her work. She now supports authors in drafting texts in English by organising writing courses and compiling style guides and lists of useful phrases (Observation T2).

### **5.4 External translators: Relief and enjoyment**

The members of the translation department express relief in the fact that they can draw on a relatively large pool of good external translators and proofreaders with whom they have worked for many years (Interview T1). These include two of their own former colleagues, whose familiarity with the organisation and internal know-how make them highly valuable resources (Observation T4).

Further assistance is obtained through cross-institutional collaboration with other translation departments. In exchange programmes lasting several weeks, colleagues from other institutions have the opportunity to “join” the translation

department and work alongside the inhouse translators. According to T2, having an additional translator support them for a couple of weeks is very helpful (Observation T2).

T3 enjoys the reciprocal nature of this programme, which has already given her the opportunity to work on another organisation's annual report on four occasions. She describes this as "a great way to work" and a chance to "learn so much" (Interview T3). She is likewise pleased that her initial concerns about joining the organisation have proved unfounded:

When I accepted this job, I had the horrifying vision that I'd be sitting in front of a computer all day long with no contact to other people. [...] I'm glad that this nightmarish thought does not reflect reality. We actually have a lot of contact with the authors, the people in our own team or translators in other institutions. (Interview T3)

### **5.5 Readers: Disappointment**

The emotion of disappointment refers here to the feeling that you experience when you realise that something is not what you hoped it would be. Perhaps due to her quality awareness for proofreading and translation, T5 often has the feeling that she is to some extent working for "nothing" (Interview T5). While the translation department puts great effort into delivering texts that are useful and readily understandable for the target reader, she notes that:

[...] with some types of publication, you think from the outset that no one will ever read this. That often makes it difficult from the sense perspective, so you sometimes just have to concentrate more on your pay slip and think, "okay, I've finished that now, it's part of my work, I get paid for it, and it pays the bills". I find that quite hard because I love creating a beautiful text but when you know full well that it will probably only be read by about three people, you find yourself wondering whether it's a meaningful use of your time. But it's always a very mixed bag. (Interview T5)

On rare occasions, T4 also finds herself wondering whether she couldn't have chosen a more meaningful job (Interview T4). "But that usually passes again", she adds, "then a nice piece of work comes along, and you think, 'oh yes, people will want to read this' or that you have made a meaningful contribution to something" (Interview T4).

## **6 Discussion**

While the qualitative study described in this article assigns high priority to the ecological validity of its design, it is also a case study based on a small number of participants, and its findings cannot thus be generalised. Nonetheless, since articles on the topic of emotions are still thin on the ground in translation studies, and few empirical studies as yet address the social embeddedness of emotions in the translation workplace, we felt that a qualitative approach was best suited to obtaining an initial insight into the situative, emotional processes related to social aspects in the translation process.

As can be seen from the examples presented in Section 5, the translators in our study experience a wide range of emotions in relation to their social network in the course of translation processes. While the emotions triggered by working in their own team and with external translators are largely positive, those relating to the institution-wide social network are somewhat mixed.

Our study also provides examples of how translators apply strategies to deal with emotions. In this regard, emotions can be seen as elements of strategies to meaningfully interact with the environment and reconfigure relationships. A review of the negatively valenced emotions triggered by the social network indicates that these did not result in acquiescence but rather in the translators becoming proactive and seeking solutions to problems. Strategies to manage stressful situations within the team have likewise proved effective despite the differences of opinion regarding manageable workload. Strategies have also emerged to deal with developments beyond the translators' control, i.e., the growing importance of English as a lingua franca and the corresponding increase in proofreading and editing tasks due to the switch to mostly English-language publications in the organisation.

## **7 Conclusions**

From an emotional embeddedness perspective, our results provide examples of how the emotions experienced by translators can be studied as “enacted in the interaction with [their] environment” (Wilutzky et al. 2011: 285) and as “forms of skillful engagement with the world [...] scaffolded by [...] and dynamically coupled to an environment” (Griffiths & Scarantino 2009: 438).

Our study also indicates that qualitative ethnographic methods like participatory observation and semi-structured expert interviews can be included in the methodological toolkit of translation scholars studying the emotional aspects of



translation, alongside other tools like focus groups, translators' own narrations in diaries or "fictive love and break-up letters" (Ruokonen & Koskinen 2017) and controlled experiments using, for example, verbal reports or psychophysiological measures. Specifically, workplace research using participatory observation and semi-structured expert interviews provides a way to take account of the translators' social environments and processes. Accordingly, it could be used in future studies to explore how emotions are triggered in translators embedded in different social contexts, thus helping to fill the gaps in our knowledge when it comes to translation and emotions, especially in the situated context.

## **Appendix A Interview guideline**

### **Initial topics**

career history, length of time in translation department, responsibilities

### **A.1 Processes**

- Ask about the interviewee's own tasks: How would you describe the work you do? (e.g. translation, reviewing, copyediting, etc.) → What things have you had to do in the course of your work/as part of your job?
- Ratio: What do you do the most? What do you do less often?
- How would you describe a normal working day for you?
- Process: Describe how you complete a project from start to finish.

### **A.2 Contacts and relationships**

- Who are you in regular contact with (contact persons, clients, etc.)?
- Can you describe what happens when you are contacted by or contact a client? What can cause friction?

### **A.3 Artefacts**

- What do you need to do your work? What are your most important tools?
- What could you not work without?
- How important are the tools that you use?
- Software: Which software tools do you consider to be very/less useful?

#### A.4 Emotions

- How do you feel at work? What do you like a lot? What do you like less?  
→ Describe your emotions in various work steps/tasks.
- Which tasks do you like doing more than others?
- If you could change something, what would it be?
- Are there tools/software programmes that support you well and/or make your work easier? Are there any that make your work more difficult?

## References

- Damasio, Antonio R. 1994. *Descartes' error: Emotion, reason, and the human brain*. New York: Putnam.
- Damasio, Antonio R. 1999. *The feeling of what happens. Body and emotion in the making of consciousness*. New York: Harcourt Brace.
- Davou, Bettina. 2007. Interaction of emotion and cognition in the processing of textual material. *Meta* 52(1). 37–47. DOI: 10.7202/014718ar.
- Ehrensberger-Dow, Maureen & Riitta Jääskeläinen. 2018. Ergonomics of translation: Methodological, practical, and educational implications. In Helle V. Dam, Matilde Nisbeth Brøgger & Karen Korning Zethsen (eds.), *Moving boundaries in translation studies*, 132–150. Oxon: Routledge.
- Ekkekakis, Panteleimon. 2012. Affect, mood, and emotion. In Gershon Tenenbaum, Robert C. Eklund & Akihito Kamata (eds.), *Measurement in sport and exercise psychology*, 321–332. Champaign: Human Kinetics.
- Flick, Uwe. 2002. *Qualitative Sozialforschung: Eine Einführung*. 6th edn. Reinbek bei Hamburg: Rowohlt (Rowohlts Enzyklopädie 55654).
- Gläser, Jochen & Grit Laudel. 2009. *Experteninterviews und qualitative Inhaltsanalyse als Instrumente rekonstruierender Untersuchungen*. 3rd edn. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Gläser, Jochen & Grit Laudel. 2013. Life with and without coding: Two methods for early-stage data analysis in qualitative research aiming at causal explanations. *Forum Qualitative Sozialforschung* 14(2). <http://www.qualitative-research.net/index.php/fqs/article/view/1886>.
- Griffiths, Paul & Andrea Scarantino. 2009. Emotions in the wild: The situated perspective on emotion. In Philip Robbins & Murat Aydede (eds.), *The Cambridge handbook of situated cognition*, 437–453. Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo, Delhi: Cambridge University Press.

- Helfferich, Cornelia. 2014. Leitfaden- und Experteninterviews. In Nina Baur & Jörg Blasius (eds.), *Handbuch Methoden der empirischen Sozialforschung*, 559–574. Wiesbaden: Springer VS.
- Hubscher-Davidson, Séverine. 2009. Personal diversity and diverse personalities in translation: A study of individual differences. *Perspectives: Studies in Translatology* 17(3). 175–192. DOI: 10.1080/09076760903249380.
- Hubscher-Davidson, Séverine. 2013. The role of intuition in the translation process: A case study. *Translation and Interpreting Studies* 8(2). 211–232. DOI: 10.1075/tis.8.2.05hub.
- Jääskeläinen, Riitta. 1996. Hard work will bear beautiful fruit: A comparison of two think-aloud protocol studies. *Meta* 41(1). 60–74. DOI: 10.7202/003235ar.
- Kußmaul, Paul. 1991. Creativity in the translation process: Empirical approaches. In Kitty M. Van Leuven-Zwart & Ton Naaijken (eds.), *Translation studies: The state of the art. Proceedings of the first James S. Holmes symposium on translation studies*, 91–101. Amsterdam: Rodopi.
- Laukkanen, Johanna. 1996. Affective and attitudinal factors in translation processes. *Target* 8(2). 257–274. DOI: 10.1075/target.8.2.04lau.
- Lehr, Caroline. 2014. *The influence of emotion on language performance: Study of a neglected determinant of decision-making in professional translators*. Université de Genève. (Doctoral dissertation).
- Mayring, Philipp. 2002. *Einführung in die qualitative Sozialforschung. Eine Anleitung zu qualitativem Denken*. 5th edn. Weinheim, Basel: Beltz Verlag.
- Risku, Hanna, Florian Windhager & Matthias Apfelthaler. 2013. A dynamic network model of translatorial cognition and action. *Translation Spaces* 2. 151–182. DOI: 10.1075/ts.2.08ris.
- Rojo, Ana & Marina Ramos. 2016. Can emotion stir translation skill? Defining the impact of positive and negative emotions on translation performance. In Ricardo Muñoz Martín (ed.), *Reembedding translation process research*, vol. 128 (Benjamins Translation Library), 107–129. Amsterdam, Philadelphia: John Benjamins.
- Rojo, Ana & Marina Ramos. 2014. The impact of translators' ideology on the translation process: A reaction time experiment. *Minding Translation: MonTI Special Issue* 1. 247–271. DOI: 10.6035/MonTI.2014.ne1.8.
- Ruokonen, Minna & Kaisa Koskinen. 2017. Dancing with technology: Translators' narratives on the dance of human and machinic agency in translation work. *The Translator* 23(3). 310–323. DOI: 10.1080/13556509.2017.1301846.
- Scherer, Klaus. 2005. What are emotions? And how can they be measured? *Social Science Information* 44(4). 695–729. DOI: 10.1177/0539018405058216.

- Selting, Margret, Peter Auer, Dagmar Barth-Weingarten, Jörg Bergmann, Pia Bergmann, Karin Birkner, Elizabeth Couper-Kuhlen, Arnulf Deppermann, Peter Gilles, Susanne Günthner, Martin Hartung, Friederike Kern, Christine Mertzluft, Christian Meyer, Miriam Morek, Frank Oberzaucher, Jörg Peters, Uta Quasthoff, Wilfried Schütte, Anja Stukenbrock & Susanne Uhmann. 2011. A system for transcribing talk-in-interaction: GAT 2 (translated and adapted for English by Elizabeth Couper-Kuhlen and Dagmar Barth-Weingarten). *Gesprächsforschung: Online-Zeitschrift zur verbalen Interaktion* 12. 1–51. <http://www.gespraechsforschung-ozs.de/heft2009/px-gat2.pdf>.
- Stephan, Achim, Sven Walter & Wendy Wilutzky. 2014. Emotions beyond brain and body. *Philosophical Psychology* 27(1). 65–81. DOI: 10.1080/09515089.2013.828376.
- Tirkkonen-Condit, Sonja & Johanna Laukkanen. 1996. Evaluations: A key towards understanding the affective dimension of translational decisions. *Meta* 41(1). 45–59. DOI: 10.7202/002360ar.
- Wilutzky, Wendy, Sven Walter & Achim Stephan. 2011. Situierete Affektivität. In Jan Slaby, Achim Stephan, Henrik Walter & Sven Walter (eds.), *Affektive Intentionalität. Beiträge zur welterschließenden Funktion der menschlichen Gefühle*, 283–320. Paderborn: mentis.

# Name index

- Agarwal, Abhaya, 5  
Allen, Jeffrey H., 33, 80  
Angelelli, Claudia V., 83  
Arcedillo, Manuel, 34  
Arshad, Syed, 4  
Asteriadis, Stylianos, 4  
Aziz, Wilker, 36, 40
- Baayen, R. Harald, 154  
Bahdanau, Dzmitry, 34  
Bajo, M. Teresa, 4  
Balling, Laura Winther, 106  
Bartłomiejczyk, Magdalena, 118  
Bates, Douglas, 154  
Baumert, Andreas, 49  
Beißwenger, Michael, 68, 69  
Benedek, Mathias, 10  
Bentivogli, Luisa, 34, 51  
Bojar, Ondřej, 34, 35, 80  
Bowker, Lynne, 67  
Braga, Camila, 125  
Bredel, Ursula, 48  
Bruker, Astrid, 48, 50  
Brysaert, Marc, 118  
Buch-Kromann, Matthias, 80, 81, 91,  
94  
Buettner, Ricardo, 146, 158
- Cadwell, Patrick, 89  
Calixto, Iacer, 80  
Carl, Michael, 5, 80, 81, 87, 91, 94, 105,  
106, 136, 158
- Carpenter, Patricia A., 158  
Castilho, Sheila, 34, 80  
Cech, Jan, 8  
Chen, Fang, 2, 3, 10, 11  
Chen, Siyuan, 4  
Chmiel, Agnieszka, 164  
Cho, Kyunghyun, 34  
Chomsky, Noam, 124  
Christoffels, Ingrid K., 121  
Cintas, Jorge Díaz, 146  
Ciro, Jairo Buitrago, 67  
Clifton, Charles, 163  
Colina, Sonia, 83  
Cop, Uschi, 107
- Daems, Joke, 5  
Dam, Helle V., 159  
Damasio, Antonio R., 175  
Daniels, Karlheinz, 48  
Davou, Bettina, 174, 175  
De Almeida, Giselle, 80, 91, 125  
De Groot, Annette M.B., 117  
Demberg, Vera, 4, 8  
Densmer, Lee, 89  
DePalma, Donald A., 33, 80  
Depraetere, Ilse, 97  
Dietterich, Thomas G., 14, 19, 23  
Doherty, Stephen, 5, 8, 91, 105  
Dragsted, Barbara, 80, 81, 105, 158  
Duyck, Wouter, 118
- Eghbal-Azar, Kira, 111

*Name index*

- Ehrensberger-Dow, Maureen, 160, 173  
Ekkekakis, Panteleimon, 175  
Elming, Jakob, 91  
Epps, Julien, 4  
Eskenazi, Michael A., 76  
Everdell, Ian, 70  
  
Fazly, Afsaneh, 50  
Ferreira, Aline, 116, 117, 119, 120, 128,  
136, 138  
Fiederer, Rebecca, 54  
Fišer, Darja, 68  
Flanagan, Marian, 89  
Fleiss, Joseph L., 40, 41  
Flick, Uwe, 177  
Folk, Jocelyn R., 76  
Forcada, Mikel L., 36  
Fox, John, 129, 154  
Fox, Wendy, 106  
  
García, Adolfo M., 121  
García, Ignacio, 80, 92  
Gaspari, Federico, 34, 80  
Gehring, Jonas, 11  
Germann, Ulrich, 36  
Giagkou, Maria, 83  
Giménez, Jesús, 55  
Ginsberg, Jay P., 9, 11  
Gläser, Jochen, 177, 178  
Glatz, Daniel, 50, 51  
Goldberg, Joseph H., 8  
González, Meritxell, 55  
Göpferich, Susanne, 80, 124  
Graesser, Arthur C., 165  
Griffiths, Paul, 176, 184  
Grimshaw, Jane, 48  
Grosjean, François, 122  
Grucza, Sambor, 163  
Guasch, Marc, 121  
  
Guerberof, Ana, 33, 34  
Gutermuth, Silke, 48  
  
Haapalainen, Eija, 9  
Halliday, Michael, 125  
Han, Aaron Li-feng, 55  
Hansen-Schirra, Silvia, 48, 51, 55, 146,  
160, 163  
Heine, Antje, 49  
Helfferich, Cornelia, 177  
Herbig, Nico, 2, 6–13, 16, 24, 25, 27  
Hockey, Robert, 4  
Hoffmann, Ludger, 49  
Hong, Jangman, 154  
Hossain, Gahangir, 9  
Hu, Ke, 89  
Hubscher-Davidson, Séverine, 173, 174  
Hutchins, William J., 53  
Hvelplund, Kristian Tangsgaard, 70,  
80, 87, 89, 164  
  
Iani, Cristina, 9  
Ipsen, Helene, 159  
Iqbal, Shamsi T., 4  
Isabelle, Pierre, 35  
Itti, Laurent, 136  
  
Jääskeläinen, Riitta, 173, 174  
Jakobsen, Arnt Lykke, 81, 105, 125, 163  
Jensen, Kristian T. H., 95, 105, 119, 163  
Jespersen, Otto, 48  
Jia, Yanfang, 80, 91  
Junczys-Dowmunt, Marcin, 35  
Just, Marcel A., 158  
  
Kaernbach, Christian, 10  
Kaiser-Cooke, Michèle, 83  
Kibler, Amanda K., 138  
Kienle, Andrea, 69  
Kitchener, Karen Strohm, 148, 158

- Klein, Denise, 117  
Koehn, Philipp, 34, 37  
Koglin, Arlene, 5  
Kokanova, Elena, 164, 165  
Koponen, Maarit, 6, 34, 80, 82, 90–92  
Kościuczuk, Tomasz, 83, 119  
Koskinen, Kaisa, 185  
Kotval, Xerxes P., 8  
Kowler, Eileen, 163  
Kramer, Arthur F., 4  
Krings, Hans P., 4, 34, 81, 90  
Kroll, Judith F., 116, 122  
Kuhn, Jonas, 50  
Kußmaul, Paul, 174  
Kuznetsova, Alexandra, 154
- La Heij, Wido, 117  
Lacruz, Isabel, 4, 8  
Larsen-Freeman, Diane, 137  
Läubli, Samuel, 105  
Laudel, Grit, 177, 178  
Laukkanen, Johanna, 173, 174  
Lavie, Alon, 5  
Lehr, Caroline, 174  
Lenth, Russell, 154  
Lin, Chin-Yew, 7  
Lommel, Arle, 33, 80  
López, Belem G., 122  
Lörscher, Wolfgang, 122, 123
- Maaß, Christiane, 48, 146  
Macizo, Pedro, 4  
Mack, David J., 26  
Malmkjær, Kirsten, 124  
Mann, William C., 128  
Marzouk, Shaimaa, 51, 55  
Masselot, François, 34  
Massey, Gary, 160  
Matthiessen, Christian M.I.M., 125
- Mayring, Philipp, 177  
Mazur, Iwona, 164  
Mellinger, Christopher Davey, 5  
Mesa-Lao, Bartolomé, 80, 87, 89, 91  
Mester, Armin, 48  
Minson, Christopher T., 123  
Mitchell, Linda G., 80  
Moorkens, Joss, 5, 8, 34, 80, 82, 91, 92, 94, 96, 105  
Mulder, Lambertus J. M., 4
- Netflix, 150  
Newmark, Peter, 120  
Nielsen, Jakob, 106  
Nitzke, Jean, 80, 82, 94, 95, 105  
North, Ryan, 50  
Nunes Vieira, Lucas, 80
- O’Brien, Sharon, 4, 5, 8, 34, 54, 80, 82, 89, 91, 92, 94, 96, 105, 111, 125  
Och, Franz J., 7  
Orrego-Carmona, David, 106, 146, 149
- Paas, Fred G.W.C., 3, 5, 7, 16  
PACTE, 123, 124  
Pal, Santanu, 2, 6–13, 16, 24, 25, 27  
Palumbo, Giuseppe, 83  
Papineni, Kishore, 6, 35  
Paradis, Michel, 122  
Parra Escartín, Carla, 34  
Paulsen Christensen, Tina, 89  
Pavlović, Nataša, 95, 118–120  
Pedersen, Jan, 147  
Peltsch, A., 137  
Pernice, Kara, 106  
Plitt, Mirko, 34  
Pokorn, Nike K., 118, 119  
Pollard, Carl J., 51

*Name index*

- Popović, Maja, 51  
Price, Cathy J., 117
- Quaresima, Valentina, 118
- Ramlow, Markus, 67  
Ramos, Marina, 174, 175  
Raptis, Spyros, 83  
Rayner, Keith, 136, 168  
Risku, Hanna, 178  
Rojo, Ana, 174, 175  
Rösener, Christoph, 52  
Rothe-Neves, Rui, 125, 126  
Rowe, Dennis W., 4, 9  
Ruokonen, Minna, 185
- Sag, Ivan A., 51  
Salmi, Leena, 91, 92  
Sánchez-Cartagena, Víctor M., 34, 35,  
51, 80  
Sanchez-Torron, Marina, 34  
Sayeed, Asad, 4, 8  
Scarantino, Andrea, 176, 184  
Schaeffer, Moritz, 80, 82, 94, 106, 146,  
158  
Scherer, Klaus, 178  
Schmaltz, Marcia, 87  
Schnitzer, Brian S., 163  
Schwieter, John W., 116, 117, 120  
Selting, Margret, 178  
Sennrich, Rico, 34, 80  
Shaffer, Fred, 9, 11  
Sharmin, Selina, 91  
Shen, John, 136  
Shi, Yu, 4  
Shreve, Gregory M., 4, 83, 124  
Simard, Michel, 80  
Smith, Michael W., 137  
Snover, Matthew, 5, 7, 55
- Somers, Harold L., 53  
Soukupova, Tereza, 8  
Specia, Lucia, 34  
Stephan, Achim, 176  
Stevenson, Suzanne, 50  
Stewart, Erika, 116  
Storrer, Angelika, 48, 50, 51, 68, 69  
Strecker, Bruno, 49  
Stuyven, Els, 4  
Sweller, John, 3
- Taboada, Maite, 128  
Tatler, Benjamin W., 111  
TAUS, 89  
Thawabteh, Mohammad Ahmad, 164  
Tirkkonen-Condit, Sonja, 174  
Toral, Antonio, 34, 35, 51, 80  
Torgimson, Britta N., 123
- Vaid, Jyotsna, 122  
Valdes, Guadalupe, 138  
van den Berg, Marten, 11  
van Genabith, Josef, 2, 24, 27  
van Merriënboer, Jeroen J.G., 3, 5, 7,  
16  
Van Orden, Karl F., 4, 8  
Vela, Mihaela, 6–13, 16, 25, 27  
Verheijen, Lieke, 69  
Verhein-Jarren, Annette, 49  
Vieira, Lucas Nunes, 2, 3, 5–7, 12, 14,  
15, 19, 25–27  
Vilar, David, 53  
Villarejo, María Viqueira, 4  
Villegas, Marta, 38  
von Polenz, Peter, 48
- Wagner, Emma, 80  
Walker, Francesco, 111  
Wang, Jingxin, 136



Weisberg, Sanford, 129  
Whyatt, Bogusława, 83, 119, 120  
Wilutzky, Wendy, 176, 184  
Winhart, Heike, 50  
Wood, Guilherme Maia de Oliveira,  
126  
  
Yamada, Masaru, 80, 92, 97  
Yamakoshi, Takehiro, 4  
Yeasin, Mohammed, 9  
  
Zifonun, Gisela, 49





# Translation, interpreting, cognition

Cognitive aspects of the translation process have become central in Translation and Interpreting Studies in recent years. Empirical and interdisciplinary studies investigating translation and interpreting processes promise a hitherto unprecedented predictive and explanatory power. This collection contains such studies which observe behaviour during translation and interpreting. The contributions cover a vast area and investigate behaviour during translation and interpreting – with a focus on training of future professionals, on language processing more generally, on the role of technology in the practice of translation and interpreting, on translation of multimodal media texts, on aspects of ergonomics and usability, on emotions, self-concept and psychological factors, and finally also on revision and post-editing. For the present publication, we selected a number of contributions presented at the Second International Congress on Translation, Interpreting and Cognition hosted by the Tra&Co Center at the Johannes Gutenberg University of Mainz.

ISBN 978-3-96110-304-1



9 783961 103041