Machine Translation manuscript No.
(will be inserted by the editor)

Multi-modal Indicators for Estimating Perceived Cognitive Load in Post-Editing of Machine Translation

Nico Herbig · Santanu Pal · Mihaela Vela · Antonio Krüger · Josef van Genabith

Received: date / Accepted: date

Abstract In this paper, we develop a model that uses a wide range of physiological and behavioral sensor data to estimate perceived cognitive load (CL) during post-editing (PE) of machine translated (MT) text. By predicting the subjectively reported perceived CL, we aim to quantify the extent of demands placed on the mental resources available during PE. This could for example be used to better capture the usefulness of MT proposals for PE, including the mental effort required, in contrast to the mere closeness to a reference perspective that current MT evaluation focuses on. We compare the effectiveness of our physiological and behavioral features individually and in combination with each other and with the more traditional text and time features relevant to the task. Many of the physiological and behavioral features have not previously been applied to PE. Based on the data gathered from 10 participants, we show that our multi-modal measurement approach outperforms all baseline measures in terms of predicting the perceived level of CL as measured by a psychological scale. Combinations of eye-, skin-, and heart-based indicators enhance the results over each individual measure. Additionally, adding PE time improves the regression results further. An investigation of correlations between the best performing features, including sensor features previously unexplored in PE, and the corresponding subjective ratings indicates that the multi-modal approach takes advantage of several weakly to moderately correlated features to combine them into a stronger model.

Keywords Cognitive Load · Multi-modality · Post-editing · Machine Translation · Physiological Measurements · Behavioral Measurements

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. We further want to thank the reviewers and editors for their very valuable feedback.

N. Herbig, A. Krüger, J. van Genabith
German Research Center for Artificial Intelligence (DFKI)
Saarland Informatics Campus
E-mail: {nico.herbig, krueger, josef.van_genabith}@dfki.de

S. Pal, M. Vela, J. van Genabith
Saarland University
E-mail: {santanu.pal, josef.vangenabith}@uni-saarland.de, m.vela@mx.uni-saarland.de

1 Introduction

Even though machine translation (MT) systems are improving rapidly, the resulting translations currently still require manual post-editing (PE) to become adequate for many tasks at hand (e.g. for publishing). While current MT systems are mostly evaluated in terms of closeness to an independently provided reference translation, this quality perspective neglects PE costs related to the way in which post-editors work with MT output. Enhancing this process of PE can for example be accomplished by improving computer-aided translation (CAT) tools to better support PE, but also by shifting the optimization goal for MT output towards being useful for the PE task. To measure the usefulness of MT for PE, approaches recording PE time and effort (Guerberof 2009; Zampieri and Vela 2014), quantifying in seconds and keystroke logs the difference between MT output and a human-acceptable translation, have been proposed. We argue that it is not only the amount of PE necessary or the PE time that should be considered, but the actual cognitive load (CL) *perceived* by the post-editor. 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 to process information" (Chen et al. 2016). To frame it within the model of PE effort by Krings (2001), who divided effort into temporal, cognitive, and technical aspects, we propose to focus on the cognitive PE effort.

1.1 Motivation

Especially the PE task has the potential of inducing high 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 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. While CL and MT quality are interrelated, they cannot be considered equal, a simplification often made in the translation domain (cf. Section 2.2). As an example, a long translation with a lot of string overlap with a reference may obtain a high automatic or even subjective evaluation score, but turn out to be difficult to PE and therefore cause high CL. A further difference is that CL may vary with individual post-editor, and this may even to some extent be independent from MT quality (e.g. the number of similar mistakes that have been corrected in the past may impact perceived CL, while the quality remains the same). Due to such examples, it has been argued that CL is a more decisive indicator of the overall effort expended by post-editors (Vieira 2016).

In contrast to almost all related research in the translation domain, we focus on CL as defined in psychology, where it has been well researched and is based on the assumption of a limited available working memory on which load is imposed during cognitive tasks (Chen et al. 2016; Paas et al. 2003; Paas and Van Merriënboer 1994; Sweller 1988). A key finding is that it is important to avoid too high or too low CL to keep subjects motivated and to reduce stress, exhaustion and fatigue. It is also important to note that CL significantly differs from *performance*, since humans have the ability to temporarily increase their effort in order to keep performance

high when a task becomes more demanding; this, however, comes at the cost of additional strain (Hockey 1997).

Such factors like stress and fatigue are currently not considered in MT quality measures but can influence the outcome and cost of PE in terms of required time or occurring errors. Being able to robustly measure CL during PE would enable CAT tools to intervene when high loads are detected, e.g. by suggesting breaks, or providing alternative translations, thereby avoiding overload of post-editors. The automatic capture of CL without interfering in the PE process would also 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. Furthermore, translators could better balance jobs inducing different effort, or even be paid based not only on time or words, but also on CL. To provide first steps towards these goals, in this paper we are concerned with the question of how to actually estimate CL during PE. Approaches to measure CL have been proposed in the past; however, to date there is not much literature that directly focuses on estimating CL during PE.

1.2 Contributions

Our contributions are four-fold: (1) we present an approach based on physiological and behavioral sensor data from a number of modalities and combine them in various ways with each other and with traditional text- and time-based features relevant to the task to cover a number of modalities at the same time. Several of these implemented features have not previously been explored in the translation domain. (2) We investigate 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 and Van Merriënboer 1994). The different modalities and their combinations are then compared in terms of regression performance. (3) We analyze correlations between the best performing features, including some of the unexplored features within the PE domain, and the corresponding subjective ratings to better understand what benefits a multimodal approach has. (4) We publicly release the data used for our analyses, which comprises recordings from a large variety of physiological and behavioral sensors during a PE experiment with 10 translation master's students. The results of our analyses indicate that combining multiple modalities helps in detecting CL.

2 Related Work

This section reviews the most important approaches for measuring CL within the translation domain and other domains and discusses the challenges imposed by the PE domain.

2.1 CL measurements in other domains

Cognitive load theory (Paas and Van Merriënboer 1994; Sweller et al. 1998) comes from 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). Apart from psychology, CL measurement has especially been studied in the field of human-computer interaction (HCI). The approaches can be roughly divided into four categories: subjective measures, performance measures, physiological measures and behavioral measures. *Subjective measures* are based on the assumption that subjects can self-assess and report their cognitive processes after performing a task (Paas and 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* 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).

A lot of research has been done on *physiological measurements*, which assume that human cognitive processes can be seen 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 and Sayeed 2016), and the blink behavior adapts (Van Orden et al. 2001). Furthermore, Chen and 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. Other physiological measures include respiration (Chanel et al. 2008) and brain activity (Hosseini and Khalilzadeh 2010; Solovey et al. 2012). Combinations of such brain activity measures, eye based measures, and subjective measures have also been explored in the context of subtitle processing in movies (Kruger and Doherty 2016; Kruger et al. 2018). Furthermore, the recent improvements in computer vision using deep learning allow automatic extraction of emotions from videos (Kahou et al. 2016). However, simple features such as the head pose have also been shown to correlate to CL when learning (Asteriadis et al. 2009). Last, *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).

2.2 CL measurements for translation

Compared to the HCI domain, only a few, albeit seminal, publications relevant to the cognitive dimension of modeling PE exist. 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 and 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". Mellinger (2014) focused on cognitive effort when using translation memories (TM) by correlating keystroke logs and pause metrics to translation quality ratings. One should note here that while such MT quality measures are most likely related to perceived CL, they cannot be considered equal: consider e.g. very bad MT proposals that are still very easy to PE due to the simplicity of the segments, or the contrary, a very high MT quality where spotting the error can remain difficult and induce a high CL.

The question of which sentence features affect PE effort has been researched as well. Tatsumi (2009) analyzed the relation between automatic evaluation scores and PE speed and found that especially the source sentence length and structure yield to longer PE times. Temnikova (2010) extended an existing MT error classification by ranking the error types in terms of cognitive effort based on a cognitive model of reading, working memory theory, and written error detection studies. However, an analysis of which CL these errors actually induce on editors was not performed. Koponen (2012) compared edit distances to human judgments specifying the amount of PE effort that would be necessary to achieve a useful translation. Similar scales were also proposed by Specia et al. (2010) and Callison-Burch et al. (2010), measuring quality/expected percentage that needs editing and implicitly assuming this to be equal to CL. Koponen et al. (2012) "suggest post-editing time as a way to assess some of the cognitive effort involved in post-editing". Lacruz and Shreve (2014) correlate different error types, classified into mechanical and transfer errors, to PWR, HTER (Snover et al. 2006, 2009), and user ratings of MT quality. Similarly, the work of Popovic et al. (2014) shows that "lexical and word order edit operations require most cognitive effort, lexical errors require most time, while removing additions has low impact on both quality and on time"; however, they simply considered human quality level scores as cognitive effort. To summarize, these works provide insight into which features of a MT output lead to longer PE times or worse subjective quality ratings; however, a direct link to CL in the psychological sense was not shown, but only assumed to exist.

Eye-tracking as a means to capture CL during PE has also been investigated: 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. Moorkens et al. (2015) correlated ratings of expected PE effort with temporal, technical and cognitive effort, in terms of time, TER, 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. 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 and Van Merriënboer 1994) to segments expected to pose different levels of translation difficulty and their corresponding Meteor (Lavie and Agarwal 2007) ratings. In a follow-up work, 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 principal component analysis showed that they cluster in different ways. While these works by Vieira (2014, 2016) are probably the most closely related studies, our approach differs in two important regards: (i) instead of just exploring eye, pause, and time measures, we integrate many more CL measurement methods in a multi-modal fashion that are previously unexplored in the translation domain, and (ii) we analyze how well the self-report CL ratings can be predicted based on these measurements to investigate the feasibility of automatically gathering CL values for segments through different sensors.

2.3 Challenges of translation/PE domain

The translation/PE domain poses a few challenges compared to normal CL studies. First, the task difficulty is of a subjective nature, as it depends on the translator's experience with similar texts, vocabulary, etc.; hence, the translations are not objectively hard or easy. These inter-translator differences could, however, be captured well by subjective measures. Performance measures, on the other hand, besides the problem of compensatory effects (Hockey 1997) discussed above, have the inherent problem that defining performance is by itself not easy in this domain, due to the complex inter-relation of speed and quality. Also, the frequently used dual-task design is impractical, since the focus should remain on the PE task without distraction. Second, the task of PE is very restricted: the translator does not move a lot, is focused on the screen, does not speak, etc. Thus, behavioral measures are limited to mouse and keyboard inputs. Last, any sensors should not hinder the process or make the translator feel uncomfortable, which can be an issue with two-finger GSR sensors, or any EEG sensors. Therefore, physiological measures should focus on wearables and cameras.

3 Towards a robust CL measure for PE

As stated in the introduction, we believe that the CL perceived by translators during PE should be considered more closely, since MT output nowadays often requires PE and only considering the number of changes needed may not be an accurate measure of the effort involved (Koponen 2016). By focusing on the CL during PE, we aim for improved motivation to work and avoidance of boredom, exhaustion and stress. Adding this CL-based perspective on PE of MT to the commonly used but arguably oversimplifying BLEU (Papineni et al. 2002) perspective on MT quality should lead to a better approximation of actual PE cost.

Thus, we need a method to robustly measure CL in PE. The research literature provides a lot of studies in other domains (cf. Section 2.1); however, the question remains which of the related approaches are applicable here. Within the translation domain (cf. Section 2.2) only a few of these approaches have been tested and the focus was mostly not on CL but on perceived MT quality. To test which

measuring approaches can actually reflect different levels of CL in PE, we gather data, which can be combined in a multi-modal fashion, from a variety of sensors during PE. As a ground truth, we use the subjective ratings of perceived CL per segment of each individual post-editor to also capture inter-translator differences. A combination of a set of the gathered sensor data is then correlated to these subjective ratings by regression analysis predicting the rating from the data. The goal is to be able to automatically infer the CL from the raw sensor data during PE to avoid interruptions by asking for these ratings. Ideally, this should work using as few and as commonly used sensors as possible to prevent overhead and make it more feasible in practice. In this section, we present the steps we have taken so far for building a robust CL measure for PE.

To assess data from multiple modalities during PE, we implemented a framework combining several sensors that show correlations with CL in other domains, as well as other sensors that we considered interesting possible indicators of CL in PE. A node.js server, running on the same machine as the PE is done on, retrieves data via web sockets and stores it to a database. The system is event-based; thus, whenever a sensor acquires data, it is sent as a JSON event to the server. To calculate higher-level features based on a combination of raw data during runtime, it is also possible to subscribe to specific events, process them, and send the resulting high-level feature back to the server.

Our most basic sensor is a keylogger storing all keyboard and mouse input during PE. The higher-level pause features APR and PWR by Lacruz et al. (2012) are automatically calculated from the keyboard events. Our software also listens to the shortcut to switch to the next segment within the CAT tool and intervenes by showing a pop-up asking for a subjective CL rating. As a subjective rating scale for CL, we decided to use the one proposed by Paas and Van Merriënboer (1994), since it focuses on CL and not on quality, has been widely used and verified in many application areas, can be answered quickly as it contains only a single question (in contrast to NASA-TLX (Hart and Staveland 1988)), and allows ratings on a 9-point scale, thereby offering a sufficiently wide range to select from. The single question is 'In solving or studying the preceding problem I invested' with answer possibilities from 'very, very low mental effort' to 'very, very high mental effort'.

We integrate the remote Tobii eye tracker 4C, since it is cheap, offers high-quality data and can therefore be considered as a candidate for real-world usage. With it, we record the raw gaze data, detect the amount of blinking (Blinks), and compute the average fixation amount (Fix_{avg}) and average saccade durations ($\text{SaccDur}_{\text{avg}}$), all of which have been shown to be indicators of CL. Furthermore, we calculate the probability of visual search (Goldberg and Kotval 1999) ($\text{SearchProb}_{\text{avg}}$), which was used to find user interface flaws, hoping that it might also help determine CL. We did not use the pupil diameter, because it requires more expensive hardware, has a long latency, and is less feasible to measure under changing illumination.

For cardiovascular measures, we integrate a Polar H7 heart belt communicating with the computer via Bluetooth Low Energy. It measures the heart rate and the RR interval, which is the length between two successive Rs (basically the peaks) in the ECG signal. Based on this, we calculate the often-used CL and stress measures of heart rate variability (Rowe et al. 1998), in particular the root mean square of successive RR interval differences (RMSSD) and the standard deviation of NN intervals (SDNN). Since the SDNN uses NN intervals, which normalize across the

RR intervals and thereby smooth abnormal values, we expect it to be influenced to a lesser degree by outliers, but at the same time to react more slowly to changes in HRV. As with the other features, the calculated values are normalized per participant and averaged per segment ($RMSSD_{avg}$, $SDNN_{avg}$). This normalization is achieved by projecting all values of a participant to the interval [0,1].

We integrate the Microsoft Band v2, a small bracelet offering a variety of sensors, including a galvanic skin response (GSR) sensor. As more and more people wear such devices in their daily lives, we argue that it could also be accessible during PE in the near future. As described in detail in Chen et al. (2016), three features are calculated from the raw data: the accumulated and average GSR per segment after normalizing per participant (GSR_{acc}, GSR_{avg}), and the equivalent to GSR_{avg} in the frequency domain ($FreqGSR_{avg}$). Similar to Chen et al. (2016), normalization is achieved here by dividing each value of the participant by the participant's average GSR value.

Two web-cams are integrated into the system. The first one simply records images at a fixed interval. These are then sent to an emotion recognition API like Microsoft Cognitive Service[1], returning a simple JSON format with the likelihood of each of the basic emotions based on a trained neural network. The basic emotion values are normalized per participant and the mean is calculated per segment ($EmotionName_{avg}$). The second web-cam is used to calculate the eye aspect ratio, which indicates the openness of the lids. For this, we re-implemented the work of Soukupova and Cech (2016) and average the values per segment (EAR_{avg}). Even though both web-cam based features have not been shown to be an indicator of CL in the literature, they are included because intuitively a link might exist and the simplicity of using web-cams would make the CL measurement easily applicable in practice.

Last, a Kinect v2 captures the body posture. We hypothesize that post-editors come closer to the screen for hard-to-edit translations. The distance to the head is normalized per participant and the mean distance per segment is calculated ($HeadDist_{avg}$).

Apart from the sensors, we need to generate state-of-the-art translations for our experiments that contain realistic error types. For this we adapted the ConvS2S neural machine translation (NMT) system (Gehring et al. 2017) trained on English-German parallel data from the WMT 2017 translation task. We use an ensemble of four expert ConvS2S NMT models with different random weight initializations. To mitigate the label bias problem (Lafferty et al. 2001), each model was trained separately to decode from left-to-right and right-to-left, i.e., we achieve a left-to-right and right-to-left decoding symmetry for MT. Finally, we re-score hypotheses by interpolating left-to-right and right-to-left scores with uniform weights. Before training our NMT model, we preprocessed words into subword units (Sennrich et al. 2016). We followed the best hyper-parameter settings as described in Gehring et al. (2017). During translation (i.e., at the decoding time) we set the beam size to 5. The overall performance achieved by our NMT system is 29.5 in BLEU and 60.1 in TER on the WMT 2017 test set. Compared to the best system in WMT-2017 (Sennrich et al. 2017), ConvS2S achieves +1.2 BLEU and -1.1 TER absolute points. These state-of-the-art results should therefore properly reflect the types of errors currently occurring in MT outputs.

[1] https://azure.microsoft.com/services/cognitive-services

4 Experiment

We conducted an experiment to see if and how we can automatically determine the CL perceived during PE and whether our multi-modal approach facilitates the CL measurement process[2]. All data used throughout this experiment is publicly available at `http://mmpe.dfki.de/data/MTJournal2019`.

4.1 Text selection

Similar to Vieira (2016), we used a subset of the WMT 2017 news translation task test set as texts for this study. After using our NMT system, we extracted 300 sentences and their translations, 100 each within different TER score intervals[3]. All segments had a length of ≤ 35 words. Out of these 300 sentences, we extracted 60 segments based on error rules to ensure different difficulties are represented in this set. For this, we categorized the errors contained as being either errors of lexical choice, containing mistranslated words or errors in fluency, or errors in word order. By selecting sentences containing these error types and combinations thereof, we hoped to induce different levels of CL on the participants.

To further reduce the amount of segments and to ensure that these actually can cause different levels of CL on the participants, we performed a pre-study (with counterbalanced segment order). Two German natives with a similar English skill level, as both are in the same translation science master's program, participated and translated the 60 segments. As described above, a pop-up appeared after each segment asking for a subjective CL rating. We used the resulting 2 times 60 segment ratings to pick 30 for the final study. For this selection, we filtered out segments with disagreement >3 on the 9-point Likert scale, meaning that they had at least a similar judgment. To pick 30 sentences, the remaining sentences were ordered by average rating, and we removed multiple segments with equal average ratings to achieve an equal rating spread. The hope was that this well-distributed set of CL perceptions among the participants of this pre-study leads to transferable ratings in the final study. Note however, that we did not use the pre-study ratings as the CL labels for the following actual study, but only to perform this pre-selection of segments. In the main study we again ask the participants for CL ratings, and use their individual ratings for the analysis to capture inter-participant differences.

All participants in the final study used these same 30 segments; however, the order is randomized to avoid ordering effects. While using WMT data, which consists of independent segments instead of complete texts, prevents us from analyzing the effects of textual (i.e. cross-sentential) coherence and cohesion on CL, it allows us to perform this randomization of segment order which would not make sense with a complete text. Since each participant receives the same segments in a different order, potential effects such as feeling tired towards the end of the experiment do not always affect the same segments and therefore balance out.

[2] The study was approved by the university's ethical review board and the data protection officer.

[3] As TER intervals we used [35-50], [60-70], and [80-95].

4.2 Apparatus

For the study, the post-editor is equipped with a Microsoft Band v2 on her right wrist, the heart belt on her chest, and an eye tracker, as well as two web-cams 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 SDL Trados Studio 2017 translation environment. We chose Trados for this study as it is by far the most used CAT tool in professional applications.

4.3 Participants

The experiment participants were 10 native German speakers enrolled in the translation master's degree program, who had already attended a CAT tools class where they had completed the SDL certification program including practice sessions. From that class, all of them were familiar with MT concepts and PE. Overall, 7 female and 3 male paid students, aged 22 to 32 (average, 25.9), participated.

Prior to the actual experiment, the participants were asked to fill out a data protection form and a basic questionnaire gathering demographics as well as language skills and translation/PE experience. Furthermore, 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 unnecessary editing. Concrete time limits were not stated. The reason for clearly specifying how detailed the corrections should be was to ensure a similar PE process across participants; other specifications would also have been valid for such an experiment. 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.

4.4 Subjective CL Ratings

All 9 CL ratings were used during the experiment; however, 89.7% of the ratings were within the range 3 to 7 (inclusive) while the extreme cases were only rarely chosen (see Figure 1 for the rating distribution). We also observe rating differences between post-editors, with an average standard deviation across segments of 1.3 (minimum 0.8, maximum 2.1). Note that we use these individual CL ratings for the remaining analyses to also capture the differences in CL perceptions between participants. A reason for the non-uniform, rather normal rating distribution could be the strong wording chosen by the authors of the scale we used to assess perceived CL (Paas and Van Merriënboer 1994): 'very, very high/low mental effort' is something that we believe users simply do not identify themselves with often. Even though we invested work in finding segments that we expected to induce very, very low or high mental effort through the pre-study, the inter-personal differences seem to simply be too high to ensure this. These inter-personal rating differences also show why CL and the BLEU perspective of MT quality cannot be considered equal, since the latter is an objective measure, while perceived CL is an inherently subjective variable and depends on how individuals cope with variation in the demands of a task (Vieira 2016).

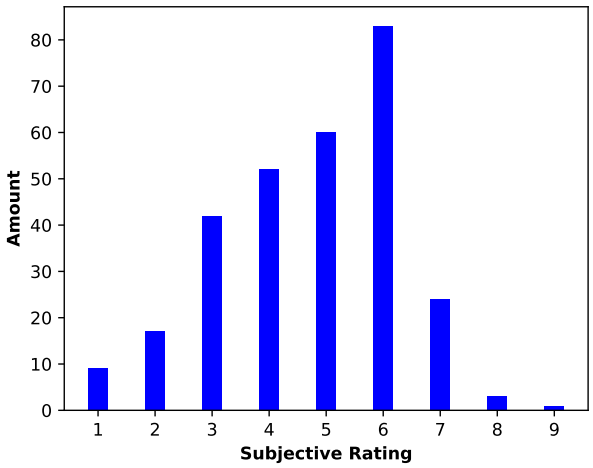


Fig. 1: Segment distribution across subjective CL scale

4.5 Evaluation Method

Based on these subjective ratings and the sensor data corresponding to these ratings, we conduct an analysis consisting of two parts. First, we investigate how well the CL perceived by the individual participant can be predicted from different modalities and whether a combination of modalities improves the accuracy (cf. Section 4.6). Second, we look at the concrete features that performed well in this first stage and analyze their correlations with subjectively measured CL. This second stage provides further insights into reasons for and against using multi-modality (cf. Section 4.7).

For both analyses, we designed four categories of feature sets, which are compared against each other: (1) *time*-based features, (2) *text*-based features, (3) *sensor*-based features, and (4) a *combination* of the previous three.

Here, the *time* features are the post-editing time (PeTime) or the length-normalized post-editing time (LnPeTime); the *text* features consist of smoothed BLEU, HBLEU (Lin and Och 2004), TER, HTER (Snover et al. 2009), and sentence length (SL), as well as all combinations thereof. 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 translation and calculate the amount of necessary edits to transform the hypothesis into the reference, while HBLEU and HTER perform the same calculation, but this time between the hypothesis translation (the MT output) and the post-edited translation. For the *sensor* features, we analyze the different features that were engineered on the raw sensor data (see Section 3). The modalities heart, eyes, skin, keyboard, body posture, and emotions are evaluated individually and combined. For the *combinations*, we combine these *sensor* combinations with the *time* and *text*-based features.

4.6 Multi-Modal CL Detection – Regression Analysis

The goal of this stage is to learn a function that best fits the implemented *time*, *text*, *sensor*, and *combined* features to the CL as reported by each participant on the subjective 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, because we focus on the CL perceived by each individual and not on any general measure of quality. Apart from comparing the different regression models against each other, we also compare each model to two simple baselines: (1) always predicting the mean subjective rating ($\text{SubjCL}_{\text{avg}}$), and (2), always predicting the median subjective rating ($\text{SubjCL}_{\text{median}}$).

Since different features and combinations of features require different types of functions to best approximate them locally (e.g. not all of them show linear, polynomial, or radial relations), we train not only one, but several regression algorithms making different assumptions about the underlying function space: a support-vector regressor (SVR) with a radial basis function kernel, and 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[4] using the default parameters and feature normalization. In this way, for each feature and group of features we obtain locally optimal results before comparing them and drawing conclusions on the usefulness of the features involved. While this approach might miss some ideal hyper-parameter combination, it offers a reasonably wide range of function spaces to choose from and, furthermore, we did not want our results to be biased and possibly be distorted by the use and limitations of a single classifier (and with it the class of functions that can be learned).

Please note that the rating scale used (Paas and Van Merriënboer 1994) is ordinal; however, the outputs of the regressors can be continuous. The reason is that we explicitly decided to use the scale as it was designed and verified without any alterations, but did not see value in forcing the models to output ordinals because their target value, CL, spans a continuous space. To avoid over-fitting, all regression functions use regularization or averaging, and we perform cross-validation. Missing data values for features are replaced by the mean of the feature values across all participants and segments.

We report the results of the individual features, of combining features within a modality, and of combining features across modalities. Feature combination is always achieved using simple vector concatenation. Whenever the space of possible feature combinations becomes too large, 1000 samples of random feature combinations of a maximum of 5 and a maximum 10 features per combination are used instead of all possible combinations. For the *sensor* and *combined* feature sets, we ensure that features of different modalities are combined: for the *sensor* features, features of multiple sensor modalities are mixed, and for the *combined* sets, we ensure that at least one feature of *time* or *text* is combined with one or multiple *sensor*-based features.

[4] http://scikit-learn.org/

For all of these feature combinations, we train each of the above regressors using a 10-fold stratified cross-validation, which also considers the inbalanced rating distribution. For each regressor, the average test mean square error (MSE) is computed across the 10 folds. This average score is then compared across regressors as it is a good measure for our actual goal: predicting the 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.

For each reported model, we also perform a 5 by 2 cross-validation 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.

Since we expect that more information helps predict perceived CL, we hypothesize that *combinations* perform best, followed by *sensor*-, then *text*-, and last, *time*-based features.

4.6.1 Results

The regression results are presented in Table 1. It is divided into the five categories *baselines*, *time*, *text*, *sensor* and *combined* features. First of all, one should note that the results for 10-fold and 5 by 2-fold cross-validations are rather similar, which indicates robustness of the models that is also reflected in the small standard deviations. We compare each 5 by 2-fold cross-validation MSE score using a univariate ANOVA with all models as conditions and calculate the contrasts to the mean and median baselines as references. Both ANOVAs violated the sphericity assumption but still showed strong significance ($p < 0.01$) after Greenhouse-Geisser correction of the degrees of freedom. The table shows that all models are significantly better than the median baseline, and that most but not all models are also significantly better than the mean baseline (after Bonferroni correction).

Apart from comparing the models against the baselines, we also perform pairwise comparisons between the best models of the categories *time*, *text*, *sensor*, and *combination*, which we report in Table 2. For the pairwise comparisons we use the modified paired t-tests as described in Dietterich (1998).

For *time* features, we notice that PeTime performs better than its length-normalized alternative, and that both are significantly better than the two baselines. In contrast, the results for the *text*-based features do not differ as much from each other, and are closer to the baselines, where BLEU, HTER, and HBLEU are not significantly better than the mean baseline. Note in particular that contrary to our expectation the results are worse than those for the *time* features.

The *sensor* features are again separated into the individual modalities. The combined eye-based features show the best results, followed by the skin, keyboard, and then heart. Inferred emotions and body posture considered individually show worse results. Regarding inferred emotions, we only report the best emotion and the best combined set of emotions, as all others had very similar results and in general the MSE's are very close to the baselines, indicating that these features in this simple form do not perform well.

The last section among the sensor features shows that using a combination of multiple modalities improves results considerably compared to each modality used alone, and that this combination also performs better than the *time* and *text*-based features. Here, the best result for up to 10 features and the best result for

up to 5 combined features are reported, even though several other combinations with similar results were found among the sampled features. The last section in the table shows the results when *combining* not only *sensor* modalities, but also incorporating *time* and *text* features. These results are also better than those of multi-modal *sensor* features. Again, the best results for up to 10 and up to 5 feature combinations are reported.

We further use the 5 by 2 cross-validation results in combination with a modified t-test (Dietterich 1998) followed by Bonferroni-Holm corrections to test the differences between the best models of *time*, *text*, *sensor*, and *combined* features for significance. Table 2 shows that indeed the *combined* approach is significantly better than *time* and *text*, and that *sensor* is better than *text*; however the other pairs reported in the table do not show significant differences.

4.6.2 Discussion

Although the concrete ratings differ between post-editors, the methods to measure CL, especially the multi-modal ones, are apparently transferable across participants. When comparing time and text features, we are surprised to see that PeTime seems to be the better, albeit not significantly better, measure of perceived CL, which also performs quite well in general. The sentence length and therefore length normalization does not seem to provide further insights in terms of CL. Interestingly, the H-based text features do not improve results compared to BLEU/TER as we have expected, and even contrarily, do not beat the simple mean baseline on our dataset. A reason for this could be that CL does not focus on how much needs to be edited, but on how difficult it is to do so, which strengthens the need for CL detection. Inspecting the data in further detail, we find 60 out of 260 cases where multiple participants rated the same segment as equally tough while having an editing difference of more than 30 HBLEU. This supports our above argument, that several cases exist where strong differences in editing behavior do not impact the CL perception.

While the heart features all significantly outperform the baseline, they generally show similarly bad results as the text features. Based on the literature, we were expecting to find better results here. In comparison to this, combining several eye features yields the best results among all individual modalities, and also better results than any *time* or *text* feature. Interestingly, the amount of blinking alone already shows good results and is better than eye-tracking data using only web-cam data (i.e. EAR_{avg}).

Combinations of GSR-based features or the accumulated GSR value also work comparatively well, however, we had expected better results based on the literature here. Since smartwatches are spreading and often include GSR sensors, this data is especially interesting because it could be easily read by future CAT tools. For the keyboard features we see only small differences between PWR and APR, and the combination of both does not boost the model's performance. Based on the findings by Lacruz et al. (2012), we also expected better results for these features.

The normalized distance to the participant's head does not perform better than text-, time- or many of the sensor-based features and while being significantly better than the baseline, the gains are diminishing small. Maybe more complex features on the human body posture can provide better results in the future. Emotions also do not perform better than most of the other features and the gains

Features	MSE 1x10-CV↓(Reg.)	MSE 5x2-CV↓ (SD)
Baselines		
$SubjCL_{avg}$	2.466 (-)	2.465 (0.093)††
$SubjCL_{median}$	2.540 (-)	2.538 (0.093)**
Time Features		
PeTime	1.891 (Ridge)	1.878 (0.061)**††
LnPeTime	2.052 (Lasso)	2.037 (0.091)**††
Text Features		
BLEU	2.330 (RF)	2.380 (0.118)††
TER	2.340 (RF)	2.350 (0.159)*††
HTER	2.311 (EN)	2.383 (0.174)†
HBLEU	2.341 (EN)	2.384 (0.150)††
SL	2.437 (Ridge)	2.444 (0.087)*††
HBLEU, TER, SL	2.261 (Ridge)	2.321 (0.165)*††
Sensor Features		
Heart		
$RMSSD_{avg}$	2.285 (Ridge)	2.282 (0.054)**††
$SDNN_{avg}$	2.352 (Ridge)	2.379 (0.078)**††
$RMSSD_{avg}$, $SDNN_{avg}$	2.304 (SVR)	2.309 (0.057)**††
Eyes		
Blinks	2.034 (Ridge)	2.040 (0.062)**††
Fix_{avg}	2.276 (SVR)	2.292 (0.131)**††
$SaccDur_{avg}$	2.415 (Lasso)	2.421 (0.122)††
$SearchProb_{avg}$	2.462 (Lasso)	2.247 (0.094)††
EAR_{avg}	2.424 (Ridge)	2.438 (0.093)**††
Blinks, Fix_{avg}, $SearchProb_{avg}$, EAR_{avg}	1.704 (RF)	1.803 (0.175)**††
Skin		
GSR_{avg}	2.462 (Lasso)	2.461 (0.093)††
GSR_{acc}	2.181 (Lasso)	2.185 (0.041)**††
$FreqGSR_{avg}$	2.402 (Ridge)	2.383 (0.082)*††
GSR_{avg}, GSR_{acc}, $FreqGSR_{avg}$	2.074 (Ridge)	2.117 (0.079)**††
Keyboard		
APR	2.307 (Ridge)	2.311 (0.139)**††
PWR	2.259 (SVR)	2.265 (0.128)**††
APR, PWR	2.219 (Ridge)	2.247 (0.139)**††
Body Posture		
$HeadDist_{avg}$	2.445 (SGD)	2.460 (0.095)**††
Emotions		
$Anger_{avg}$	2.430 (SGD)	2.445 (0.089)**††
$Anger_{avg}$, $Neutral_{avg}$, $Sadness_{avg}$, $Surprise_{avg}$	2.383 (RF)	2.420 (0.101)**††
Combined Sensors		
Keyboard (TER) Eye (Blinks, Fix_{avg}, $SaccDur_{avg}$, EAR_{avg}) Skin (GSR_{acc}, GSR_{avg}, $FreqGSR_{avg}$) Heart ($SDNN_{avg}$, $RMSSD_{avg}$)	1.512 (RF)	1.639 (0.153)**††
Eye (Blinks, Fix_{avg}, EAR_{avg}) Skin (GSR_{acc}, GSR_{avg})	1.595 (RF)	1.646 (0.115)**††
Combined Features		
Time (PeTime) Keyboard (APR, PWR) Eye (Blinks, Fix_{avg}, EAR_{avg}, $SaccDur_{avg}$, $SearchProb_{avg}$) Skin ($FreqGSR_{avg}$) Heart ($RMSSD_{avg}$)	1.434 (Ridge)	1.487 (0.069)**††
Time (PeTime) Skin ($FreqGSR_{avg}$) Eye (Blinks, $Fixs_{avg}$) Heart ($RMSSD_{avg}$)	1.490 (Ridge)	1.508 (0.084)**††

Table 1: Feature evaluation results for 10-fold and 5 by 2-fold cross-validation (CV) with standard deviation (SD). An asterisk (*) in the right column indicates a significant difference to $SubjCL_{avg}$, while a dagger (†) indicates a significant difference to $SubjCL_{median}$. */† represent $p < 0.05$, **/†† represent $p < 0.01$ after Bonferroni correction.

Time: PeTime	**Time**		
Text: HBLEU, TER, SL	$\tilde{t} = 2.79$	**Text**	
Sensors: Keyboard (TER) Eye (Blinks, Fix_{avg}, $\text{SaccDur}_{\text{avg}}$, EAR_{avg}) Skin (GSR_{acc}, GSR_{avg}, $\text{FreqGSR}_{\text{avg}}$) Heart ($\text{SDNN}_{\text{avg}}$, $\text{RMSSD}_{\text{avg}}$)	$\tilde{t} = -2.07$	$\tilde{t} = -7.06$**	**Sensors**
Combined: Time (PeTime) Keyboard (APR, PWR) Eye (Blinks, Fix_{avg}, EAR_{avg}, $\text{SaccDur}_{\text{avg}}$, $\text{SearchProb}_{\text{avg}}$) Skin ($\text{FreqGSR}_{\text{avg}}$) Heart ($\text{RMSSD}_{\text{avg}}$)	$\tilde{t} = -10.75$**	$\tilde{t} = -4.59$*	$\tilde{t} = -0.55$

Table 2: Pairwise comparisons between the best models of *time*, *text*, *sensors*, and *combined* features. * shows significance with $p < 0.05$, while ** means $p < 0.01$ after Bonferroni-Holm correction. $\tilde{t}$ is the test statistics for the modified paired t-test (Dietterich 1998).

compared to the baseline, albeit significant, have limited practical use. Again, further investigation and more complex features than the normalized mean might improve this in the future.

Combining the different sensor modalities improves the results, showing the advantage of our multi-modal approach. This is in line with Vieira (2016)'s discussion after analyzing the correlations between eye, keyboard, time, and subjective measures, stating that "different measures may be more sensitive to different nuances of cognitive effort, which would imply that, while a single construct, cognitive effort might have different facets". Our combined *sensor* modalities improve (insignificantly) over *time* and (significantly) over *text* features (cf. Table 2), but also seem better than any individual modality. When combining across *time*, *text*, and *sensor* features, even better results are achieved, which significantly outperform both *time* and *text* features. Generally, these results show that combining multiple modalities of CL indicators improves the regression quality, especially in comparison to each individual modality.

To summarize, while almost all individual features statistically outperform the baselines, the gains of most features are small; thus, the only practically really interesting features are PeTime, the combination of several eye features, and in particular the combination of features from several modalities. Regarding our hypothesis stated earlier, we could show better results for *combined* than for *sensor* features, which again outperformed *time*- and *text*-based features. However, con-

trary to our expectations, *time* was a considerably better measure than *text*. One should note that these results were achieved without optimizing feature preprocessing, that no hyper-parameter tuning was applied, and that simple random sampling of feature sets was used, because we were only interested in a fair comparison between the methods and not in finding the best possible model. Using a more informed approach might therefore decrease the MSE's in the future.

4.7 Why Multi-Modality Helps – Correlation Analysis

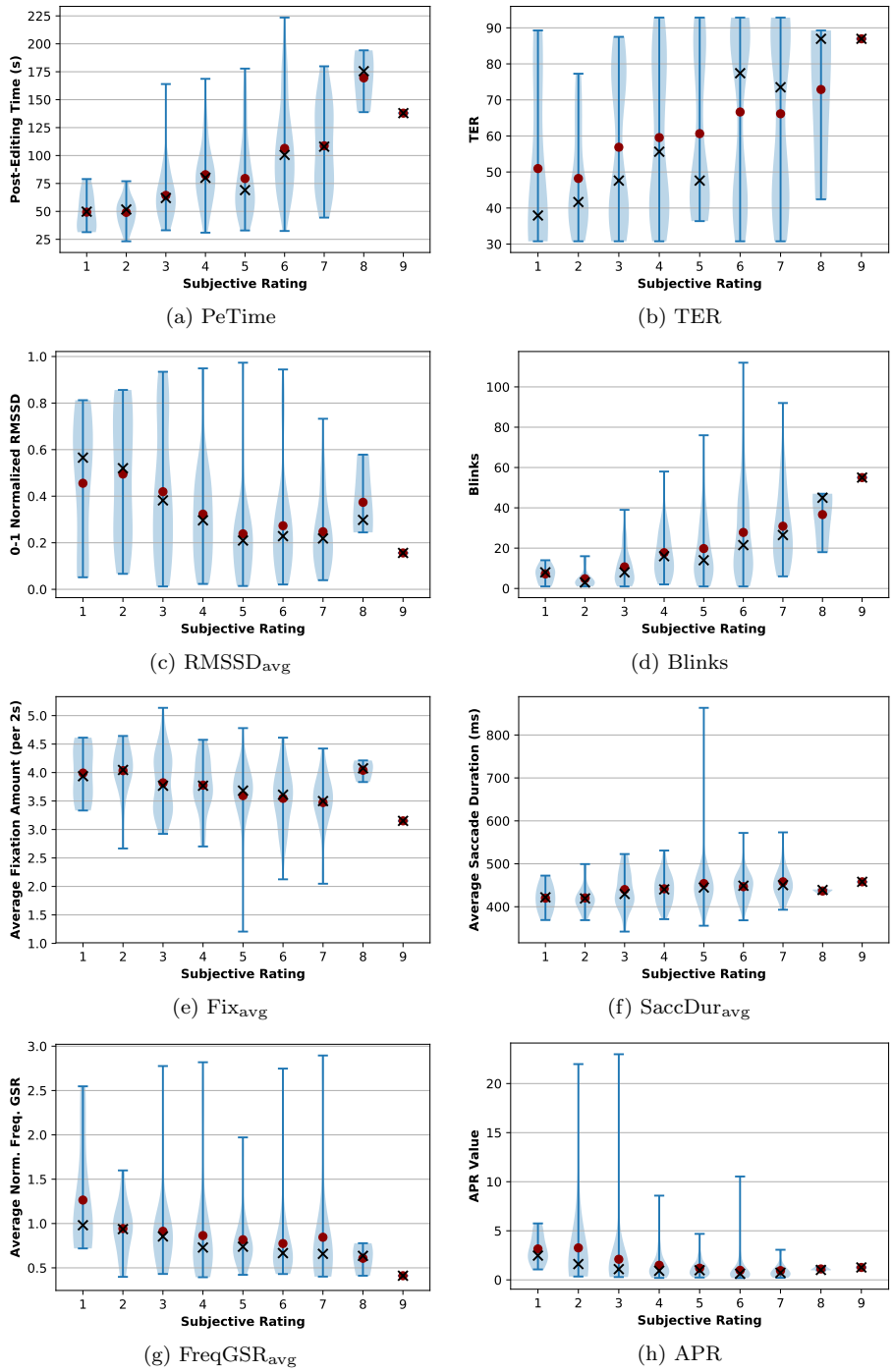
After inspecting the overall performance of different modalities and their combinations in terms of regression analysis, we now inspect the individual features in more detail. For space reasons, however, we cannot discuss every single feature. Instead, we focus on some of the features used by the best-performing regressor in the *combined* feature sets, and additionally the TER feature to also include a *text*-based feature.

4.7.1 Results

Figure 2 shows violin plots of the individual feature values plotted against the subjective CL ratings provided by each participant for the segments on which those features were calculated. Inspecting the course of the means (circles) or medians (crosses), we notice that there is a certain dependence between the individual features and their corresponding ratings. At the same time, we can clearly see a lot of noise around those means/medians (note that the limits in violin plots are the minimum and maximum values).

An analysis of Spearman's correlations between those features and the corresponding subjective ratings yields further insights into why our various regressors perform differently. To interpret the correlation coefficients, we use the interpretation of Corder and Foreman (2009), stating that values around ± 0.1 can be considered as weak, values around ± 0.3 as moderate, and values around ± 0.5 as strong correlations.

As can be seen in Table 3, PeTime strongly correlates ($+0.505$) with the subjective ratings, which explains why the regressor trained solely on that feature already performs quite well. This can also be seen in the corresponding plot, showing an upwards tendency with only a moderate amount of noise. The text feature TER on the other hand shows a lot of noise and a strong divergence between means and medians. The correlation coefficient of $+0.276$ can be interpreted as moderate. Contrary to the results for TER, there is a negative correlation for the heart feature $\text{RMSSD}_{\text{avg}}$ (-0.220) that is weak to moderate. For the eye features Blinks, Fix_{avg} and $\text{SaccDur}_{\text{avg}}$, we find strong positive ($+0.453$), moderate negative (-0.262), and weak to moderate positive correlations ($+0.193$), respectively. For skin features, we can observe moderate negative correlations (-0.264) with subjective CL ratings. One should note here that for the plot and calculation the imaginary part of this feature was dropped. Last, for the keyboard-based feature APR we can also observe moderate negative correlations (-0.308). All reported Spearman correlations are statistically significant with p-values < 0.001.



(a) PeTime (b) TER

(c) $\text{RMSSD}_{\text{avg}}$ (d) Blinks

(e) Fix_{avg} (f) $\text{SaccDur}_{\text{avg}}$

(g) $\text{FreqGSR}_{\text{avg}}$ (h) APR

Fig. 2: Violin plots for different feature values per subjective rating (x-axis). The circles indicate the means, the crosses the medians.

Feature	**Spearman's ρ**	**Interpretation**	**p-value**
PeTime	+0.505	Strong	< 0.001
TER	+0.276	Moderate	< 0.001
$\text{RMSSD}_{\text{avg}}$	−0.220	Weak to moderate	< 0.001
Blinks	+0.453	Strong	< 0.001
Fix_{avg}	−0.262	Moderate	< 0.001
$\text{SaccDur}_{\text{avg}}$	+0.193	Weak to moderate	< 0.001
$\text{FreqGSR}_{\text{avg}}$	−0.264	Moderate	< 0.001
APR	−0.308	Moderate	< 0.001

Table 3: Spearman's correlation results between different features and subjective CL ratings.

4.7.2 Discussion

These results show why multi-modality helps: apart from PeTime and Blinks, all reported correlations are weak to moderate, so by themselves not sufficient for good subjective CL detections. However, each modality provides a little more insight into the overall CL perception. Therefore, combining features of several modalities in a single regressor increases its performance. This is also why the best regressor of the eye features (cf. Table 1), or the regressors of *combined* features, show better results than the regression model trained solely on PeTime, even though the latter correlates more strongly. The combination with this strongly correlated PeTime that was used in the best model of the *combined* feature sets then naturally improves performance compared to the models of *sensor*-based features. Note however, that Spearman correlations can only capture monotonic correlations, thus more complex, e.g. bell-shaped, or even concave relationships cannot be analyzed using this method.

In practice, of course, one should consider what modalities are available and feasible and stack these to achieve better accuracy. Freelance translators probably do not have eye-tracking devices at home; however, as smart watches and fitness trackers are becoming more and more common, an integration of CL detections based on skin and heart data gathered through such devices could be a good and simple addition to CAT tools. Translation companies with fixed workstations might even consider investing in consumer eye trackers like the one used in this study, as the eye features seemingly perform best in this setting. Naturally these modalities should be combined with the easy-to-integrate keyboard- and time-based features that do not require any additional hardware, to increase the CL detection accuracies further.

4.8 Limitations

The presented results are subject to the following limitations: the data sample is relatively small, since only 10 subjects participated in our study, and the participants were translation master's students and not experts with several years of work experience. Next, while we performed cross-validation 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 and cannot expect the existing model to work immediately. Furthermore, the choice of sentences, upon which our two test participants roughly agreed, might lead to different results than evaluating the approach in-the-wild. Moreover, our prediction approach is rather indirect: using sensor measurements, we predict the subjectively assessed CL, which we assume to be a good proxy for actual CL based on the literature. While the rating scale used has been utilized in a large variety of experiments, participants may still have had different interpretations of the scale's labels that might have biased the results. One should also note that our eye tracker only samples at 90 Hz (as opposed to 240 Hz), which could affect the peak velocity reconstruction and thereby saccades (Mack et al. 2017). Last, due to the high variability across subjects, mixed effect regression models (Demberg and Sayeed 2016) might provide further interesting findings in the future.

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 the related works in the translation domain, 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, several of the implemented physiological and behavioral features, e.g. heart rate variability or eye aspect ratio, have not previously been explored in PE. In our study, PE time correlates strongly with perceived CL; however, text-based features show weaker performance. Among the sensor modalities, eye-based features (in particular the blink amount) show the best results, but combining multiple modalities like those based on the skin, eye, etc. improves results further, showing the advantages of a multi-modal approach. Using such a combination of modalities, we can estimate CL during PE without interrupting the actual process through manual ratings.

Currently, all our features are calculated on all data available per segment, which is sufficient to predict perceived CL after finishing the segment. However, in contrast to the time and text features, almost all of our sensor features are averaged across the raw data within a segment. Therefore, in the future we want to investigate, whether we can detect the level of CL even before finishing a segment, by calculating running averages instead. Furthermore, more detailed investigations of the features, e.g. in terms of data filtering approaches or hyper-parameter tuning, will be investigated to make better use of the available data than the regression approach chosen for this work. The long-term goal is to be able to decrease the perceived CL, and thereby stress and exhaustion, during PE. One approach could be to automatically fine-train MT systems on the user's CL measurements to produce less demanding outputs. Another possibility would be to intervene in the PE process within CAT tools when high loads are detected, e.g. 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.

References

Arshad S, Wang Y, Chen F (2013) Analysing mouse activity for cognitive load detection. In: Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration, ACM, pp 115–118

Asteriadis S, Tzouveli P, Karpouzis K, Kollias S (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

Callison-Burch C, Koehn P, Monz C, Peterson K, Przybocki M, Zaidan OF (2010) Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In: Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics, Association for Computational Linguistics, pp 17–53

Chanel G, Rebetez C, Bétrancourt M, Pun T (2008) Boredom, engagement and anxiety as indicators for adaptation to difficulty in games. In: Proceedings of the 12th International Conference on Entertainment and Media in the Ubiquitous Era, ACM, pp 13–17

Chen F, Zhou J, Wang Y, Yu K, Arshad SZ, Khawaji A, Conway D (2016) Robust Multimodal Cognitive Load Measurement. Springer

Chen S, Epps J (2013) Automatic classification of eye activity for cognitive load measurement with emotion interference. Computer Methods and Programs in Biomedicine 110(2):111–124

Corder GW, Foreman DI (2009) Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach. Wiley

Demberg V, Sayeed A (2016) The frequency of rapid pupil dilations as a measure of linguistic processing difficulty. PloS one 11(1):1–29

Dietterich TG (1998) Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation 10(7):1895–1923

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

Gehring J, Auli M, Grangier D, Yarats D, Dauphin YN (2017) Convolutional sequence to sequence learning. arXiv

Goldberg JH, Kotval XP (1999) Computer interface evaluation using eye movements: Methods and constructs. International Journal of Industrial Ergonomics 24(6):631–645

Guerberof A (2009) Productivity and quality in the post-editing of outputs from translation memories and machine translation. International Journal of Localization 7(1)

Hart SG, Staveland LE (1988) Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: Advances in Psychology, vol 52, Elsevier, pp 139–183

Hockey GRJ (1997) Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. Biological Psychology 45(1):73–93

Hosseini SA, Khalilzadeh MA (2010) Emotional stress recognition system using EEG and psychophysiological signals: Using new labelling process of EEG signals in emotional stress state. In: International Conference on Biomedical Engineering and Computer Science, IEEE, pp 1–6

Iqbal ST, Zheng XS, Bailey BP (2004) Task-evoked pupillary response to mental workload in human-computer interaction. In: Extended Abstracts on Human Factors in Computing Systems, ACM, pp 1477–1480

Kahou SE, Bouthillier X, Lamblin P, Gulcehre C, Michalski V, Konda K, Jean S, Froumenty P, Dauphin Y, Boulanger-Lewandowski N, et al. (2016) Emonets: Multimodal deep learning approaches for emotion recognition in video. Journal on Multimodal User Interfaces 10(2):99–111

Koponen M (2012) Comparing human perceptions of post-editing effort with post-editing operations. In: Proceedings of the Seventh Workshop on Statistical Machine Translation, Association for Computational Linguistics, pp 181–190

Koponen M (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 M, Aziz W, Ramos L, Specia L (2012) Post-editing time as a measure of cognitive effort. In: AMTA Workshop on Post-Editing Technology and Practice, pp 11–20

Kramer AF (1991) Physiological metrics of mental workload: A review of recent progress. Multiple-Task Performance pp 279–328

Krings HP (2001) Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes, vol 5. Kent State University Press
Kruger JL, Doherty S (2016) Measuring cognitive load in the presence of educational video: Towards a multimodal methodology. Australasian Journal of Educational Technology 32(6)
Kruger JL, Doherty S, Fox W, De Lissa P (2018) Multimodal measurement of cognitive load during subtitle processing. Innovation and Expansion in Translation Process Research p 267
Lacruz I, Shreve GM (2014) Pauses and cognitive effort in post-editing. Post-Editing of Machine Translation: Processes and Applications p 246
Lacruz I, Shreve GM, Angelone E (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, pp 21–30
Lafferty JD, McCallum A, Pereira FCN (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning, pp 282–289
Lavie A, Agarwal A (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 CY, Och FJ (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
Mack DJ, Belfanti S, Schwarz U (2017) The effect of sampling rate and lowpass filters on saccades–a modeling approach. Behavior Research Methods 49(6):2146–2162
Mellinger CD (2014) Computer-Assisted Translation: An Empirical Investigation of Cognitive Effort. Kent State University
Moorkens J, O'Brien S, da Silva IA, de Lima Fonseca NB, Alves F (2015) Correlations of perceived post-editing effort with measurements of actual effort. Machine Translation 29(3-4):267–284
Mulder L (1992) Measurement and analysis methods of heart rate and respiration for use in applied environments. Biological Psychology 34(2):205–236
O'Brien S (2005) Methodologies for measuring the correlations between post-editing effort and machine translatability. Machine Translation 19(1):37–58
O'Brien S (2006a) Eye-tracking and translation memory matches. Perspectives: Studies in Translatology 14(3):185–205
O'Brien S (2006b) Pauses as indicators of cognitive effort in post-editing machine translation output. Across Languages and Cultures 7(1):1–21
Paas F, Tuovinen JE, Tabbers H, Van Gerven PW (2003) Cognitive load measurement as a means to advance cognitive load theory. Educational Psychologist 38(1):63–71
Paas FG, Van Merriënboer JJ (1994) Instructional control of cognitive load in the training of complex cognitive tasks. Educational Psychology Review 6(4):351–371
Papineni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the ACL, pp 311–318
Popovic M, Lommel A, Burchardt A, Avramidis E, Uszkoreit H (2014) Relations between different types of post-editing operations, cognitive effort and temporal effort. In: Proceedings of the 17th Annual Conference of the European Association for Machine Translation, pp 191–198
Rowe DW, Sibert J, Irwin D (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, pp 480–487
Sennrich R, Haddow B, Birch A (2016) Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics
Sennrich R, Birch A, Currey A, Germann U, Haddow B, Heafield K, Miceli Barone AV, Williams P (2017) The University of Edinburgh's neural MT systems for WMT17. In: Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers, pp 389–399
Shi Y, Ruiz N, Taib R, Choi E, Chen F (2007) Galvanic skin response (GSR) as an index of cognitive load. In: Extended Abstracts on Human Factors in Computing Systems, pp 2651–2656

Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J (2006) A study of translation edit rate with targeted human annotation. In: Proceedings of the Association for Machine Translation in the Americas, pp 223–231

Snover M, Madnani N, Dorr B, Schwartz R (2009) Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In: Proceedings of the 4th Workshop on Statistical Machine Translation, pp 259–268

Solovey E, Schermerhorn P, Scheutz M, Sassaroli A, Fantini S, Jacob R (2012) Brainput: Enhancing interactive systems with streaming fNIRS brain input. In: Proceedings of the Conference on Human Factors in Computing Systems, ACM, pp 2193–2202

Soukupova T, Cech J (2016) Real-time eye blink detection using facial landmarks. In: 21st Computer Vision Winter Workshop, pp 1–8

Specia L, Raj D, Turchi M (2010) Machine translation evaluation versus quality estimation. Machine Translation 24(1):39–50

Stuyven E, Van der Goten K, Vandierendonck A, Claeys K, Crevits L (2000) The effect of cognitive load on saccadic eye movements. Acta Psychologica 104(1):69–85

Sweller J (1988) Cognitive load during problem solving: Effects on learning. Cognitive Science 12(2):257–285

Sweller J, Van Merrienboer JJ, Paas FG (1998) Cognitive architecture and instructional design. Educational Psychology Review 10(3):251–296

Tatsumi M (2009) Correlation between automatic evaluation metric scores, post-editing speed, and some other factors. The Twelfth Machine Translation Summit pp 332–339

Temnikova IP (2010) Cognitive evaluation approach for a controlled language post–editing experiment. In: Proceedings of the International Conference on Language Resources and Evaluation

Van Orden KF, Limbert W, Makeig S, Jung TP (2001) Eye activity correlates of workload during a visuospatial memory task. Human Factors 43(1):111–121

Vieira LN (2014) Indices of cognitive effort in machine translation post-editing. Machine Translation 28(3-4):187–216

Vieira LN (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 MV, Zapirain BG, Zorrilla AM (2012) A stress sensor based on galvanic skin response (GSR) controlled by ZigBee. Sensors 12(5):6075–6101

Yamakoshi T, Yamakoshi K, Tanaka S, Nogawa M, Park SB, Shibata M, Sawada Y, Rolfe P, Hirose Y (2008) Feasibility study on driver's stress detection from differential skin temperature measurement. In: Engineering in Medicine and Biology Society, IEEE, pp 1076–1079

Zampieri M, Vela M (2014) Quantifying the influence of MT output in the translators' performance: A case study in technical translation. In: Proceedings of the EACL Workshop on Humans and Computer-assisted Translation, pp 93–98