Chapter 1

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

Nico Herbig[a,b,c], Santanu Pal[a,b,c], Antonio Krüger[a,b,c] & Josef van Genabith[a,b,c]

[a]German Research Center for Artificial Intelligence (DFKI) [b]Saarland University [c]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.

Nico Herbig, Santanu Pal, Antonio Krüger & Josef van Genabith. 2021. Multi-modal estimation of cognitive load in post-editing of machine translation. In Tra&Co Group (ed.), *Translation, interpreting, cognition: The way out of the box*, 1–32. Berlin: Language Science Press. DOI: 10.5281/zenodo.4545029

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[1] 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.

[1] 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$^{\text{wave}}$), 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$^{\text{count}}$). 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[2] to calculate higher level skin conductance features on the Empatica raw data. It provides us with "global" features, namely the mean value (Leda_{avg}) and the maximum positive deflection ($\text{Leda}_{\text{MaxDefl}}$), and "through-to-peak (TTP)/min-max" analysis, namely the number of significant (i.e. above-threshold) skin conductance responses (SCRs) ($\text{Leda}_{\text{TTP.nSCR}}$), the sum of SCR amplitudes ($\text{Leda}_{\text{TTP.AmpSum}}$) of significant SCRs, and the response latency ($\text{Leda}_{\text{TTP.Lat}}$) 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 ($\text{Leda}_{\text{CDA.nSCR}}$, $\text{Leda}_{\text{CDA.AmpSum}}$, $\text{Leda}_{\text{CDA.Lat}}$). Furthermore, the average phasic driver ($\text{Leda}_{\text{CDA.SCR}}$), the area of phasic driver ($\text{Leda}_{\text{CDA.ISCR}}$), as well as the maximum value of phasic activity ($\text{Leda}_{\text{CDA.PhasMax}}$) and the mean tonic activity ($\text{Leda}_{\text{CDA.Ton}}$) 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.

[2] 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 $\text{GSR}_{\text{range}}$.

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 kΩ 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 (SubjCL$_{\text{avg}}$).

Overall, we compare two approaches for training regression models.

The first approach uses only the above measures to predict 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.[3]

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.[4] 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,

[3] 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.

[4] 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 appraoch", 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 ($\text{mean} = 40.4, \text{SD} = 9.7$). 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 ($\text{mean} = 12.1, \text{SD} = 3$). 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 ($\text{SD} = 2.1, \text{min} = 1, \text{max} = 9$). On a 5-point scale ranging from very bad to very good, they judged their knowledge of CAT tools as good ($\text{mean} = 4.2, \text{SD} = 0.9$), their experience with Trados as good ($\text{mean} = 4.4, \text{SD} = 0.7$), their general knowledge of translation as very good ($\text{mean} = 4.8, \text{SD} = 0.4$), and their PE knowledge as good ($\text{mean} = 3.8, \text{SD} = 1.0$).

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.

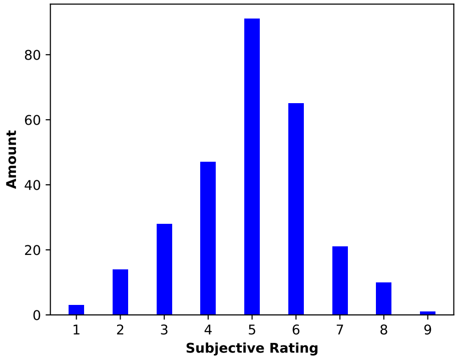


Figure 1.1: Rating distribution across subjective CL scale.

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 SubjCL$_{avg}$ after Bonferroni correction.

		MSE	
	Features	1x10-CV↓(Reg.)	5x2-CV↓ (SD)
Baseline	SubjCL$_{avg}$	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.

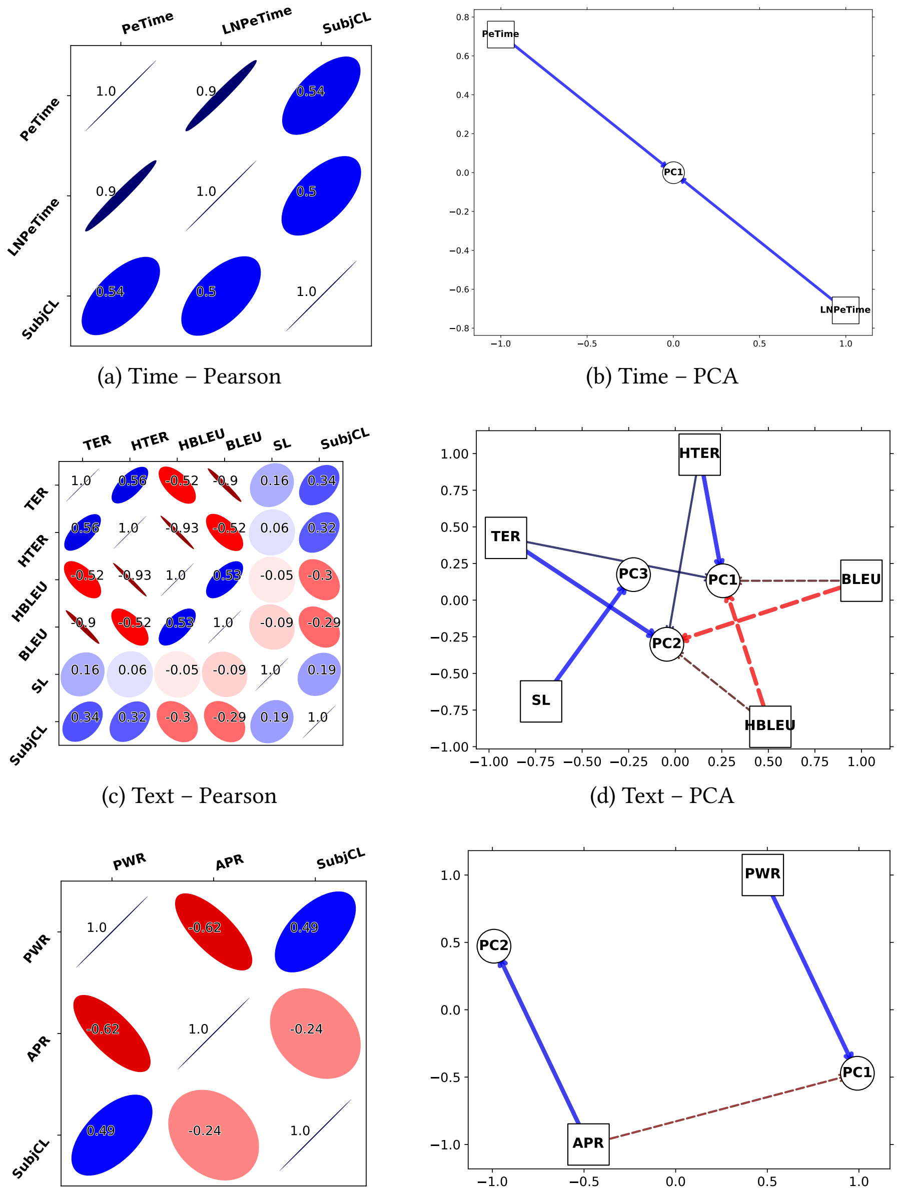
4.3 Pairwise correlations and PCA

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 SubjCL$_{\text{avg}}$ after Bonferroni correction.

		MSE (L: LMEM, R: RF)	
	Features	1x10-CV↓(Reg.)	5x2-CV↓ (SD)
Baseline	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)*



(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.