SAARLAND UNIVERSITY

Faculty of Mathematics and Computer Science
Department of Computer Science
Dissertation

Multi-Modal Post-Editing of Machine Translation

Dissertation zur Erlangung des Grades des
Doktors der Ingenieurwissenschaften (Dr.-Ing.)
der Fakultät für Mathematik und Informatik
der Universität des Saarlandes

vorgelegt von
(M.Sc.) Nico Herbig
Saarbrücken
2022

Date of the colloquium:	May 12, 2022
Dean:	Professor Dr. Jürgen Steimle
Chairman of the examination board:	Professor Dr. Jürgen Steimle
Reporter:	Professor Dr. Antonio Krüger
	Professor Dr. Josef van Genabith
	Professor Dr. Antonio Toral
Scientific Assistant:	Dr. Florian Daiber

Notes on style:
Most of the work presented in this dissertation was done in collaboration with other researchers or students, therefore, the scientific plural “we” is used. Furthermore, the pronouns “she” and “he”, as well as the possesive pronouns “her” and “his”, are used interchangeably and are not intended to reveal the gender of participants. All hyperlinks provided in this dissertation were last accessed on January 29, 2022.

Acknowledgements

There are many people to whom I am truly thankful, and without whom this thesis would not have been possible. If I forgot anyone, please accept my apology.

First of all, I want to thank both Antonio Krüger and Josef van Genabith for the opportunity to work on this topic, for the valuable feedback and guidance throughout the years, not just regarding the investigated topics, but also as mentors. I also want to thank Antonio Toral for reviewing this dissertation.

Furthermore, I want to thank the German Research Foundation (DFG) and the German Research Center for Artificial Intelligence (DFKI) for funding me throughout my dissertation. I also want to thank the Saarbrücken Graduate School of Computer Science for supporting my work and Margaret De Lap for proofreading my publications.

I really enjoyed the working atmosphere we had at our chair, and want to thank all colleagues for it! Now, in Corona-times, I often look back sentimentally and miss all the fun we had on our floor and when traveling to conferences or seminars. Thanks to all of you also for the fruitful discussions, feedback, and ideas, which helped me shape my work. In particular, I want to thank André Zenner and Santanu Pal. André, who was always there since the first semester of our Bachelor's until now at the end of our PhDs. Even though we nowadays work in slightly different areas, you always gave me valuable advice and feedback. And Santanu, who was my strongest collaborator regarding this dissertation. Thanks for helping me get started in the area of machine translation, and for working together through the nights before our deadlines. I further want to thank all research assistants and Master's students who supported my work throughout the years, in particular, Tim Düwel, Mossad Helali, Raksha Shenoy, Rashad Albo Jamara, and Lea Eckhart.

I also want to thank my family for their perpetual support throughout my whole life, which brought me where I am now. A big thank also to my friends, for your aid and support. Last but not least, I want to thank Nina for being there all these years, for supporting and understanding me. Thank you.

Abstract

As Machine Translation (MT) quality continues to improve, more and more translators switch from traditional translation from scratch to Post-Editing (PE) of MT output, which has been shown to save time and reduce errors. Instead of mainly generating text, translators are now asked to correct errors within otherwise helpful translation proposals, where repetitive MT errors make the process tiresome, while hard-to-spot errors make PE a cognitively demanding activity. Our contribution is three-fold: first, we explore whether interaction modalities other than mouse and keyboard could well support PE by creating and testing the MMPE translation environment. MMPE allows translators to cross out or hand-write text, drag and drop words for reordering, use spoken commands or hand gestures to manipulate text, or to combine any of these input modalities. Second, our interviews revealed that translators see value in automatically receiving additional translation support when a high Cognitive Load (CL) is detected during PE. We therefore developed a sensor framework using a wide range of physiological and behavioral data to estimate perceived CL and tested it in three studies, showing that multi-modal, eye, heart, and skin measures can be used to make translation environments cognition-aware. Third, we present two multi-encoder Transformer architectures for Automatic Post-Editing (APE) and discuss how these can adapt MT output to a domain and thereby avoid correcting repetitive MT errors.

Zusammenfassung

Angesichts der stetig steigenden Qualität maschineller Übersetzungssysteme (MÜ) post-editieren (PE) immer mehr Übersetzer die MÜ-Ausgabe, was im Vergleich zur herkömmlichen Übersetzung Zeit spart und Fehler reduziert. Anstatt primär Text zu generieren, müssen Übersetzer nun Fehler in ansonsten hilfreichen Übersetzungsvorschlägen korrigieren. Dennoch bleibt die Arbeit durch wiederkehrende MÜ-Fehler mühsam und schwer zu erkennende Fehler fordern die Übersetzer kognitiv. Wir tragen auf drei Ebenen zur Verbesserung des PE bei: Erstens untersuchen wir, ob andere Interaktionsmodalitäten als Maus und Tastatur das PE unterstützen können, indem wir die Übersetzungsumgebung MMPE entwickeln und testen. MMPE ermöglicht es, Text handschriftlich, per Sprache oder über Handgesten zu verändern, Wörter per Drag & Drop neu anzuordnen oder all diese Eingabemodalitäten zu kombinieren. Zweitens stellen wir ein Sensor-Framework vor, das eine Vielzahl physiologischer und verhaltensbezogener Messwerte verwendet, um die kognitive Last (KL) abzuschätzen. In drei Studien konnten wir zeigen, dass multimodale Messung von Augen-, Herz- und Hautmerkmalen verwendet werden kann, um Übersetzungsumgebungen an die KL der Übersetzer anzupassen. Drittens stellen wir zwei Multi-Encoder-Transformer-Architekturen für das automatische Post-Editieren (APE) vor und erörtern, wie diese die MÜ-Ausgabe an eine Domäne anpassen und dadurch die Korrektur von sich wiederholenden MÜ-Fehlern vermeiden können.

List of Publications

Large sections of this dissertation, including text passages, figures, tables but also the ideas, applications, studies, results, and conclusions have already been published individually. The following list shows these publications clustered by type and states where they appear in the dissertation.

Full conference papers

- **Nico Herbig**, Santanu Pal, Josef van Genabith, and Antonio Krüger. 2019a. Multi-modal approaches for post-editing machine translation. In *Conference on Human Factors in Computing Systems*, pages 1–11. Association for Computing Machinery. (**chapter 3** and **chapter 6**)

- **Nico Herbig**, Tim Düwel, Santanu Pal, Kalliopi Meladaki, Mahsa Monshizadeh, Antonio Krüger, and Josef van Genabith. 2020b. MMPE: a multi-modal interface for post-editing machine translation. In *Annual Meeting of the Association for Computational Linguistics*, pages 1691–1702. Association for Computational Linguistics. (**chapter 4**)

- **Nico Herbig**, Tim Düwel, Mossad Helali, Lea Eckhart, Patrick Schuck, Subhabrata Choudhury, and Antonio Krüger. 2020a. Investigating multi-modal measures for cognitive load detection in e-learning. In *Conference on User Modeling, Adaptation and Personalization*, pages 88–97. Association for Computing Machinery. (**chapter 7** and **section 8.3**)

- Santanu Pal, Hongfei Xu, **Nico Herbig**, Sudip Kumar Naskar, Antonio Krüger, and Josef van Genabith. 2020. The Transference architecture for automatic post-editing. In *International Conference on Computational Linguistics*, pages 5963–5974. The COLING Organizing Committee. (**chapter 11**)

- Rashad Albo Jamara, **Nico Herbig**, Antonio Krüger, and Josef van Genabith. 2021. Mid-air hand gestures for post-editing of machine translation. In *Annual Meeting of the Association for Computational Linguistics*, pages 6763–6773. Association for Computational Linguistics. (**section 5.8**)

- Raksha Shenoy, **Nico Herbig**, Antonio Krüger, and Josef van Genabith. 2021. Investigating the helpfulness of word-level quality estimation for post-editing machine translation output. In *Conference on Empirical Methods in Natural Language Processing*, pages 10173–10185. Association for Computational Linguistics. (**section 5.9**)

Journal articles

- **Nico Herbig**, Santanu Pal, Mihaela Vela, Antonio Krüger, and Josef van Genabith. 2019a. Multi-modal indicators for estimating perceived cognitive load in post-editing of machine translation. *Machine Translation*, 33(1–2):91–115. (**chapter 7** and **section 8.1**)

Book chapters

- **Nico Herbig**, Santanu Pal, Antonio Krüger, and Josef van Genabith. 2021. Multi-modal estimation of cognitive load in post-editing of machine translation. In *Translation, Interpreting, Cognition: The Way Out of the Box*, pages 1–32. Language Science Press. (**chapter 7** and **section 8.2**)

Short conference papers

- **Nico Herbig**, Patrick Schuck, and Antonio Krüger. 2019d. User acceptance of cognition-aware e-learning: an online survey. In *International Conference on Mobile and Ubiquitous Multimedia*, pages 1–6. Association for Computing Machinery. (**chapter 9**)

Demo papers

- **Nico Herbig**, Santanu Pal, Tim Düwel, Kalliopi Meladaki, Mahsa Monshizadeh, Vladislav Hnatovskiy, Antonio Krüger, and Josef van Genabith. 2020c. MMPE: a multi-modal interface using handwriting, touch reordering, and speech commands for post-editing machine translation. In *Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 327–334. Association for Computational Linguistics. (**chapter 4**)

Shared task papers

- Santanu Pal, **Nico Herbig**, Antonio Krüger, and Josef van Genabith. 2018. A transformer-based multi-source automatic post-editing system. In *Conference on Machine Translation*, pages 827–835. Association for Computational Linguistics. (**chapter 10**)
- Santanu Pal, Hongfei Xu, **Nico Herbig**, Antonio Krüger, and Josef van Genabith. 2019. USAAR-DFKI–The Transference architecture for English–German automatic post-editing. In *Conference on Machine Translation*, pages 124–131. Association for Computational Linguistics. (**chapter 11**)

Workshop papers

- **Nico Herbig**, Santanu Pal, Tim Düwel, Raksha Shenoy, Antonio Krüger, and Josef van Genabith. 2020d. Improving the multi-modal post-editing (MMPE) CAT environment based on professional translators' feedback. In *AMTA Workshop on Post-Editing in Modern-Day Translation*, pages 93–108. Association for Machine Translation in the Americas. (**chapter 5**)

ArXiv papers

- **Nico Herbig**, Santanu Pal, Josef van Genabith, and Antonio Krüger. 2019. Integrating artificial and human intelligence for efficient translation. *CoRR*, abs/1903.02978:1–4. (**chapter 1**, **chapter 2**, and **chapter 14**)

Supervised Master's theses

- Rashad Albo Jamara. 2021. Using hand gestures for text editing tasks in post-editing of machine translation. Master's thesis, Saarland University. Advisor: **Nico Herbig**. (**section 5.8**)

- Raksha Shenoy. 2021. ImpoWord MTQE: impact of word-level machine translation quality estimation on post-editing effort. Master's thesis, Saarland University. Advisor: **Nico Herbig**. (**section 5.9**)

- Atika Akmal. 2021. IPE: enhancing visualization of multiple alternatives in interactive post-editing. Master's thesis, Saarland University. Advisor: **Nico Herbig**.

- Shiya Wang. 2021. The study of multi-proposals in post-editing. Master's thesis, Saarland University. Advisor: **Nico Herbig**.

Contents

Part I

Introduction and Background

This initial part introduces the topic of post-editing machine translated text and motivates the research questions addressed in this dissertation (chapter 1). Afterwards, chapter 2 presents the background and related works that are relevant for the remaining parts.

Chapter 1
Introduction

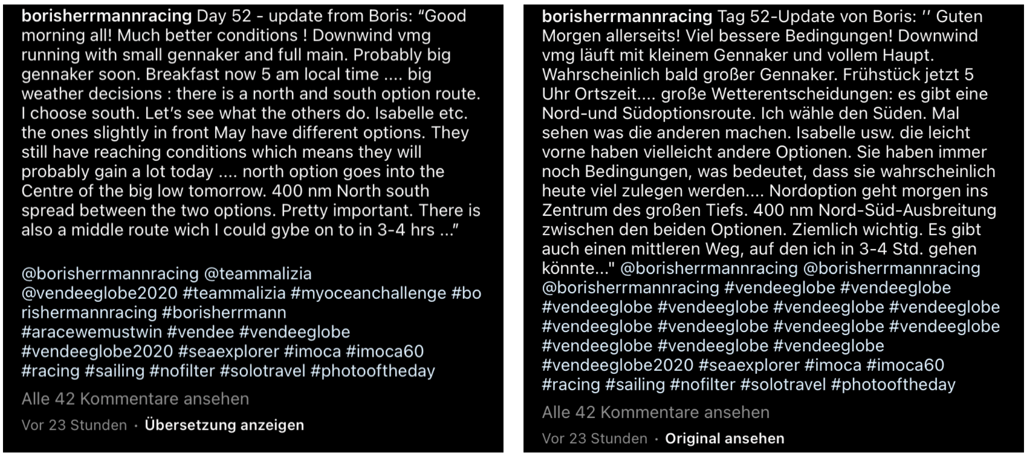
This chapter introduces the task of Post-Editing (PE), which this thesis addresses by considering and conducting research at the intersection of Machine Translation (MT) and Human-Computer Interaction (HCI). We then focus on current problems in PE MT and thereby motivate our work. Afterwards, our main research questions and the approaches used to tackle these are presented. In the end, we provide an overview of the main contributions this thesis makes, and present the thesis structure, linking individual chapters to the research questions.

Parts of this chapter are based on Herbig et al. (2019b).

1.1 Post-Editing at the Intersection of Machine Translation and Human-Computer Interaction

Traditional translation from scratch dates back thousands of years, when tribes had to communicate with each other, first verbally through interpreters, then also in written form through translators. Fully automatic language translation is therefore a long anticipated goal pushed early on through pioneers like Warren Weaver (Weaver, 1953). The science fiction literature also presented a variety of solutions to communicate across languages, e.g., the talking robot C-3PO in Star Wars or the Babel Fish in Douglas Adams' Hitchhikers' Guide to the Galaxy. Even though the problem is far from being solved, Machine Translation generated by computers is nowadays integrated into several end-user applications, ranging from translating posts on social media (see Figure 1.1) to slides in Power Point (see Figure 1.2).

These automatic translations help to handle the the vast amount of text that is being produced every day in an interconnected world; the market size for global



borisherrmannracing Day 52 - update from Boris: "Good morning all! Much better conditions ! Downwind vmg running with small gennaker and full main. Probably big gennaker soon. Breakfast now 5 am local time .... big weather decisions : there is a north and south option route. I choose south. Let's see what the others do. Isabelle etc. the ones slightly in front May have different options. They still have reaching conditions which means they will probably gain a lot today .... north option goes into the Centre of the big low tomorrow. 400 nm North south spread between the two options. Pretty important. There is also a middle route wich I could gybe on to in 3-4 hrs ..."

@borisherrmannracing @teammalizia @vendeeglobe2020 #teammalizia #myoceanchallenge #borishermannracing #borisherrmann #aracewemustwin #vendee #vendeeglobe #vendeeglobe2020 #seaexplorer #imoca #imoca60 #racing #sailing #nofilter #solotravel #photooftheday

Alle 42 Kommentare ansehen

Vor 23 Stunden · **Übersetzung anzeigen**

(a) English original post.

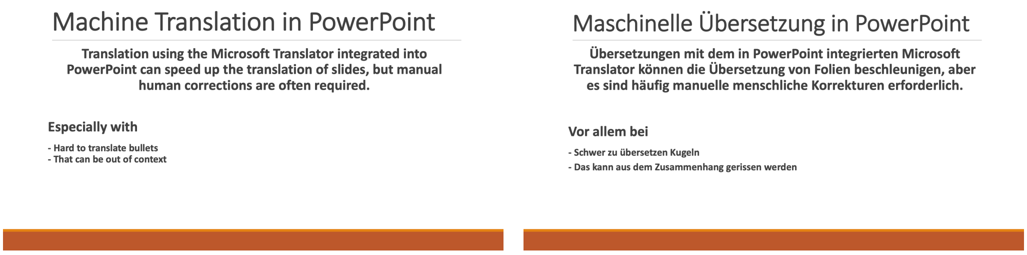
borisherrmannracing Tag 52-Update von Boris: '' Guten Morgen allerseits! Viel bessere Bedingungen! Downwind vmg läuft mit kleinem Gennaker und vollem Haupt. Wahrscheinlich bald großer Gennaker. Frühstück jetzt 5 Uhr Ortszeit.... große Wetterentscheidungen: es gibt eine Nord-und Südoptionsroute. Ich wähle den Süden. Mal sehen was die anderen machen. Isabelle usw. die leicht vorne haben vielleicht andere Optionen. Sie haben immer noch Bedingungen, was bedeutet, dass sie wahrscheinlich heute viel zulegen werden.... Nordoption geht morgen ins Zentrum des großen Tiefs. 400 nm Nord-Süd-Ausbreitung zwischen den beiden Optionen. Ziemlich wichtig. Es gibt auch einen mittleren Weg, auf den ich in 3-4 Std. gehen könnte..." @borisherrmannracing @borisherrmannracing @borisherrmannracing #vendeeglobe #vendeeglobe #vendeeglobe #vendeeglobe #vendeeglobe #vendeeglobe #vendeeglobe #vendeeglobe #vendeeglobe #vendeeglobe #vendeeglobe #vendeeglobe2020 #seaexplorer #imoca #imoca60 #racing #sailing #nofilter #solotravel #photooftheday

Alle 42 Kommentare ansehen

Vor 23 Stunden · **Original ansehen**

(b) German translated post.

Figure 1.1: Machine Translation in social media, here Instagram post by *borisherrmannracing* from 30th of December 2020.



Machine Translation in PowerPoint

Translation using the Microsoft Translator integrated into PowerPoint can speed up the translation of slides, but manual human corrections are often required.

Especially with

- Hard to translate bullets
- That can be out of context

(a) Original slide in English.

Maschinelle Übersetzung in PowerPoint

Übersetzungen mit dem in PowerPoint integrierten Microsoft Translator können die Übersetzung von Folien beschleunigen, aber es sind häufig manuelle menschliche Korrekturen erforderlich.

Vor allem bei

- Schwer zu übersetzen Kugeln
- Das kann aus dem Zusammenhang gerissen werden

(b) Translated slide in German.

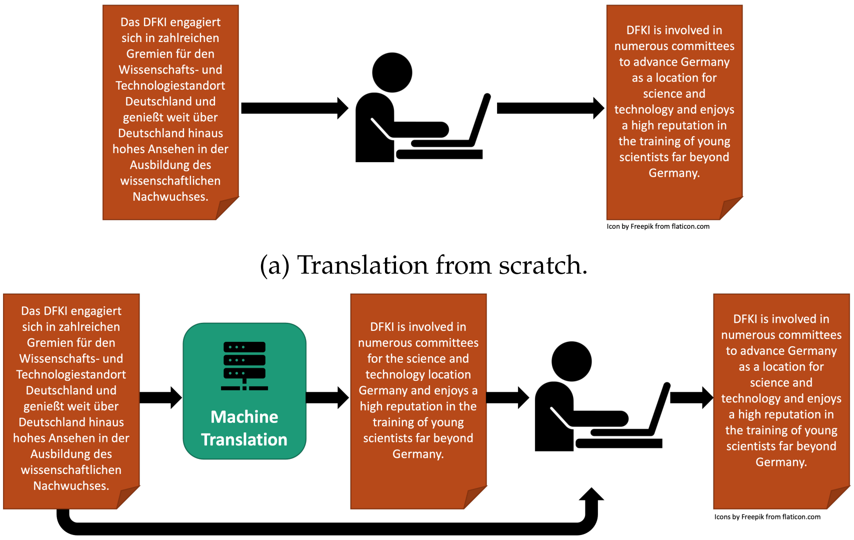
Figure 1.2: Machine Translation in PowerPoint.

language services has doubled in size over the last 10 years[1]. For the social media example, it is often sufficient to grasp the gist of what is being said without understanding every detail, which is a requirement met by most modern-day MT in frequent language pairs. In other cases, e.g., the Power Point example above, a higher level of quality is required and thus humans need to be involved to guarantee output that is of publishable quality. Nevertheless, MT can serve as a starting point that is then "post-edited" towards the final translation.

This Post-Editing process of Machine Translation output is what our work focuses on. In traditional translation, translators gradually create the target text by manually translating the source text. In PE, the source text is first given to a MT system to produce a translation proposal. The human translator then considers both the original source text and the MT's proposal to identify and

[1] From 23.5 billion USD in 2009 to 49.6 billion USD in 2019 according to Statista , see `https://www.statista.com/statistics/257656/size-of-the-global-language-services-market/`

correct mistakes, as well as select, adapt, and recombine useful text fragments. Figure 1.3 shows a comparison between traditional translation from scratch and PE. While the strong improvements in MT have only been achieved in recent years, the concept of PE dates back long time, with books being published on the topic as early as 1982 (Lawson, 1982).


Das DFKI engagiert sich in zahlreichen Gremien für den Wissenschafts- und Technologiestandort Deutschland und genießt weit über Deutschland hinaus hohes Ansehen in der Ausbildung des wissenschaftlichen Nachwuchses.
DFKI is involved in numerous committees to advance Germany as a location for science and technology and enjoys a high reputation in the training of young scientists far beyond Germany.
Icon by Freepik from flaticon.com
Das DFKI engagiert sich in zahlreichen Gremien für den Wissenschafts- und Technologiestandort Deutschland und genießt weit über Deutschland hinaus hohes Ansehen in der Ausbildung des wissenschaftlichen Nachwuchses.
Machine Translation
DFKI is involved in numerous committees for the science and technology location Germany and enjoys a high reputation in the training of young scientists far beyond Germany.
DFKI is involved in numerous committees to advance Germany as a location for science and technology and enjoys a high reputation in the training of young scientists far beyond Germany.
Icons by Freepik from flaticon.com

(a) Translation from scratch.

(b) PE of MT.

Figure 1.3: Traditional translation from scratch vs. PE of MT (icons form Freepik by flaticons.com).

The motivation for PE is that editing the potentially erroneous output of translation technologies may substantially boost productivity compared to a manual translation from scratch and that it can even reduce errors (Green et al., 2013). PE thus combines the advantages of both Artificial Intelligence (AI) and human intelligence: While the AI is good at quickly proposing draft translations of nowadays often high quality, a human with high proficiency in source and target language is needed to ensure that the meaning is presented in translation, to analyze lexical and semantic nuances, and to understand the segment of text in a large (con)text, including the target audience, their cultural background etc. That way, the human can be inspired by the machine, but uses their expert knowledge in combination with human creativity to improve the MT proposal. Thus, the task changes from mostly text production to comparing and adapting machine-generated proposals, or put differently, from control to supervision. As it will probably take a long time until machines are able to solve translation for every possible domain and every language pair without the help of humans, researching better synergies between human and AI translators is very important.

Fruitful human-machine collaboration in translation was described by Martin Kay, who wrote a visionary paper on “The Proper Place of Men and Machines in Language Translation” already back in 1980 (later republished in Kay (1997)):

> "A computer is a device that can be used to magnify human productivity. Properly used, it does not dehumanize by imposing its own Orwellian stamp on the products of the human spirit and the dignity of human labor but, by taking over what is mechanical and routine, it frees human beings for what is essentially human. Translation is a fine and exacting art, but there is much about it that is mechanical and routine and, if this were given over to a machine, the productivity of the translator would not only be magnified but his work would become more rewarding, more exciting, more human."

1.2 Motivation & Problem Statement

Kay's vision of human-machine collaboration has become a reality in many regards, yielding productivity gains in several translation domains and across levels of expertise (Aranberri et al., 2014; Läubli et al., 2013; Zampieri and Vela, 2014a), and also reducing errors (Green et al., 2013). According to a large-scale study involving more than 7000 translators and interpreters conducted by Common Sense Advisory (CSA), a major translation think tank, around 34% of translators offer PE of MT (Pielmeier and O'Mara, 2010).

Even though PE becomes more widespread mainly due to productivity gains, its popularity among translators is often rather low, as instead of Kay's suggestion of work becoming "more rewarding" as the machine is taking over "mechanical and routine" tasks, translators, who used to feel like creative writers now feel degraded to "assembly line workers" (LeBlanc, 2017), fixing often repetitive errors of the MT (Guerberof-Arenas, 2013; O'Brien and Moorkens, 2014). A reason could be the strong focus of MT research on improving the MT itself that ignores the fact that human translators still have to work with the output of systems as long as fully reliable automatic translation is not possible. This mismatch was also outlined by Vieira and Specia (2011), stating that while MT research is growing, the "integration with human translation, however, does not seem to have advanced in the same proportions". One should also note that MT systems are trained to produce output most similar to a provided reference translation, thereby aiming for a perfect translation, which might not necessarily be the translation that is best suited when humans post-edit it.

How can these tedious aspects of PE be mitigated, such that translators can indeed focus on the more exciting and rewarding aspects that Kay (1997) envisioned?

One possibility is to better reflect the PE process and its particularities in so called Computer-Aided Translation (CAT) tools, which are used by most professional translators in their day-to-day work (Van den Bergh et al., 2015). CAT tools minimally show a text window for the source text and one supporting text editing operations for the target, which can initially be either empty (translation from scratch), or populated with a translation proposal (PE). Usually however, CAT tools offer a wide range of additional features supporting the translation process and maximizing the potential of these workbenches is thus one of the priorities for both the research community and industry (Mesa-Lao, 2012).

Naturally, PE not only changes the interaction patterns within CAT environments, but also the cognitive dimension of translation. It requires awareness not only of the sentence in the original language, but also of the error-prone MT output or possibly even multiple such outputs, Translation Memory (TM) matches, the surrounding context, the text's domain and its terminology, the target audience and their cultural background, consistency within translation projects, also with colleagues, etc. "Too much information, like too little information, can lead to confusion, stress and unnecessary effort on the part of readers" (Byrne, 2006). Thus, techniques to capture the cognitive dimension of PE are required, to ensure that the variety of translation aids does support but not overwhelm the human.

Apart from improving the interface itself, one can instead focus on the MT and try to adapt it to a domain, customer, or the translator: Except for actual errors, the MT can perform the same lexical or stylistic choices over and over again, which the current translator might disagree with, thus requiring similar modifications throughout the text and inducing a feeling of assembly line work. As retraining a full MT engine requires lots of data, better approaches to learn from post-edits might improve the overall experience.

1.3 Research Questions

The overarching research hypothesis that this thesis addresses can be summarized as follows:

> **The PE process can be better supported by moving away from conventional mouse and keyboard-based interaction towards a novel multi-modal CAT environment, which further considers Cognitive Load (CL) and learns from previous post-edits to avoid correcting repetitive mistakes, thereby making a complex and cognitively challenging task easier.**

To explore this hypothesis, we divide it into the following research questions:

RQ1 Can (combinations of) novel interaction modalities like touch, digital pen, or speech enhance common PE operations like insertions, deletions, replacements, or reorderings?

RQ2 Can multi-modal combinations of sensing devices be used to better estimate CL during PE, with the overall goal of reducing the cognitive demand imposed on translators?

RQ3 Can we learn from post-edits to adapt the MT output to certain domains or translators, such that subsequent PE becomes quicker and less repetitive?

Since translators in PE do not produce as much text as they did in translation from scratch, but instead identify and correct errors in MT output, the question arises whether traditional mouse and keyboard input is still the most suitable interaction modality. Thus, **RQ1** explores if PE can be improved through the use of corrective speech commands, handwriting as done in copy editing, finger touch as known from tablets, or combinations thereof. As certain operations in PE like deletions or reorderings might be better supported by some modalities than by others, we further investigate in which scenarios which modality can provide a suitable alternative to mouse and keyboard. This RQ is approached in a user-centered fashion, initially asking professional translators how they would envision such a multi-modal environment, then implementing a prototype based on this initial requirements analysis and testing it with professional translators. Finally, the feedback from the prototype evaluation was used to further improve the system.

Apart from *explicit* multi-modal input to edit MT output (RQ1), **RQ2** addresses *implicit* multi-modal sensor input to better model CL of translators during PE. Here, combinations of eye, skin, heart, typing, and other indicators are used to explore how closely CL during PE can be predicted. This forms the basis of cognition-aware CAT environments that consider the user's current state. Furthermore, better detection of CL can allow a more accurate understanding of the drivers of CL during PE.

Finally, **RQ3** addresses the problem that MT systems only learn input-output mappings between source text and final translations, thereby ignoring the manual effort that post-editors apply to improve MT output. Training full MT systems not only requires a lot of time, but especially data, so adding a few post-edited segments and retraining the MT usually does not sway the weights in a big MT model, such that it keeps making the same mistakes or wrong stylistic choices, leading to repetitive mistakes and a feeling of assembly line work. In RQ3, the concept of Automatic Post-Editing (APE) is therefore addressed. APE aims to learn repetitive mistakes or paraphrasing from post-edits, to then automatically correct such errors of the MT before showing the output to the human translator, who should then be free to focus on the more exciting and rewarding parts of PE.

1.4 Approach & Methods

For exploring whether explicit multi-modal input facilitates common PE operations (RQ1), we start our research with an elicitation study (Vatavu and Wobbrock, 2015), asking professional translators which interaction modalities they believe to be suitable for which PE tasks. Since elicitation studies do not bias participants by any pre-defined concepts, mock-ups, or prototypes, it was shown that these kinds of studies lead to natural interfaces with a high level of immediate usage (Wobbrock et al., 2005). We afterwards iteratively prototype the findings of our study into a fully usable CAT tool focusing on multi-modal input, which we realease open-source. In a further experiment with professional translators using our prototype, we test every modality for every PE operation in a structured way to capture timing and ratings, but also conduct semi-structured interviews (Longhurst, 2003) to capture subjective feedback. Thus, RQ1 is addressed with an exploratory, user-centered design process (Vredenburg et al., 2002), involving users early on and continuously learning from their feedback.

For understanding how the cognitive dimension of PE can be better modeled (RQ2), we start by interviewing professional translators how the PE process could be improved by adapting to estimated CL. We then build a sensor framework, combining input from a variety of modalities, that allows capturing potentially relevant features of CL. With it, we conduct three data capturing experiments with different users (students and professionals) and domains (PE and e-learning to analyze transferability). Based on the captured data, predictive models are trained and correlations are investigated to explore to which degree CL can be estimated with our framework. Finally, we conduct a user acceptance study, asking potential users of such cognition-aware systems how big assumed benefits would need to be in order to share data of certain sensors (Acquisti et al., 2013).

For RQ3, we use publicly available data of post-edits and explore how these can be best used to improve MT output. As designing such APE systems requires many reiterations and explorations on thousands of sentences, a human evaluation within the design process is unfortunately not feasible. Therefore, we rely on automatic quality metrics. Our evaluation metric strictly follows that of a shared task including publicly available datasets (Chatterjee et al., 2018), thereby allowing a fair comparison to related works. We further discuss the implications of having APE systems in real PE settings by surveying the APE literature.

1.5 Contributions

This thesis provides contributions at the intersection of Human-Computer Interaction (HCI) and Natural Language Processing (NLP), and aims to combine recent advances in both fields to push the boundaries how PE is nowadays conducted. Our contributions can be summarized into three sub-fields: explicit multi-modal input for PE, modeling the cognitive dimension of PE through implicit multi-modal sensor input, and learning from post-edits in the form of APE, to reduce manual effort and especially repetitive corrections.

Regarding *explicit multi-modal input for PE*, we explore the design space of modalities which are intuitively appropriate for which operation, and how the operations (e.g., reorder) should be performed with the various modalities, by conducting an elicitation study with professional translators. Furthermore, we present findings from an interview on hardware setup and interface design of

CAT tools. Guided by these theoretical contributions, we provide the technical contribution of a prototype implementing a broad range of interaction modalities for PE operations, which was iteratively improved based on user feedback. Finally, a study investigating the different modalities for the different operations in a structured test, provides guidance on which modalities CAT tools should focus on and what the advantages and disadvantages of the different interaction modalities are.

Regarding the *cognitive dimension of PE through implicit multi-modal sensor input*, we contribute a study on professional translator's ideas on how interfaces can adapt to measured CL or how MT systems can be improved based on measured CL. As a technical contribution, we provide a unified framework accessing a broad range of CL measures from different sensor modalities. Our data analysis framework also provides detailed analysis possibilities, including a variety of pre-processing, ML-based CL prediction model training and visualization steps. This framework was used in different stages of development in three studies to explore how well CL can be estimated based on this variety of sensor inputs. Last, we contribute users' perceptions on sharing sensor data for the purpose of CL adaptations, which, in combination with the findings of which sensor modalities work well for the task, can guide practical implementations of CL measurement.

Regarding *learning from post-edits in the form of APE*, we contribute two model architectures combining both source text and MT proposal to reduce errors in the latter, and analyze them on publicly available datasets. We further theoretically discuss how such systems could be applied to support translators by learning from limited amount of data, and adapting to stylistic changes by paraphrasing.

1.6 Outline of this Thesis

The remainder of the thesis is structured as shown in Figure 1.4.

Chapter 2 presents the background and related work and thereby provides a solid foundation for the presented research. First, we explain the different types of Machine Translation and focus on their differences for the PE process. Then we discuss how this MT output can be used for PE, what the PE workflow looks like, how PE impacts effort, time and quality, and what translators think about PE in general. Afterwards, we look at CAT environments, what features they offer, what their limitations are, and especially focus on the degree to which multi-modal input has already been explored in CAT tools. Afterwards, we shift to the topic of user modeling, where we particularly focus on the concept of Cognitive Load and approaches to measure and adapt to CL both within and outside of the translation domain. Finally, we talk about Automatic Post-Editing, its basic assumptions, the different architectures, and to which degree they have already been integrated into CAT tools.

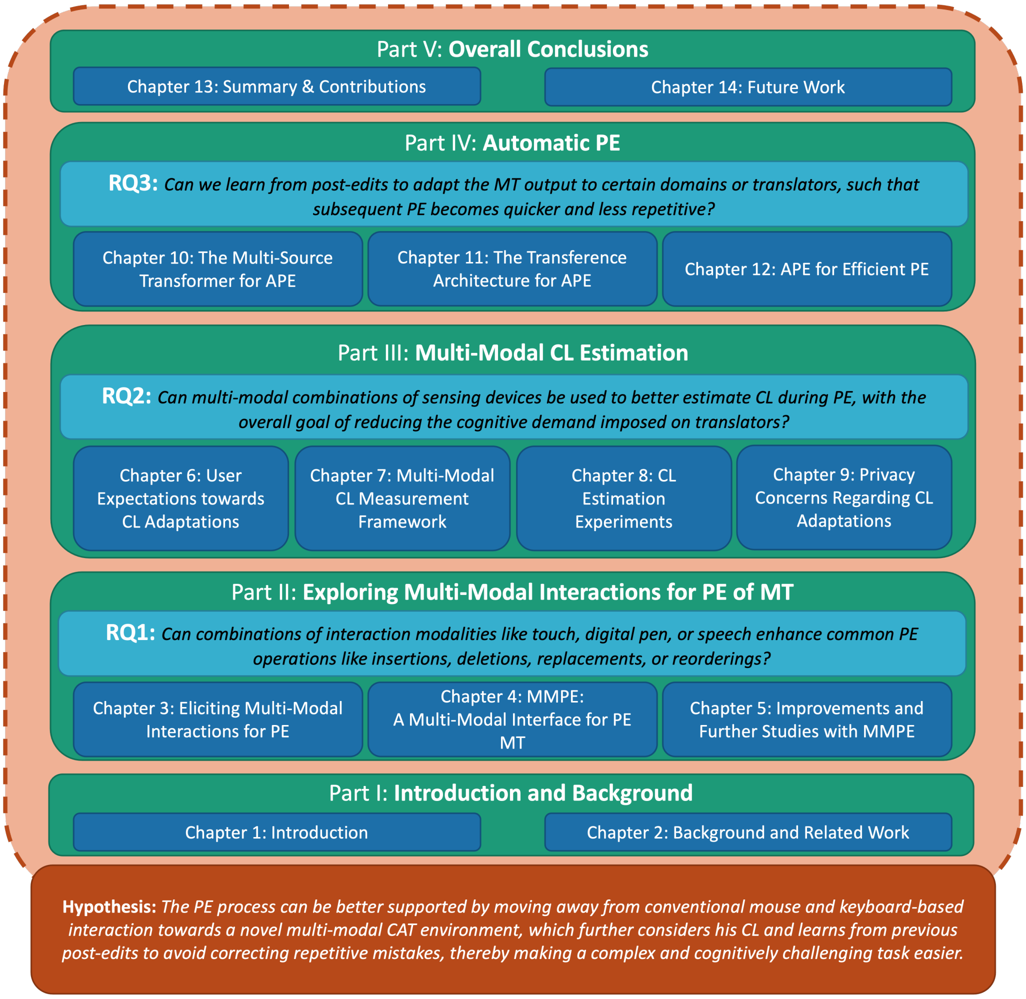

Part V: **Overall Conclusions**
Chapter 13: Summary & Contributions
Chapter 14: Future Work
Part IV: **Automatic PE**
RQ3: *Can we learn from post-edits to adapt the MT output to certain domains or translators, such that subsequent PE becomes quicker and less repetitive?*
Chapter 10: The Multi-Source Transformer for APE
Chapter 11: The Transference Architecture for APE
Chapter 12: APE for Efficient PE
Part III: **Multi-Modal CL Estimation**
RQ2: *Can multi-modal combinations of sensing devices be used to better estimate CL during PE, with the overall goal of reducing the cognitive demand imposed on translators?*
Chapter 6: User Expectations towards CL Adaptations
Chapter 7: Multi-Modal CL Measurement Framework
Chapter 8: CL Estimation Experiments
Chapter 9: Privacy Concerns Regarding CL Adaptations
Part II: **Exploring Multi-Modal Interactions for PE of MT**
RQ1: *Can combinations of interaction modalities like touch, digital pen, or speech enhance common PE operations like insertions, deletions, replacements, or reorderings?*
Chapter 3: Eliciting Multi-Modal Interactions for PE
Chapter 4: MMPE: A Multi-Modal Interface for PE MT
Chapter 5: Improvements and Further Studies with MMPE
Part I: **Introduction and Background**
Chapter 1: Introduction
Chapter 2: Background and Related Work
Hypothesis: *The PE process can be better supported by moving away from conventional mouse and keyboard-based interaction towards a novel multi-modal CAT environment, which further considers his CL and learns from previous post-edits to avoid correcting repetitive mistakes, thereby making a complex and cognitively challenging task easier.*

Figure 1.4: Thesis structure relating chapters and parts to research questions and the research hypothesis.

Part II then focuses on RQ1, investigating how explicit multi-modal interactions can facilitate PE of MT. Chapter 3 starts with an elicitation study, to understand which interactions with which modalities professional translators find intuitive for which tasks. Chapter 4 then presents the prototype built upon these initial findings, and evaluates it with professional translators to ascertain which implemented modality is suitable for which PE operation. Furthermore, a lot of ideas on how the prototype can be improved are discussed, which are then implemented and presented in chapter 5. The same chapter also describes further research conducted with the prototype: first our investigations of mid-air hand gestures during PE, followed by a study on the helpfulness on word-level Quality Estimation (QE) for PE of MT.

Part III focuses on RQ2, the question of how implicit multi-modal sensor input can be leveraged to estimate and adapt to CL. Chapter 6 starts by presenting findings from an interview with professional translators on how CAT tools could be improved based on estimated CL, which sets the overarching goal towards which the other chapters contribute. Chapter 7 then presents the multi-modal CL estimation framework that we have built to conduct studies, which combines a wide set of features from different modalities to understand the users' cognitive states. This framework is then explored in 3 different studies in chapter 8: Starting with a smaller set of CL measures, an initial study is conducted with translation students. After that, the full framework is explored with professional translators, and finally, to also test it in a completely different domain, it is evaluated in the context of e-learning with computer science students. Based on these three studies, we learn how well which CL measures perform, and which degree of accuracy can be reached with a multi-modal approach. Finally, chapter 9 presents an online study to investigate which of the various measures are considered potentially invasive and privacy-critical, and therefore reach higher or lower user acceptance in practical applications.

RQ3 is then addressed in Part IV, where we first present two APE architectures proposed for learning from post-edits to avoid repetitive mistakes (see chapter 10 and chapter 11), and then theoretically discuss how APE can be integrated in the CAT process to really leverage its potential for efficient PE (see chapter 12).

Finally, Part V gives an overall conclusion of this thesis and summarizes the main take-aways. We also discuss various opportunities for future work that arise from the findings of this thesis.

Chapter 2
Background and Related Work

This chapter reviews the literature most relevant to this thesis. We first discuss different types of machine translation and especially focus on overall quality and common errors. Afterwards, the post-editing process is discussed in detail, including the involved effort, achieved time savings, resulting text quality, and translators' perceptions in comparison to traditional translation from scratch. We then review the computer-aided translation domain, showcasing common features and available software tools, and particularly focus on multi-modal input possibilities. After that, we switch towards the concept of cognitive load and corresponding measuring techniques, where we again focus on multi-modal techniques. Finally, automatic approaches to post-editing are reviewed and we discuss how these can be leveraged to avoid repetitive mistakes. We conclude by summarizing the research gaps and discussing how this thesis aims to fill them.

2.1 Machine Translation

"Machine Translation is a sub-field of computational linguistics that aims to automatically translate text from one language to another using a computing device" (Garg and Agarwal, 2019). Even though this definition is very simple, the complexity of natural languages makes implementing a high-quality MT system very challenging. The knowledge required encompasses not only syntax and semantics, but also grammar, culture, context, commonsense knowledge, idioms, etc. in both languages, thus, requiring years of study for humans and making it complex for computational linguists to create high-quality MT systems. After briefly reviewing the history of MT, we discuss the main MT paradigms, present human and automatic MT evaluation methods, and discuss the level of quality achieved by the different MT paradigms.

2.1.1 History

Hutchins (2007) summarizes the history of machine translation and distinguishes two kinds of demands: machine translation for assimilation, and machine translation for dissemination, where the former can be "imperfect, lexically awkward and stylistically crude", while the latter should produce "publishable-quality translation". Whereas early approaches focused on MT for assimilation, later approaches tend to work towards MT for dissemination.

An early pioneer of MT, named Petr Troyanskii, approached the USSR Academy of Sciences already in 1939 and proposed to work on machine translation, however, after years of fruitless discussion the contact was lost again (Hutchins and Lovtskii, 2000). Other early proposals include Warren Weaver's famous memorandum from 1949, where he proposed ideas to move away from limited word-by-word translation schemes (Hutchins, 2000; Weaver, 1999). Three years later, in 1952, Weaver's and others' ideas were discussed at the first MT conference at MIT organized by Yehoshua Bar-Hillel (Hutchins, 2007). Back then, the view was that "full automation of good quality translation was a virtual impossibility, and that human intervention either before or after computer processes (known from the beginning as pre- and post-editing respectively) would be essential" (Hutchins, 2007). The first journal on the topic also appeared in 1953 under the common term "Mechanical Translation" (Hutchins, 2007).

MT research grew and showed some improvements, however, in 1966 a major setback happened: The Automatic Language Processing Advisory Committee (1966) (ALPAC) examined the status quo and found that MT is slower, less accurate, leading to lower comprehension and at the same time being more costly than human translation, concluding that "there is no immediate or predictable prospect of useful machine translation". Instead, the ALPAC recommended to focus on basic research in computational linguistics and to develop machine aids for translators. This ended most of MT research for over a decade (Hutchins, 2007) until the field gained popularity again. In terms of approaches to MT, earlier research focused on Rule-Based Machine Translation (RBMT), followed by example-based approaches in the 1980s, Statistical Machine Translation (SMT) in late 1980s (due to an increase in computational power), and the shift to deep learning after gaining popularity in the 2010s (Garg and Agarwal, 2019). We will discuss these paradigms in more detail in the following sections.

2.1.2 Rule-Based & Example-Based Machine Translation

Rule-Based Machine Translation (RBMT) was the main research focus in the 1970s (Garg and Agarwal, 2019). As the name suggests, RBMT is driven by a set of rules that are used to parse the input text and generate output text. Overall, there are three types of RBMT (Garg and Agarwal, 2019): direct approaches, transfer-based approaches, and interlingual approaches. Direct systems map input to output directly, while transfer based approaches analyze the source regarding morphology and syntax to translate the sentence. Simply speaking, these approaches parse the structure of the input sentence, look up appropriate rules which transform the structure into the target language, then use a dictionary to translate each word, and potentially apply post-processing to adapt word forms etc. The last approach, interlingual RBMT, instead transforms the source into a language-independent intermediate representation, which is then used to generate the target.

RBMT has several advantages: On the one hand, it only requires dictionaries and rules, but no big corpora of parallel text. If rules exist for rare sentences, these can also be handled well, while statistical or neural approaches often have problems with infrequent phenomena. Furthermore, rules can be used independent of the text domain, the source analysis for a certain language is independent of different target languages, and similarly the generation of target sentences from an intermediate representation is independent of different source languages. One can even start supporting new languages by adapting rules from similar languages. Due to the rules, the explainability of RBMT is also very high.

However, the generation of rules requires detailed linguistic analysis and lots of manual effort. It is also complex to write rules that work well for ambiguous or idiomatic expressions, and the stronger the structural differences of languages, the more complicated the creation of rules becomes (Garg and Agarwal, 2019).

Since this is especially true for English and Japanese, Nagao (1984) proposed a different approach to "mechanical translation". The idea arose from the human translation process: Humans do not perform deep linguistic analyses before translation, but instead decompose the sentence into fragments, translate those individually, and compose the translated fragments into an overall sentence. The translation of each fragment can be done by analogy, so the proposed example-based MT can be seen as a form of case-based reasoning, where known solutions to (sub-)problems are reused to solve new problems. Nagao thus proposed to learn word and sentence mappings from data, instead of manually creating and applying rules.

2.1.3 Statistical Machine Translation

Statistical Machine Translation (SMT) takes the idea of learning from corpora instead of curating rules to the next stage. SMT considers the translation problem as a purely statistical problem, where the goal is to find the sentence T in the target language, that has the highest probability of being the translation for a sentence S in the source language, thus maximizing $P(T|S)$. Using Bayes law, this can rewritten as $\underset{T}{\mathrm{argmax}} P(T|S) = \underset{T}{\mathrm{argmax}} P(T) * P(S|T)$ where $P(T)$ is referred to as the language model, while $P(S|T)$ is referred to as the translation model. As argued by Brown et al. (1993), pioneering in the field of SMT, dividing the problem into the inverse (source given target) and language model might seem unintuitive, but has advantages compared to solving $P(T|S)$ directly: $P(S|T)$

can concentrate on well-formed sentences only, as $P(T)$ assigns low probabilities to ill-formed ones, thereby making the outcome of $P(S|T)$ irrelevant for these cases. Thus, applying Bayes law simplifies the task of the translation model $P(S|T)$ compared to $P(T|S)$ but requires an additional language model $P(T)$.

Language models need only output the probability of a word sequence, which can be broken down to the probability of a certain word occurring after previous words. When fixing the size of the considered context, language models can thus be as simple as counting how often a word appears after a certain sequence to get a maximum likelihood estimate (Garg and Agarwal, 2019), however, nowadays much better approaches to language modeling using Transformer-based neural networks like BERT (Devlin et al., 2019) have been proposed. Even when using language models to focus on well-formed sentences, the search space remains huge, thus, SMT approaches can leverage the source text and heuristics to guide the search, with the usual trade-off that a wider exploration of the search space costs time but can improve the quality.

SMT started with word-based approaches, but later shifted to phrase-based approaches (Garg and Agarwal, 2019): Word-based approaches like the IBM models (Brown et al., 1988, 1990, 1993) use individual words as the unit of translation and therefore cannot understand words within their surrounding context, often resulting in a poor lexical choice. Therefore, Och et al. (1999) introduced Phrase-Based Statistical Machine Translation (PBSMT), which uses phrases as the basic translation unit. In contrast to word-based approaches, that rely on simple alignments (i.e., a mapping which source word corresponds to which target word), phrase-based approaches align phrases of M source words to N target words. This allows a translation in context, where each pair of phrases in source and target has a probability of the one being the translation of the other, that can be learned from a large bilingual corpus (Garg and Agarwal, 2019). Until 2016, SMT was still the main paradigm used in the popular Google Translator[2].

Compared to RBMT, SMT does not involve creating and maintaining rules based on linguistic knowledge, but instead learns these patterns automatically from data. Furthermore, the language model ensures that the output sounds rather fluent. However, the dependence on a large corpus is also a disadvantage, as it needs to be created and must match the text domain. Other problems arise from the statistical procedure, e.g., proper names might be overwritten by more likely translations in the training corpus.

2.1.4 Neural Machine Translation

One of the first papers proposing to solely rely on neural networks without any component from SMT was Kalchbrenner and Blunsom (2013). In the following years, the percentage of Neural Machine Translation (NMT) approaches

[2] https://www.blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/

submitted to the yearly shared task hosted by the Conference/Workshop on Machine Translation (WMT) increased until becoming the de-facto standard. Most approaches follow a so-called encoder-decoder architecture, where in early RNN-based versions of this (see below), the encoder reads each input token one word at a time to produce an intermediate representation, which is then used by the decoder to generate the target translation.

For neural networks to work well, words need to be encoded into vector space. To avoid the curse of dimensionality and very sparse representations that would occur when simply encoding each word in the vocabulary as a one-hot vector, the idea arose to model all words in a smaller dimensional but dense vector space, and learning to group similar words next to each other (Bengio et al., 2003; Mikolov et al., 2013). The grouping of similar words within such word embeddings help to predict suitable translations for so far unseen data that is similar to other data in the embedding space. Word embeddings also allow vector arithmetic: A common example depicted in Figure 2.1 is that the high-dimensional vector space can also encode relationships like gender or royalty, thus, knowing how man relates to $king$ (royalty) and man relates to $women$ (gender), one can calculate the word representation for $queen$ $(women + (king - man))$. Nowadays, neural machine translation models are often trained on sub-word level to avoid out-of-vocabulary words and learn e.g., what the stem and ending of words are, thereby also reducing the vocabulary size due to reusability, and thus, improving MT quality for rare words (Sennrich et al., 2016).

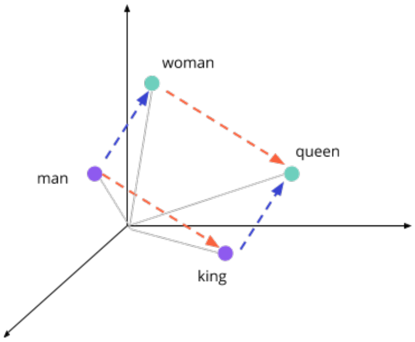


Figure 2.1: Vector arithmetic on word embeddings (taken from Unbabel[3]).

Originally, Recurrent Neural Networks (RNNs), often using Long-Short Term Memory (LSTM) layers, were used for NMT, as they can memorize important parts of the sentence that are required to correctly translate later tokens, for example, for coreference resolution. However, for long sentences, this memory functionality often did not work as anticipated, leading to a reduced accuracy for longer sentences. A reason for the decreased performance is that the encoder-decoder architecture requires all information to be compressed in an intermediate representation of fixed size, which was shown to reduce the performance for longer sentences (Cho et al., 2014). To solve this, Bahdanau et al. (2014) propose

[3] https://resources.unbabel.com/blog/ai-talking-understanding

to introduce so-called attention mechanism, by mapping the source sentence into a sequence of vectors instead of a fixed size vector. The decoder can then "attend" to certain parts of this sequence while decoding, thus, it can learn that some source words are more relevant to the generation of the current target word than others, thereby facilitating longer sentences.

Later on, Gehring et al. (2017) proposed fully Convolutional Neural Networks (CNN), relying on gated linear units (Dauphin et al., 2017) and residual connections (He et al., 2016a), with a separate attention module per decoder layer, and achieved the new state-of-the-art in MT. In contrast to RNNs, CNNs do not need to be trained sequentially from beginning to end, allowing better parallelization and therefore faster training or the use of much more training data. The earlier layers in CNNs capture interactions of nearby words, while later layers can also capture long-range dependencies.

As in RNNs, this CNN model also used attention mechanisms only to help the model focus on certain parts of the input. Shortly thereafter, Vaswani et al. (2017) showed that indeed "attention is all you need", thus, that attention mechanisms are powerful in themselves. The proposed "Transformer" architecture has the advantage of faster training due to parallelization and performs particularly well at modelling dependencies without regard to their distance in the input and output sequences. The encoder and decoder both use multi-head self-attention to compute representations of their inputs, and also compute multi-head vanilla-attentions between encoder and decoder representations. Since then, many NLP problems have been addressed using Transformer-based architectures, e.g., in pretrained systems like BERT (Devlin et al., 2019) or GPT-3 (Brown et al., 2020).

Compared to SMT, NMT models are trained end-to-end and do not rely on a variety of separately built components. There is also no separation of the problem into language model and translation model. Models trained on one domain can further be fine-tuned towards other domains using much less data (Luong and Manning, 2015), thus, showing a degree of generalization. Even though the training times are long, NMT requires much less memory than SMT.

A problem with the use of fixed-size vocabularies in NMT is that rare tokens might not be covered, leading to the creation of "unknown" tokens in the translation. The phrase table in SMT systems also memorizes and translates such rare words. However, as discussed above, approaches like Byte-Pair Encoding (BPE) (Sennrich et al., 2016) operate on the sub-word level to reduce or avoid out-of-vocabulary problems.

2.1.5 Decoding

As in many other NLP tasks that require text generation (e.g., text summarization), a decoding process is required to generate a target sequence for an input sequence. Given the whole source sequence and a partly generated target sequence, the MT model generates a probability distribution for the next token of the target sequence. The decoding process now determines how these probability distributions given by the MT model are used to generate the overall translation result for a source sequence. With thousands of possible tokens at each decoding step (depending on the vocabulary size), the search problem is exponential in the length of the sequence, making it an NP-complete problem (Knight, 1999). Since exhaustive search for such problems is unfeasible, heuristics need to be employed to get high quality sequences even if only small parts of the search space are explored.

The most simple decoding approach that can be used is greedy search, where at each step in the process the most likely token is chosen. Naturally this approach is extremely fast, but at the same time it cannot find sequences that rely on a temporarily worse choice to get an overall better score. Consider an exemplary 3-token scenario, where the highest probability for the first token is 0.5, followed by a highest probability of 0.7 for the second and third token. Overall, this sequence achieves a likelihood of $0.5 * 0.7 * 0.7 = 0.245$. Whereas choosing another initially sub-optimal token with probability 0.4 might lead to the next most likely words having probability 0.8 and 0.9, giving us an overall likelihood of $0.4 * 0.8 * 0.9 = 0.288$.

To tackle this issue, a frequently used heuristic is beam search (Koehn et al., 2003), which expands all possible next tokens in a breadth-first fashion and keeps the k most likely ones in each step. The beam size (or width) k defines how many traces through the graph are explored in parallel: bigger values lead to better results at the cost of decoding speed. Note that beam search with beam size 1 is identical to greedy search. In practice, values around 5 are commonly chosen as this turned out to be a good trade-off between quality and speed.

Naturally, decoding is a whole research field of its own, where many popular search algorithms like A* (Och et al., 2001) can be deployed that make different pruning decisions to restrict the search space. An interesting example is diverse beam search (Vijayakumar et al., 2016), which adds a diversity penalty to avoid too similar sequences within the beam size, thereby generating more diverse outputs than pure standard beam search decoding.

2.1.6 Evaluating Machine Translation Quality

Now that we have seen different approaches to MT, this section focuses on ways to judge if a MT output is of good quality. Traditionally, MT systems were evaluated by humans, which is why we start our overview with human evaluation methods. However, this process is expensive and time-consuming. Therefore, several automatic evaluation schemes have been proposed that correlate highly with human evaluations, but can be run cheaply and efficiently, thereby allowing model tuning.

Human Evaluation

Traditional human evaluation of MT as conducted in ALPAC (1966) comprises *intelligibility* and *fidelity* on 9 point scales. *Intelligibility* here refers to the automatic translation being understandable and reading like normal target language, which is judged mono-lingually in the target language. *Fidelity* on the other hand is assessed with access to the source and captures whether the translation really retains the meaning of the source without distortion. Back then, ALPAC used a direct assessment in the form of a rating scale for intelligibility, but used an indirect approach for fidelity where raters saw the source only after having seen the translation, and were asked to rate how informative the source was. The more informative the source was, the less of the meaning was already captured in the target. To compensate inter-rater variation, evaluation schemes often employ multiple raters.

Human evaluation in the 1990s by the Advanced Research Projects Agency (ARPA) was instead focusing on the measures *adequacy*, *fluency*, and *comprehension* (Han et al., 2016). *Adequacy* here is similar to the traditional fidelity metric and captures how much information from the source is retained and is therefore measured with access to the source on a simple 5 point rating scale. However, in contrast to the ALPAC approach, it is measured directly. *Fluency* on the other hand is comparable to the original intelligibility metric and captures if the translation is a well-formed target language sentence, again rated on a 5 point scale. *Comprehension* captures if the reader understands the conveyed information and is measured by having raters reply to 6 questions with 6 answer possibilities that were created with the reference translation.

Direct assessment, as proposed by Graham et al. (2017), asks raters to rate the adequacy on a scale from 0 to 100, where raters have a slider and can only see the anchors on the left and right. It can be either by ranking the adequacy in comparison to a human reference translation, or by asking for how adequate the sentence is to the source (requiring proficiency in both languages). This approach is nowadays frequently used in the annual Conference/Workshop on Machine Translation (WMT).

Various other scales have also been proposed, but what they have in common is that the MT proposal is either judged on its own, or compared to the source or a reference translation (Läubli et al., 2020). To score the quality, humans either rate the MT proposal or post-edit it according to some criteria.

Apart from these rating approaches, the different MT outputs can also be compared to one another based on segment ranking. In earlier versions of WMT, a complete ranking of the different MT outputs was conducted for each source sentence (Callison-Burch et al., 2011). Later evaluation schemes instead display the source and reference along only 5 MT outputs that need to be ranked, thus, yielding 10 pairwise comparisons (if there are no ties) (Bojar et al., 2014). An advantage of such relative ranking based approaches compared to direct ratings is that the inter- and intra-annotator agreement is much higher (Callison-Burch et al., 2007). However, relative ranking does not provide any information on the magnitude of the differences between two systems (Läubli et al., 2020), which is why direct assessment remains a widespread approach that was also used by later versions of WMT.

Automatic Evaluation

While humans are the target users of MT output and therefor the gold-standard in judging MT output, doing a proper MT quality evaluation with humans is very time-consuming and expensive. Especially when training an MT system, to tune parameters one would need feedback on the quality of thousands of validation (and later test) sentences, which is not feasible with human translators. Therefore, a whole research field in MT aims to design automatic metrics that highly correlate with human judgments. These automatic measures can usually be employed either at sentence level (comparing the MT output with a (set of) reference translation(s)), or at the corpus level (doing the same across a whole corpus), where the latter usually gives more reliable results. Most of the metrics rely on string comparison between the MT output and one or several known reference translations, e.g., word overlap or the edits required to transform one into the other. Working on a string level implies focusing on lexical similarity only, however, some more advanced metrics also consider linguistic features like named entities, paraphrasing, or part of speech (Han et al., 2016).

Banerjee and Lavie (2005) highlight a set of requirements for automatic MT evaluation metrics: high correlation to human evaluation, sensitive to differences in MT quality, "consistent (same MT system on similar texts should produce similar scores), reliable (MT systems that score similarly can be trusted to perform similarly) and general (applicable to different MT tasks in a wide range of domains and scenarios)".

In the following, we will outline a few of the common automatic evaluation metrics, which compare an MT hypothesis with (an) externally provided reference translation(s), some of which will be used throughout this thesis.

Word Error Rate (WER) Su et al. (1992) present what is later called Word Error Rate (WER), which computes the minimal number of edits required to transform the MT output into a reference translation. As edits, the work considers deletions, insertions, and replacements, and uses a dynamic programming approach to efficiently compute their metric. One should however note that reorder is not explicitly considered, but would appear as a delete operation followed by an insert, thus, making word order errors especially expensive.

Position-Independent Word Error Rate (PER) To overcome this problem, Tillmann et al. (1997) introduce the Position-independent word Error Rate (PER), which ignores the word order and simply compares for identical words in both strings. Every non-matched word is considered a replacement, and every additional/missing word is considered a deletion/insertion. Thus, it is guaranteed to be less or equal compared to WER but sentences ordered wrongly do not count as errors.

Translation Edit Rate (TER) Another approach to incorporate word order, that is still widely used today, is the so-called Translation Edit Rate (Snover et al., 2006). It is similar to WER, but adds a movement of single or groups of words as an additional edit operation to replace, insert, and delete. In terms of cost, moving a group of words has equal cost of 1, same as moving a single word, or deleting/inserting/replacing a single word. Extensions of Translation Edit Rate (TER) have also been proposed, a popular one being TER-Plus (TERp) (Snover et al., 2009), which additionally considers stem matches, synonym matches, and phrase substitutions.

BiLingual Evaluation Understudy (BLEU) Instead of viewing the similarity problem as an edit distance to transform one string into another, BiLingual Evaluation Understudy (BLEU) (Papineni et al., 2002) compares the n-gram overlap between the strings. In principle, BLEU is based on the well-known precision metric, counting how many n-grams are correctly output by the MT system. However, a predicted word can not count as correct more often than it occurs in any reference (called modified unigram precision). The 1-gram overlap can also be considered as adequacy, while longer n-grams represent the fluency aspect. BLEU is one of the most widely used MT evaluation metrics, as it is easy to compute and was one of the first metrics which correlates highly with human evaluation. However, one should note that sentence-level BLEU scores might not always be accurate, and that averaging across a test corpus is usually required for reliable results, or as the authors put it: “quantity leads to quality”.

National Institute of Standards and Technology (NIST) NIST (Doddington, 2002), named after the National Institute of Standards and Technology, is based on BLEU but weights n-grams by informativeness. Here, informativeness is simply defined as rareness, thus, rarer n-grams get higher weight. This has the advantage that common n-grams which are easier to generate correctly do not pull up the score, while rare n-grams, which are much harder to generate correctly, are rewarded with higher scores.

Metric for Evaluation of Translation with Explicit ORdering (METEOR) Compared to BLEU and NIST, Metric for Evaluation of Translation with Explicit ORdering (METEOR) does not only focus on precision, but also on recall, thus, the degree to which “the translation covers the entire content of the translated sentence” (Banerjee and Lavie, 2005). BLEU instead only has a so-called “brevity penalty" to penalize very short translations, which captures some degree of recall but naturally not its entirety. Instead, METEOR is based on the harmonic mean between unigram precision and unigram recall, with an adapted weighing where recall counts much more than precision. METEOR also considers synonyms, word stems, and a measure of fragmentation to capture the order of words.

2.1.7 Comparison of Machine Translation Paradigms

Now that we have seen which MT paradigms exist and how quality of MT systems can be determined, we give an overview of the comparative literature.

As already said before, the current state-of-the-art NMT yields higher **automatic evaluation** metric scores than SMT (Castilho et al., 2019). Toral and Sánchez-Cartagena (2017) compare NMT and PBSMT for 9 language pairs, and find better automatic scores for 7 out of 9 pairs. Wu et al. (2016b) agree and highlight that methods like operating on the sub-word level employed in NMT lead to even stronger improvements over PBSMT for morphologically rich languages.

Castilho et al. (2017) agree that NMT results in terms of automatic evaluations "are very promising, however human evaluations show mixed results": NMT increases the **fluency**, but in terms of **adequacy** and **PE effort** no paradigm clearly outperformed the other in their study on 3 different domains. This however partly contradicts Bentivogli et al. (2016), who report that the time required for PE NMT was reduced by 26% in comparison to SMT. Regarding fluency, Toral and Sánchez-Cartagena (2017) agree, finding that NMT output is "more fluent and more accurate in terms of word order compared to [PBSMT]".

Apart from investigating which paradigm overall shows the best results, researchers have investigated the **linguistic features** leading to these gains and under which circumstances NMT might not be the best. Bentivogli et al. (2016) report fewer lexical, and morphological errors, and find that NMT improvements get stronger for lexically richer text. Furthermore, NMT output has less word order error and particularly performs well on placing verbs correctly. These reordering improvements were also shown in the study by Toral and Sánchez-Cartagena (2017), where NMT also lead to improved reordering results across 9 investigated language pairs. A recent study by Díaz et al. (2020) investigated the main error types remaining in modern NMT across a large variety of language pairs and showed that the major problems fall into the category of "lexical semantic", comprising e.g., word ambiguity, unknown words or dialectical variants.

Another linguistic analysis of the paradigms compares not only SMT and NMT, but also RBMT (Macketanz et al., 2017) using English-German in the IT-domain. Interestingly, using their test suite of linguistic phenomena, they find that the overall average performance of the three paradigms is very similar. SMT would be best for terminology and quotation marks, while RBMT outperforms all others on compounds, verbs and phrasal verbs. On such verbs and phrasal verbs, NMT is apparently better than SMT, however, worse than RBMT. Both RBMT and

NMT are best on function words, long distance dependencies, as well as verb tense, mood, and aspect. Their overall conclusion therefore is that RBMT seems to handle given linguistic phenomena better than other systems, followed by NMT and then SMT. This also shows that NMT can learn rules coded in RBMT better than SMT. One should however note, that the explored NMT system was rather premature, so these findings might not be fully reliable.

Koponen et al. (2019) analyze **editing patterns** in PE and find that RBMT requires most deletions, SMT most insertions, while NMT has the greatest number of word-form changes and word-substitutions. Overall they conclude that even though the edit types are different for the different MT paradigms, these differences are "not necessarily reflected in process-based effort indicators".

For **sentence length**, Bentivogli et al. (2016) find that NMT outperforms PBSMT on all sentence lengths, however, here, Toral and Sánchez-Cartagena (2017) disagree and show that SMT performs better for segments longer than 40 words.

Overall there is consensus that NMT yields the best overall results (assuming there is enough training material for the language pair and domain), however, the error types are different for the MT paradigms, and there is no consensus on the impact of MT paradigm on PE effort. We will however revisit the topic of PE effort in a later section.

2.1.8 Human Parity

For some language pairs, NMT developers claimed to have achieved human parity: For example, the Microsoft system for English to Chinese news translation (Hassan et al., 2018) claimed to be on a level with professional translators and significantly above crowd-sourced non-professional translation. Similar statements of human parity have been made for other language pairs like English to Czech (Popel, 2018), however, here the authors warned that the evaluation scheme might be inappropriate for such claims (Bojar et al., 2018). As statements about human parity also capture lots of media attention and question the concept of PE, we use this section to examine whether human parity has indeed been reached for some language pairs.

Toral et al. (2018a) and Läubli et al. (2020) reinvestigate the claim of the Microsoft system, showing that "the finding of human–machine parity was owed to weaknesses in the evaluation design – which is currently considered best practice in the field" (Läubli et al., 2020). For proper evaluation of human parity, one should consider three aspects: (i) the language in which the source was originally written, (ii) the proficiency of the evaluators, and (iii) the inter-sentential context, i.e., the impact of rating sentence by sentence or seeing also the document's context.

(i) aims to prevent so-called "translationese", that is, a source which is actually a human translation of a text originating from another language. Such sentences were included in the test material to half the manual effort for corpus creation: WMT usually takes half of the sentences from one language (e.g., Chinese) and half from another language (e.g., English), translates both, and uses all sentences together for both language directions, thereby increasing the sample size by a factor of 2. However, having such translationese phenomena makes it easier for MT systems to recover the translation because translationese exhibits less lexical variety, and is therefore generally easier for MT systems. The authors were therefore able to show that human parity has indeed not been achieved when considering only sentences that were translated from their original language.

Regarding (ii), expert raters have a higher inter-annotator agreement and judge the gap between Human Translation (HT) and MT to be wider, because they also consider nuances of the language that non-professionals would not. Thus, human parity analyses should rely on expert raters.

For (iii) it was shown that raters assessing adequacy on whole documents prefer HT, while viewing sentences in isolation they show no significant preference (Läubli et al., 2020). This is because inter-sentential aspects are not considered by NMT systems trained on a sentence-level. However, recent research on document-level MT systems also aims to better capture the context, thus potentially fixing this gap in the future (Kang et al., 2020).

Furthermore, explorations of human parity should ensure that the translation was done from scratch by professionals and not post-edited; that bilingual raters are employed to compare the output to the source and not to a reference; and that reference translations should not be heavily edited, as this increases fluency but reduces adequacy "to the degree that they become indistinguishable from MT in terms of accuracy" (Läubli et al., 2020).

Some of the proposed guidelines, namely to use original source text only and to provide access to the whole document when rating, are now considered from WMT 2019 onwards (Barrault et al., 2019), thereby making MT quality analyses more reliable.

2.1.9 Summary

In this section, we have briefly summarized the history of machine translation and how the research community moved from rule-based MT, to statistical MT, and now to neural MT, going hand in hand with an increasing quality. Some claims of human parity achieved by NMT have been made, however, more detailed analysis showed the importance of MT evaluation schemes, and how evaluation methods that used to be sufficiently accurate now need to be adapted to be sensitive enough for the new level of MT quality.

An investigation of the pros and cons of the paradigms has, however, shown that NMT is not the best tool in every linguistic aspect yet. Furthermore, with thousands of languages existing in the world, and plenty of unresearched text domains, one should not be fooled by impressive results of NMT on a few language pairs and in common domains. MT is still far from solved and we hope to see as much improvement in the following years as we did in the last years.

Views on MT are also changing: Whereas historically, the relationship between human and machine translation was considered adversarial due to claims the MT would replace the human, more and more stakeholders start to view the two paradigms as symbiotic (Balashov, 2020). This makes sense, as no matter how good NMT has become even under the best conditions, blindly trusting MT systems is and will not be possible in many use cases. Humans will remain in the center of the translation process, to consider cultural nuances, to ensure the MT output is not biased, and in general, to ensure a high-quality translation not just on average, but for every single sentence. So even if the translator's job is and will be strongly impacted by advances in MT research, it will not become superfluous. Therefore, this thesis focuses on enhancing the human-machine collaboration in the form of post-editing.

2.2 Post-Editing

The process of manually capturing and correcting mistakes, as well as selecting, adapting, and recombining translation candidates produced by automatic translation segments, is called Post-Editing (PE). While the strong improvements in MT have only been achieved in recent years, the concept of post-editing dates back a long time, with books being published on the topic as early as 1982 (Lawson, 1982). We will first provide a motivation for PE, then look at the PE task in more detail, inspect how it impacts time and quality, and finally present studies on translators' perceptions towards PE.

2.2.1 Motivation for Post-Editing

The motivation for PE is that correcting the often erroneous output of translation technologies may substantially boost productivity compared to a manual translation from scratch (Green et al., 2013). PE thus combines the advantages of both AI and human intelligence: While the AI is good at quickly proposing draft translations of nowadays often high quality, a human with high proficiency in source and target language is needed to ensure that the meaning is identical, to analyze lexical and semantic nuances, and to understand the segment of text in a large (con)text, including the target audience, their cultural background etc. Thus, the task changes from mostly text production to comparing and adapting machine-generated proposals, or put differently, from control to supervision.

While MT improved tremendously in recent years, there will always remain tasks for the human translator: A simple example would be an under-specified source sentence, for which a translation requires online research or even talking to the author of the original sentence to understand the intended meaning. Furthermore, many situations like legal text or package inserts for drugs will always require a human for ensuring correctness. As we have seen in the previous section, MT is closing in on human parity, however, only for very frequent language-pairs and domains with large amounts of training data, so for almost all real-life scenarios, human PE is still and will remain relevant for many years to come.

Nevertheless, the enthusiastic media reports about MT progress are pushing Language Service Providers (LSPs) to integrate MT PE into their day-to-day workflows, with 50% of LSPs already offering MT PE in 2016 (Lommel and DePalma, 2016). A large-scale study conducted by Common Sense Advisory (CSA), a major translation think tank, further showed that around 34% of participating translators offered PE of MT in 2020 (Pielmeier and O'Mara, 2010). It has therefore been argued that MT integration in the translation process is "as disruptive and revolutionary as the disruption caused by the introduction of translation memory technology in the 1990s" (O'Brien and Moorkens, 2014).

2.2.2 Post-Editing Tasks and Workflow

Two types of post-editing can be distinguished: *Light post-editing* aims to make the MT output readable, understandable, and convey the correct meaning, but does not give any quality guarantees beyond this. *Full post-editing* additionally ensures that the language style, choice of words, etc. is appropriate and therefore aims to achieve the level of quality as 'that of text that has been translated by a human linguist' (De Almeida, 2013). As the distinction into light and full PE existed for decades in which MT quality has improved drastically, nowadays the raw MT output is often good enough to understand its meaning, such that light post-editing is becoming less relevant.

While this dissertation focuses on PE of MT, other text types can also be post-edited. On the one hand it could be a HT from another translator which requires reviewing. PE of HT, however, encompasses the correction of different types of errors compared to PE of MT, e.g., typos or inconsistencies that a MT would not make (O'Brien and Moorkens, 2014). On the other hand, the initial draft could have been proposed by a Translation Memory (TM) system. Simply put, TMs are large databases containing already completed human translations which are matched against the sentence to be translated to provide a starting point for PE (see subsubsection 2.3.1). However, PE of MT and PE of TM also differ in the types of error that require correction: MT is a potentially incorrect translation of the real source sentence, whereas TM is a correct translation of a similar but usually not completely matching source sentence. Thus, when PE TM output one must particularly consider differences in the source sentences, while PE of MT requires correcting translation errors that no human translator would have made (Moorkens and O'Brien, 2017).

The kind of PE that we focus on, MT PE, substantially changes the translation workflow (Zaretskaya and Seghiri, 2018): An early comparative study between PE and manual translation by Carl et al. (2011) showed a modest improvement in quality when PE, which was also faster and had significantly more fixations on the target text than manual translation. As expected the study also showed more deletions, more navigational keys, and less insertions in the PE condition. De Almeida (2013) found that post-edits typically are reorderings, changes of capitalization, and changes of the inflection based on gender or number, however, as discussed above, the required changes heavily depend on the type of MT that has been used. Green et al. (2013) analyzed different interaction patterns during PE: While traditional translation consists of gisting, drafting, and revising phases (Carl et al., 2010), these phases are interleaved in PE (Green et al., 2013). Furthermore, Green et al. (2013) found longer pauses and a significantly reduced amount of mouse and keyboard events in the PE condition. They also showed that MT suggestions help some subjects more than others. Even though PE and HT differ that strongly, only "few professional translators have received training either in machine translation technology or in post-editing practices to date" (Moorkens and O'Brien, 2017). A manual inspection of the courses offered to translation students at universities today shows that this is now changing.

2.2.3 Effort and Time Savings

The basic assumption of PE is of course that the more work is already done by the AI, the the less work remains for the human, thus reducing effort and saving time. On the other hand, post-editors need not only read the source text, but additionally consider the translation proposal, which can have a variety of errors simultaneously combined with useful chunks of text, thus requiring continuous scanning of both texts, making a plan how to ideally use the translation proposal, and in the end apply the required changes. The original assumption that the less work remains for the human the quicker it gets, is therefore not necessarily true. This has lead to a variety of studies on PE effort and time savings in the past:

Early work by Krings (2001) found that PE decreased time compared to traditional translation by 7% when done on paper, but increased time by 20% when done on a computer screen. While the measured times may not be precise since a think-aloud protocol was used, this shows how strongly benefits gained from PE depend on interface design. Zampieri and Vela (2014a) found that PE was on average 28% faster for technical translations. Aranberri et al. (2014) show that PE increases translation throughput for both professionals and lay users (although the latter benefit more strongly from the MT), and Läubli et al. (2013) find that PE also increases productivity in realistic environments (in contrast to the many existing isolated feature experiments). Furthermore, the recent study on post-editing a novel by Toral et al. (2018b) shows that the underlying MT paradigm impacts PE productivity, where post-editing using NMT increases throughput by 36% compared to 18% for SMT. Whereas the exact speed gains depend on the chosen language pair, MT system, text domain, and the translator, the conclusion of these studies is very clear: there is a significant gain in productivity when PE compared to traditional translation.

2.2.4 Quality

After having seen that PE saves time, the obvious question is if those time savings also impact the translation quality, e.g., because post-editors might accept imperfect MT proposals to save time. This section reviews studies on PE quality in terms of correctness, preference, linguistic properties, and creativity.

Correctness

A study investigating PE quality on three different language pairs by Green et al. (2013) showed that PE indeed reduces errors. In contrast to prior work (including e.g., Krings (2001)) Green et al. (2013) payed special attention to a rigorous, and controlled analysis without any unintended factors influencing the PE time or quality. Interestingly, a study a few years earlier by Guerberof (2009) still found a comparable amount of mistakes between human translated and post-edited text. Daems (2016) agrees with Green et al. (2013) that PE is better in terms of adequacy, however, finds that manual translation is better for acceptability, meaning that it "is true to the target (con)text and audience".

Preference

In terms of preference of the final translation, we see a similar trend: In 2009, Fiederer and O'Brien (2009) found that human translations were selected as a favorite 63% of the time, whereas Green et al. (2013) found significantly more preference for PE for several translation directions just four years later. In Bowker and Ciro (2015), 49% preferred text that was PE compared to 42% preferring HT. Similarly, both the participants and an automatic classifier in Daems et al. (2017) were unable to distinguish PE from HT text.

Post-Editese

However, language is more than correctness and preference. Thus, Toral (2019) investigates a concept similar to the above discussed translationese, namely post-editese. In particular, he investigated if PE translations are different from HT and indeed finds that PEs are "simpler and more normalised and have a higher degree of interference from the source language than HT". Reinvestigating the earlier papers shows that already Fiederer and O'Brien (2009) found that participants ranked PE higher than HT for accuracy and clarity, but found HT better in terms of style. Similarly, Green et al. (2013) already showed that the translator is primed by the output of the MT system and Čulo and Nitzke (2016) found less variation in PE than in HT, attributing it to a "shining-through" of the MT to the PE version. Farrell (2018) investigated markers of MT in PE output, and found that "PEMT may lack the variety and inventiveness of HT". These findings were replicated and extended in Toral (2019), who found least lexical variety in MT (as it prefers frequent solutions), then PE (as it primes translator), and most variety in HT. He also showed that there is less variety in NMT than in SMT. Furthermore, the lexical density (percentage of content words), was lower in both PE and MT than in HT, but comparable between the two, thus arguing that PE/MT are lexically simpler. Here again PE-NMT showed an even lower density than PE-SMT. In terms of length ratio (comparison between the source and target sentence length), MT and PE are again more similar to the source than HT. Apart from this, Toral found further evidence for interference from the source, where the Part of Speech (PoS) were most similar for MT, then PE, and least for HT. Here, however, NMT had less interference from the source than SMT. Taking all of these findings together, Toral argues that in the long term, PE might lead to the target language being simplified and overly influenced by the source language.

Creativity

With increasing MT quality, creativity is frequently stated as a factor where humans will continue to outperform machines. To investigate whether this is true, Guerberof-Arenas and Toral (2020) explored the creativity and reading experience of a fictional story in the conditions MT, MT PE, and HT. The results showed that indeed the involvement of translators (HT and PE) show more creativity, however, there were no statistical differences between HT and PE, but only a trend that creativity might be higher in HT. Nevertheless and similar to other studies, translators subjectively felt that their creativity was limited in the PE condition. An additional analysis of "creative shift", i.e., translations deviating from the source through abstraction, concretisation, or modification, revealed that HT provides more novel translation that is less constrained by MT, which is in line with Toral (2019). The authors thus concluded that in HT and PE the "professional translators add the creativity factor, by providing solutions that are both novel and acceptable, that MT is lacking at present."

2.2.5 Translators' Perceptions

So far we have seen that PE increases productivity, reduces errors but also changes the linguistic properties of the final translation. Overall, the advantages have lead to a wider adoption of MT technology into translation workflows (Zaretskaya and Seghiri, 2018; Zaretskaya et al., 2016). This section analyzes whether translators also like the transition towards PE.

Older research tackling this question showed a strong dislike of translators towards PE (Lagoudaki, 2009a; Wallis, 2006), for which they are sometimes also paid less. More recent studies agree that translators are still cautious about PE and question its benefits (Gaspari et al., 2014; Koponen, 2012), reasons being that it is considered less creative and assumed to be slower than translation from scratch, but also in part because it is seen as a threat to their profession (Moorkens, 2018).

Kelly goes a step further and even calls PE 'linguistic janitorial work'[4]. One aspect that might lead to this perception is that PE tends to require repeatedly fixing similar MT errors, thus, O'Brien and Moorkens (2014) argue that the MT engine should ideally learn from human post-edits, which is one of the topics addressed in this thesis. Back in 2014 when this work was published, 56% of translators saw MT as problematic as it was "still in baby shoes" or "just horrible". They also found that MT often contains errors a human would not make, e.g., mixed gender between noun and adjective, or a singular/plural mismatch between noun and modifier. While this was true for previous MT approaches, modern NMT is usually very fluent. Other reasons for the found dislike are of a more general nature: PE is a revision task and therefore similar to revision of other translator's works, which some translators are good at and enjoy, while others do not (Mossop, 2007). Finally, an argument against PE is the fact that it is perceived to be more cognitively demanding than translation from scratch because the MT needs to be considered additionally to the source (O'Brien and Moorkens, 2014). This might also be the reason why translators report to be less productive when PE, although as also discussed above, throughput measurements clearly show the opposite (O'Brien and Moorkens, 2014). According to a large-scale study involving more than 7000 translators and interpreters conducted by Common Sense Advisory (CSA), around 34% of translators offer PE of MT, but only 3% prefer PE over pure translation or editing human translation (Pielmeier and O'Mara, 2010).

In contrast to these rather negative studies, Green et al. (2013) demonstrated that translators actually strongly prefer PE and argue that "users might have dated perceptions of MT quality". A recent study by Vela et al. (2019) found similar results, where professional translators, who were given the choice between translation from scratch, TM, and MT, chose MT in 80% of the cases, highlighting the importance but also popularity of MT PE.

One distinction in terms of attitude towards PE seems to exist between experienced translators, who exhibit rather negative attitudes and are rather reluctant to take on PE jobs (Moorkens and O'Brien, 2015) and novice translators, who have more positive views on PE and are better suited for PE jobs (Yamada, 2015). Daems (2016), however, showed that professionals believed they could produce similar quality translations with or without PE, whereas students seemed to struggle with the cognitive processing of meaning shifts.

Thus, overall there does not seem to be consensus on the translators' perceptions of PE: A lot of research suggests that translators dislike it or at least are cautious, while other works show positive opinions on PE. Daems (2016) nicely summarizes that her participants found PE useful but preferred manual translation as it is more rewarding.

[4] https://www.huffpost.com/entry/why-so-many-translators-h_b_5506533

2.2.6 Summary

This section provided an in-depth review of the post-editing process, explaining the steps involved and how improvements in MT are pushing PE into translator's day-to-day work. We have also discussed that the errors that need to be corrected changed with different MT paradigms, but are in general not comparable to the errors in TM matches or during human revision. In terms of time savings, the literature nowadays agrees that PE is indeed faster than translation from scratch, however, quality must be considered with more care. On the one hand, the amount of errors decreases in PE, but on the other hand a phenomenon called post-editese has been shown in several studies, even with modern NMT. Thus, when style matters most and lexical variety is of highest concern, one should avoid PE if the time allows it, but when correctness matters most, PE is the proper approach. For literature translation, traditional translation might thus be the better choice, while for translations of technical documents, PE would be more suitable. Finally, we discussed the human attitudes towards PE, and saw that lots of papers reported very negative opinions on the topic, especially in older works. However, we have also seen works where translators preferred PE, and have argued that one important consideration when talking about translator's perceptions is their years of experience within the profession, with novice translators, that nowadays also participate in PE courses during their studies, being much more positive towards PE.

2.3 Computer-Aided Translation

This section reviews the work environments of translators: Computer-Aided Translation (CAT) tools. The idea to focus on "cooperative man-machine systems" instead of aiming for fully automating the entire translation process was already proposed by Kay in 1980 (republished in Kay (1997)), envisioning a word processor optimized for translation called "translator's amanuensis". Whereas no CAT tools existed at that time, computers gained popularity among translators, not just for word processing, but also for the creation of glossaries, and even the first versions of TMs (Hutchins, 2007).

Nowadays, most professional translators work within fully-fledged CAT environments (Van den Bergh et al., 2015), and maximizing the potential of these workbenches is thus one of the priorities for both the research community and industry (Mesa-Lao, 2012). CAT tools minimally show a text window for the source text and one supporting text editing operations for the target, which can initially be either empty (translation from scratch), or populated with a translation proposal (PE). Usually however, CAT tools offer a wide range of additional features supporting the translation process.

We will first give an overview of the features provided by CAT environments, then talk about the available systems both from academia and industry, present studies that show the limitations of current environments and discuss user needs. Last, we will talk about approaches exploring interaction modalities other than mouse and keyboard and summarize the main take-aways. Overall, this section provides a solid foundation for the development of our own multi-modal CAT environment in Part II.

2.3.1 Features

Kay (1997) advocates a view “in which machines are gradually, almost imperceptibly, allowed to take over certain functions in the overall translation process”. This vision has nowadays become a reality, where "it is widely acknowledged that technological change - for example the widespread use of translation blogs, wikis, open-code translation software, crowdsourcing, MT, TM, cloud-based translation tools, corpora, etc. has influenced the way in which both professional and trainee translators work" (García-Aragón and López-Rodríguez, 2017).

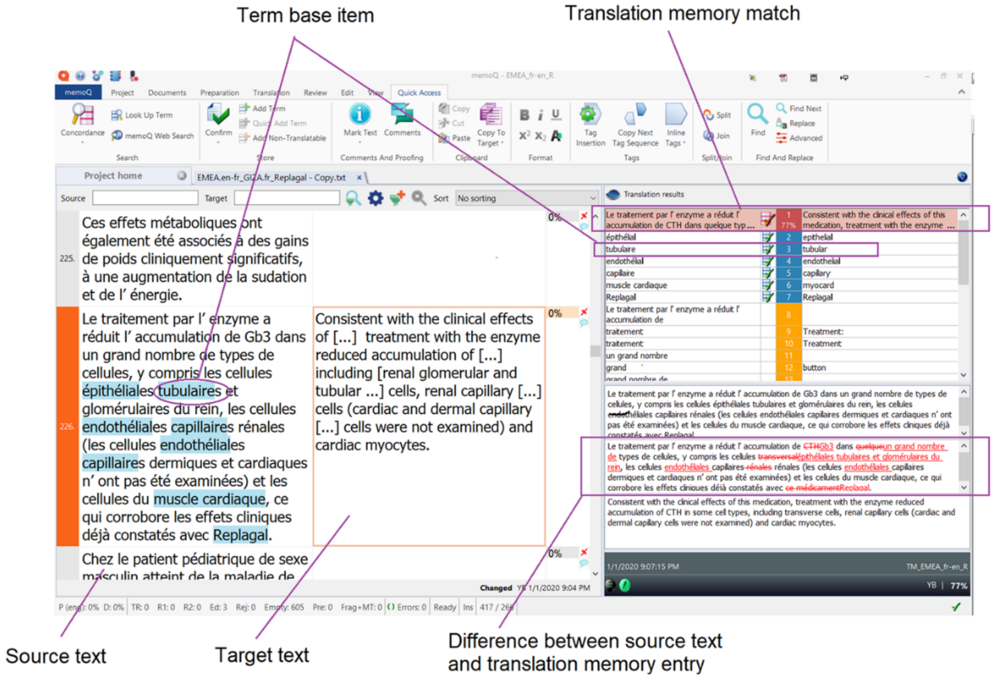
This section provides an overview of several of these interesting aspects integrated in CAT tools and discusses how these features can support translators in their day-to-day work.

Administrative Features

Translation jobs usually start with a customer request, defining the amount of text that needs to be translated, which domain it comes from, when the translations are needed, and what is paid for it. While freelance translators often directly load the raw text file formats like .docx, .pptx, .pdf into their translation environment and start translating, bigger translation jobs require more thorough planning and management. To facilitate this, CAT tools offer a variety of project management tools, like splitting tasks and distributing them among translators, assigning review tasks, where one translator reviews the work of another translator, monitoring progress, or having team communication.

Displaying Text

Even though CAT tools are designed for experts and therefore at first glance quite complex and overloaded, the basic functionality while translating is very simple: at their core, CAT tools need to visualize the source text and the continuously evolving (translation from scratch) or pre-filled (PE) target version of the text. The source and target text can be (i) segmented into individual sentences or kept in whole paragraphs, and (ii) displayed side-by-side or one above the other, with a segmented left-right visualization being the most common combination among translators (85% according to Läubli et al. (2021)). Läubli et al. (2021) explored all four combinations and showed that segmentation speeds up error identification and typing, but makes revision of super-sentential context harder. For pure copying, a top-bottom visualization led to speed gains, however, for revision tasks like PE, a left-right visualization is better suited. Similar to standard word processors like Microsoft Word, CAT tools also support styling of text (bold, underline, fontsize, etc.) and offer special functionality to apply the source text style on the target text (e.g., mark text in the target, use a hotkey while clicking on a source word whose style is then applied to the selection). In segmented visualizations, translators generally do not see the text in the layout and style of the original document. This aims to keep the focus on the text, but if needed, translators can choose to view the text in its original formatting. A horizontal text visualization can be seen in Figure 2.2 on the left (source text and target text).



225. Ces effets métaboliques ont également été associés à des gains de poids cliniquement significatifs, à une augmentation de la sudation et de l' énergie. 0%

226. Le traitement par l' enzyme a réduit l' accumulation de Gb3 dans un grand nombre de types de cellules, y compris les cellules épithéliales tubulaires et glomérulaires du rein, les cellules endothéliales capillaires rénales (les cellules endothéliales capillaires dermiques et cardiaques n' ont pas été examinées) et les cellules du muscle cardiaque, ce qui corrobore les effets cliniques déjà constatés avec Replagal.

Consistent with the clinical effects of [...] treatment with the enzyme reduced accumulation of [...] including [renal glomerular and tubular ...] cells, renal capillary [...] cells (cardiac and dermal capillary [...] cells were not examined) and cardiac myocytes. 0%

Chez le patient pédiatrique de sexe masculin atteint de la maladie de 0%

Changed YB 1/1/2020 9:04 PM

P (eng): 0% D: 0% | TR: 0 R1: 0 R2: 0 Ed: 3 Rej: 0 Empty: 605 Pre: 0 Frag+MT: 0 | Errors: 0 | Ready | Ins | 417 / 266

Le traitement par l' enzyme a réduit l' accumulation de CTH dans quelque typ...	1 77%	Consistent with the clinical effects of this medication, treatment with the enzyme ...
épithélial	2	epithelial
tubulaire	3	tubular
endothélial	4	endothelial
capillaire	5	capillary
muscle cardiaque	6	myocard
Replagal	7	Replagal
Le traitement par l' enzyme a réduit l' accumulation de	8	
traitement	9	Treatment:
traitement	10	Treatment
un grand nombre	11	
grand	12	button
grand nombre de	13	

Le traitement par l' enzyme a réduit l' accumulation de Gb3 dans un grand nombre de types de cellules, y compris les cellules épithéliales tubulaires et glomérulaires du rein, les cellules endothéliales capillaires rénales (les cellules endothéliales capillaires dermiques et cardiaques n' ont pas été examinées) et les cellules du muscle cardiaque, ce qui corrobore les effets cliniques déjà constatés avec Replagal.

Le traitement par l' enzyme a réduit l' accumulation de ~~CTH~~Gb3 dans ~~quelque~~un grand nombre de types de cellules, y compris les cellules ~~transversal~~épithéliales tubulaires et glomérulaires du rein, les cellules endothéliales capillaires ~~rénales~~ rénales (les cellules endothéliales capillaires dermiques et cardiaques n' ont pas été examinées) et les cellules du muscle cardiaque, ce qui corrobore les effets cliniques déjà constatés avec ~~ce médicament~~Replagal.

Consistent with the clinical effects of this medication, treatment with the enzyme reduced accumulation of CTH in some cell types, including transverse cells, renal capillary cells (cardiac and dermal capillary cells were not examined) and cardiac myocytes.

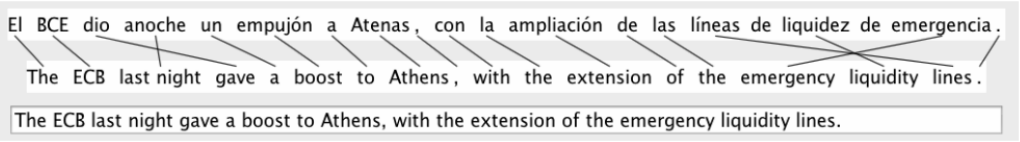
1/1/2020 9:07:15 PM TM_EMEA_fr-en_R

YB | 77%

Figure 2.2: The memoQ CAT tool, with source and evolving target segments on the left. Translation results on the right, with terms from the term base (also highlighted in the source), as well as TM matches, and changes between the actual source and the source of the TM match (taken from Balashov (2020)).

Source-Target Alignments

CAT tools with MT support can also offer alignments between source and target, indicating which parts of the source likely correspond to which parts in the MT. Such alignment tools can either be created independent of the MT system used, or use the internal alignments of the MT system to determine which links to visualize. Schwartz et al. (2015) showed that displaying alignments helps increase PE quality since it facilitates quick comparisons of even complex sentence structures. The alignments as displayed in their study can be seen in Figure 2.3.


El BCE dio anoche un empujón a Atenas , con la ampliación de las líneas de liquidez de emergencia .
The ECB last night gave a boost to Athens , with the extension of the emergency liquidity lines .

The ECB last night gave a boost to Athens, with the extension of the emergency liquidity lines.

Figure 2.3: Word alignments between source and target (taken from Schwartz et al. (2015)).

Terminology Management

Both monolingual and bilingual dictionaries are widely used tools when learning languages and translating. Therefore, the idea to integrate them into CAT tools was already proposed by Kay in 1980 (Kay, 1997), where he also argued to make dictionaries editable and share notes regarding terminology with other translators. Nowadays, CAT tools offer so-called terminology management, "allowing one to create, modify, look up, and reuse the translations of individual domain-specific terms stored in a [Term Base (TB)] or specialized glossaries" (Balashov, 2020). Such terminology management helps ensure consistency throughout a document or even across documents of the same company and allows sharing terminology across translators.

Vandeghinste et al. (2019) conducted an online questionnaire capturing how 187 professionals acquire domain-specific terminology, followed by field observations of 13 translators and terminologists. Interestingly, they found that 88% collected their terms manually, and 22% through semi-automatic term extraction programs. 52% stored their terms in their CAT TB, while 43% simply used a spreadsheet. In field observations however, only 1 out of 16 observed translators actually stored the researched term in their TB. The advantage of properly storing terms in the TB is that CAT tools can offer a variety of functionality like highlighting TB entries in the source and inserting their target language representation through hotkeys and buttons. TB entries in memoQ can be seen in Figure 2.2.

Translation Memory (TM)

Kay's visionary proposal further considered looking up compound words or sequences, which is the idea underlying Translation Memory (TM) technology that became widespread after its introduction in 1989 by Trados. In contrast to MT, TM usage is already widely accepted among professional translators to increase productivity (Screen, 2016). Simply put, TMs are large databases containing already completed human translations (both from the current user and shared among large groups of translators) which are matched against the sentence to be translated to provide a starting point for PE. A good example where TMs are useful is "prescription drug information which has mandatory section headings and standard language inside the sections, which varies from document to document mostly in the names of substances and numerical values of various parameters" (Balashov, 2020). Matches can be either exact or fuzzy, meaning that the exact or similar segments were translated before and can be used as is or as a basis for the current translation. After PE a TM match or manual translation, the source and final translation are added to the TM for reuse in the future. Whereas the basic concept of string matching for TM sounds rather simple, lots of improvements have been made over the past decades, including ways to incorporate semantic knowledge like paraphrasing (Utiyama et al., 2011) or syntactic information (Vanallemeersch and Vandeghinste, 2014). Figure 2.2 shows how TMs can be used within the widespread commercial CAT tool memoQ.

Usually, when encountering an exact match, the match can simply be used as the final translation. Depending on the match score for fuzzy matches, it might make sense to either translate from scratch, or to PE the match. There is usually an agreement between client and service provider that fuzzy matches below a specific value, e.g., 75%, need full translation from scratch and are paid as such (O'Brien and Moorkens, 2014). Often full matches are visualized in green, while "fuzzy matches are visualized in yellow, orange, or lighter shades of green" (Vieira and Specia, 2011). Apart from deciding when to use TM matches, Nayek et al. (2015) investigated the use of color coding to show similarities between input sentences and TM matches, and thereby guide the translators directly to the potentially bad parts of the translation proposals. More on this highlighting is explained in subsubsection 2.3.2.

Another interesting aspect of TM is its ability to ensure consistency: Translators are confronted with the translations chosen by their peers for similar source segments, and can therefore stick to project-wide terminology. The increased consistency alongside the higher throughput lead to a high appreciation of TM among translators (Moorkens and O'Brien, 2017).

Fuzzy matches in the TM can sometimes be automatically repaired, a concept called "fuzzy match repair". Here, the TM is used to retrieve a fuzzy match that is aligned to the current sentence. If an unmatched term is detected and either contained in the TB, is a number, named entity, or location, etc., it is automatically substituted in the fuzzy match (Balashov, 2020).

LeBlanc (2017) investigated how business practices shifted through the introduction of TMs and found that enforcing translators to use TMs decreases the quality of the resulting translations according to the perceptions of the translators. Furthermore, they feel less satisfied with their profession, feeling degraded to assembly line workers (shift in status) that are only required to become more and more productive. Thus, offering TM as an aid is considered valuable, but enforcing translators to use it and therefore also raising productivity expectations is a major downside. We can see this as a consistent pattern, that language technologies are designed to support translators and indeed do so, however, the implied expectations in terms of productivity and the reduced creativity and freedom during translation, heading from being a writer to an assembly line worker, negatively impact translators' perceptions of the technologies.

Translation Memory vs. Machine Translation

While TMs continue to be useful tools for translating segments that are highly similar to matches in the database, MT is becoming more and more relevant due to increased performance in recent years. Whereas improvements in MT are naturally essential, Church and Hovy (1993) argue that it might be even more important to create good applications using MT than improving MT itself, thus, emphasizing the vital role of CAT tools for human-AI collaboration.

With the availability of both TM and MT in modern CAT tools, editing segment by segment usually means switching back and forth between PE TM for matches above a certain threshold and PE MT when no good TM match was found (O'Brien and Moorkens, 2014). However, the threshold is not commonly agreed upon, it depends on factors like the language pair, the type of text, and its domain. Older research, therefore using outdated MT technology, suggested that PE TM matches with a fuzzy score in the 85 to 94% range is equivalent to PE MT output (Guerberof, 2008). MT and TM can also be combined, e.g., by training a binary classifier to predict if MT or TM is more suitable for PE (He et al., 2010). Apart from deciding which proposal to use, Simard and Isabelle (2009) explored a PBSMT model that leverages the top matches from TM. Another interesting combination strategy by Koehn and Senellart (2010) translates mismatched parts of a TM match using SMT to fill the gap.

Comparing PE of modern NMT to PE of TM matches, Sánchez-Gijón et al. (2019) find that PE NMT requires less editing than PE TM matches, but takes longer on average. In general, TM as a feature is currently still often valued higher than MT, with 75% of translators believing it to increase throughput and preserve consistency, while 40% think MT usage is problematic due to the amount of errors (Moorkens and O'Brien, 2017). However, as stated before, the recent study by Vela et al. (2019) showed that professional translators choose MT over TM and translation from scratch 80% of the time, even though the participants did not agree on which sentences to choose MT for. This shows that MT is gaining popularity over TM, but suggests that there are no obvious cues when to PE MT. In the long run, we believe that as MT becomes better and better for more language pairs, PE of MT will become the major theme and therefor focus on it in this dissertation. Even though using TM as a basis for sentences that are highly similar to matches in the database will always yield good results, we expect MT output to be similarly good on such frequent sentences in the future, and better than TM on less frequent ones.

Consistency Checking, Quality Assurance, & Concordance Functionality

Furthermore, consistency checkers can be run and show the human translator if source words have been translated consistently throughout the document. Quality Assurance (QA) goes a step beyond that by subsuming consistency checking, but also verifying aspects such as terminology inaccuracies (based on

TBs), non-translated segments, problems with spacing, typos in proper names, punctuation checking, and other consistency checks (Vieira and Specia, 2011).

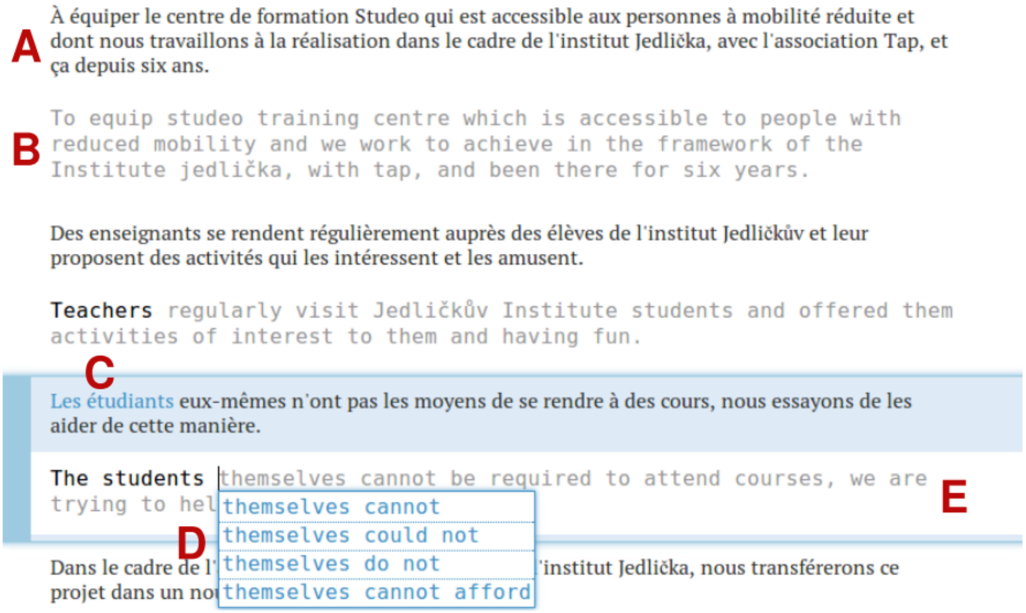
Another support tool is the concordance search, that is, looking up the correct use of terms and sub-sentences within a large corpus (e.g., the TM or a mono-/bilingual corpus) to show the correct usage and to provide ideas on how to best use the words. In their studies, Vandeghinste et al. (2019) found that translators see concordance search as an indispensable feature. Figure 2.2 further shows in orange (items 8 to 12) that CAT tools like memoQ can automatically determine the largest substring that is contained in the TM. Translators can then inspect the whole TM match and extract relevant parts of the translation for reuse.

Interactive Machine Translation (IMT)

In normal PE of MT, the MT acts first, aiming to produce the single best sentence it can, and the human then follows as a second stage and corrects the errors of the MT. Instead of this sequential process, the MT can also be used to dynamically provide the human with alternatives for the remainder of the sentence when typing a partial sentence. More generally, Interactive Machine Translation (IMT) guesses which output translation the human is aiming to produce given both MT input and manual changes. IMT thus heavily focuses on the concept of integrating the knowledge of a human expert (the translator) with a MT system, by facilitating a back and forth communication between the two. A study by Green et al. showed that translators using such interactive proposals were slightly slower, however, produced slightly higher quality translations (Green et al., 2014a,b). Figure 2.4 shows the interface used in their study, providing type-ahead predictions for a limited amount of following words in a dropdown (D), and the full proposal in gray in the editing window (E). Using the commercial CAT tool Lilt, that is based on the works by Green et al., Daems and Macken (2019) explore the influence of the MT paradigm (SMT vs. NMT) used in IMT. They find that even though SMT for IMT contains more errors than NMT, there are no significant differences in either translation time or effort, as NMT errors might be harder to detect and correct. Nevertheless, subjectively translator's preferred NMT over SMT for IMT. Knowles et al. (2019) explore a similar neural system for IMT but could not find significant performance differences compared to PE.

In the above works, the user still starts with an empty target field and types the translation while receiving (and potentially accepting) MT proposals. In contrast, the widely used DeepL[5] online translator directly fills the target with the best translation hypothesis but adds interactivity: Users can click on any word, which re-runs the decoding process at that step to receive and display a list of alternatives for this position (see Figure 2.5). Upon selecting an alternative, the decoding is continued from that position onward, again proposing the best translation hypothesis constrained by the prefix defined through the user selection. Thus,

[5] https://www.deepl.com/translator



À équiper le centre de formation Studeo qui est accessible aux personnes à mobilité réduite et dont nous travaillons à la réalisation dans le cadre de l'institut Jedlička, avec l'association Tap, et ça depuis six ans.

To equip studeo training centre which is accessible to people with reduced mobility and we work to achieve in the framework of the Institute jedlička, with tap, and been there for six years.

Des enseignants se rendent régulièrement auprès des élèves de l'institut Jedličkův et leur proposent des activités qui les intéressent et les amusent.

Teachers regularly visit Jedličkův Institute students and offered them activities of interest to them and having fun.

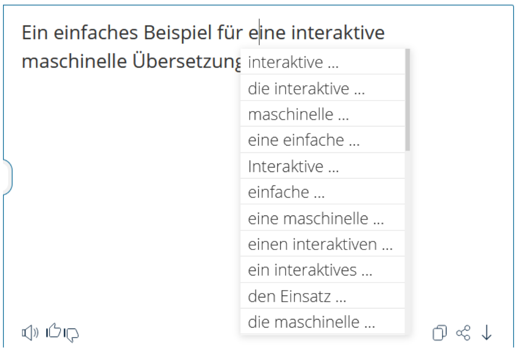
Les étudiants eux-mêmes n'ont pas les moyens de se rendre à des cours, nous essayons de les aider de cette manière.

The students themselves cannot be required to attend courses, we are trying to hel

Dans le cadre de l' 'institut Jedlička, nous transférerons ce projet dans un no

Figure 2.4: Green et al.'s interface for Interactive Machine Translation: (A) source, (B) target text, (C) source coverage of the already typed text, (D) autocomplete suggestions, and (E) full completion suggestion (taken from Green et al. (2014b)).

this is a different level of interactivity compared to the type-ahead IMT approach, where even more work is offloaded on the MT, while still offering the possibility to intervene manually.



Ein einfaches Beispiel für eine interaktive maschinelle Übersetzung

Figure 2.5: Alternatives proposed interactively in DeepL.

A recent paper by Navarro and Casacuberta (2021) proposes a similar approach, relying solely on mouse actions as bandit feedback to correct the proposed translation. The user always sees a full translation proposal and approves prefixes by positioning the cursor on a word in the sentence, which at the same time tells the system to provide an alternative (next-most likely) hypothesis for the remainder of the sentence. If this alternative is still incorrect, the user can explicitly ask for another suggestion. By simulating users, the authors showed that such a system can save a lot of typing; however, an actual study involving real humans (and correspondingly accounting not just for typing but also for cognitive effort) was not conducted.

Instead of only extending the remainder of the translation proposal after the change (i.e., to the right), Weng et al. (2019) propose to manually correct the most significant errors first, then automatically correct the sentence to the left and right of that change. The motivation behind this is that by fixing critical errors first, smaller errors might be fixed automatically by the AI. For this, their system called CAMIT, offers bi-directional decoding. The general concept was already outlined by Kay in 1980 (Kay, 1997), where he states that the big problem of cascading errors in MT could be avoided, when the human fixes the first decision and thereby resolve a whole chain of errors. When presenting concrete CAT tools in subsection 2.3.2 (both academic and non-academic ones), we will see that many of them indeed offer some form of IMT.

Quality Estimation (QE)

Quality Estimation (QE) is the task of automatically predicting the quality of a usually automatic translation (e.g., MT) without access to a reference translation (Specia et al., 2009). Note the distinction to quality metrics like BLEU or TER which rely on reference translations for comparison. Usually QE models are trained in a supervised fashion, using the source and MT proposal as features and predicting either the time it takes to PE the MT proposal (based on prior studies), some manual human rating, or an automatic quality score in comparison to a reference. Thus, QE models aim to learn patterns from the link between source, MT proposal, and corresponding quality estimate during training, so that it can provide accurate quality estimates for unseen (source, MT proposal) data pairs at prediction time. Instead of training only based on source and MT proposal, which is called *black-box* QE, one can also use internal features (e.g., its internal confidence) of the MT system, called *glass-box* QE (Biçici et al., 2013). In terms of granularity, QE systems can work on the document-level, on the sentence-level, or on the word-level (Specia and Shah, 2018). While document- and sentence-level QE require the prediction of only a single quality estimate (e.g., a percentage or binary 'OK'/'BAD' label), a word-level QE needs to output a label for every word and gap, as shown in Figure 2.6.

Having reliable QE models can help the PE process: Document-level QE can be used to quickly assess if an available MT model performs well on the client document, thus, if PE of MT in general might be suitable for the job or not. Sentence-level QE can help translators to quickly judge whether it is worth PE a particular segment or if one should fall back to translation from scratch. Word-level QE can guide the post-editor within a each sentence.

Turchi et al. (2015) were able to show that sentence-level QE (visualized by marking sentences in red/green) can indeed speed up PE, but their results were

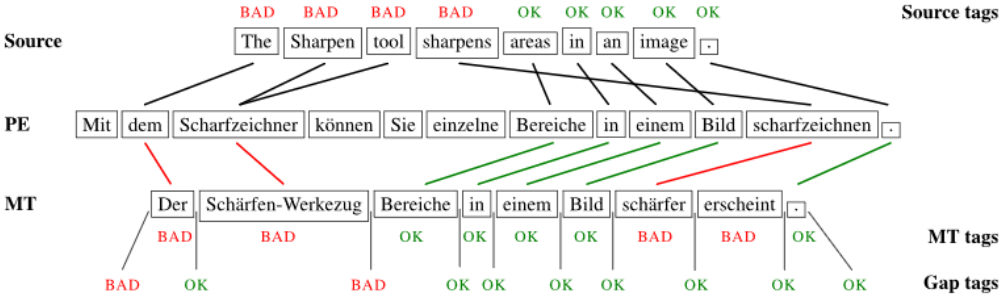


Figure 2.6: Word-level QE, with English source (top), German MT (bottom), and human PE (middle). Three types of quality tags exist: source tags for mistranslated or omitted source words, MT tags for replaced or deleted words in the MT, and gap tags for missing words in the MT (taken from Kepler et al. (2019)).

only significant for medium-length sentences (5-20 words) with a minimum quality (0.1 HTER). Using a similar traffic-light system of sentence-level QE in the Post-Editing Tool (PET, Aziz et al. (2012), see section 2.3.2), Parra Escartín et al. (2017) found that "good and accurate MT QE, is vital to the efficiency of the translation workflow, and can cut translating time and effort significantly". Teixeira and O'Brien (2017) explore the impact of expressing the scores from different sentence-level QE models as percentages between 20% and 99%. Their study with 20 professional translators indicates that just displaying sentence-level percentage scores is not enough. Instead, one should also visualize word-level QE predictions to show which parts of the MT output potentially contain errors and require special attention by highlighting words in red or green or marking gaps for missing words (Kepler et al., 2019). The IntelliCAT tool (Lee et al., 2021) offers such word-level QE as one among many features. With it, participants were indeed slightly faster, however, the differences were not significant (see section 2.3.2). Thus, further studies on word-level QE are needed to understand at what quality level it starts boosting translation performance and how it should best be visualized, which is what we do in section 5.9. This is especially interesting because wrong QE predictions could lead the human translator to assume the quality is good even when it is not, thereby speeding up the process at the cost of translation quality, or the contrary, losing time on searching for a mistake that does not exist.

Intelligibility

Another feature that has been investigated for CAT interfaces is the use of intelligibility (Coppers et al., 2018; Vandeghinste et al., 2019, 2016). The motivation is that the large variety of features offered by CAT tools makes it hard to trust a particular source, especially since the reasoning behind the chosen suggestions by the AI tools mostly remains unclear to users. Thus, visualizing justifications for the suggestions of different CAT features has been explored (see section 2.3.2).

Logging

While less important for production systems, research on CAT systems furthermore requires reliable and extensive logging functionality (Aziz et al., 2012). Most tools log mouse and keyboard input, others further track pauses in typing (Lacruz and Shreve, 2014; Lacruz et al., 2012) or eye tracking data (Alabau et al., 2013a; Carl, 2012) to get a more complete model of the user interactions. Some tools also offer a replay mode based on the log files, where researchers can view the whole translation process evolving over time (Alabau et al. (2013b), see subsubsection 2.3.2).

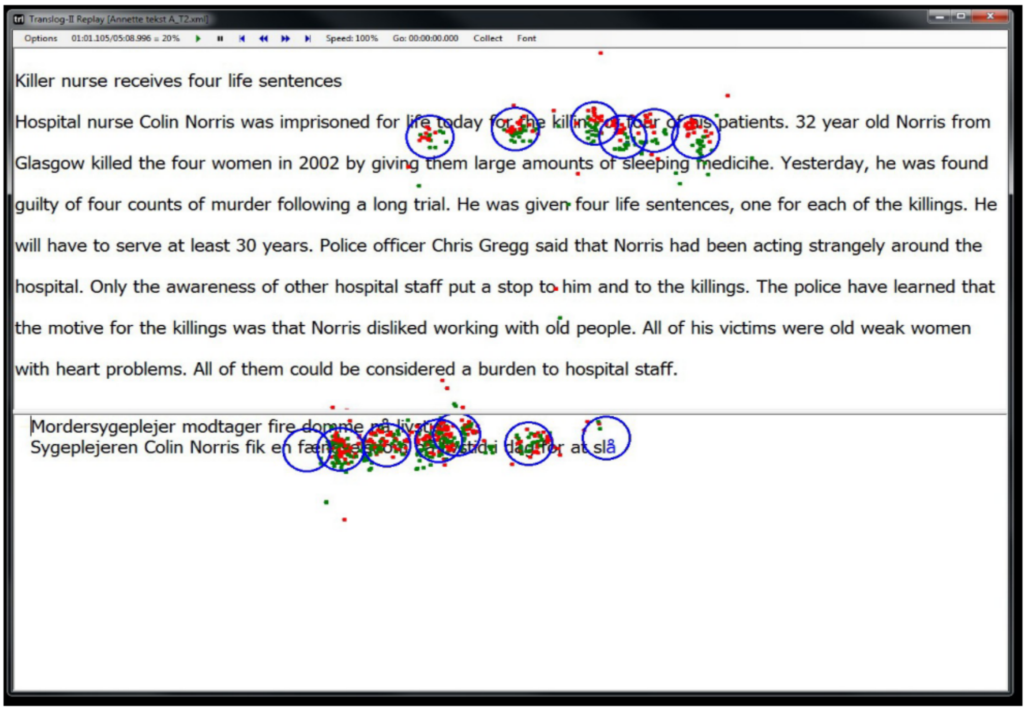
2.3.2 Research-Focused and Commercial CAT Tools

Now that we have outlined some of the features provided by CAT tools, we present concrete tools and describe their main advantages. The list provided is not intended to be complete; it first focuses on academic tools that explore specific features, and then briefly discusses the tools most widespread in industry.

Translog

Translog development started in 1995, with the idea to transform translation research by recording keystrokes and afterwards replaying them for analysis (Schou et al., 2009). The first version implemented this functionality in Microsoft DOS, however, mouse movement and clicks were not logged. Translog also allowed replay at different speeds, inspecting the source at different points in time, or viewing look-ups within the dictionary. The follow-up version, Translog 2000 then had a real user interface programmed for Windows and also recognized the mouse. In 2005 a research grant to integrate eye tracking was accepted, allowing the development of Translog 2006 using WinForms and C#, which offered Unicode and XML support, better log files, and most importantly the recording and playback of eye gazes. Since then, Translog has been used in a variety of studies on translation process research, but also for training and teaching translators.

The follow-up project, Translog-II (Carl, 2012) (see Figure 2.7), is a Windows application to "record and study reading and writing processes on a computer". As its predecessor, it can not only record sessions, but also replay them afterwards for in-depth analysis. User input is classified as insertion, deletion, navigation, copy/cut/paste, return or mouse operations. The main improvements to the earlier version lie in the direct connection to an eye tracker and better analysis of its data. In particular, Translog-II records gaze positions, fixations, and a mapping of fixations to characters on screen. Apart from replay, some statistics about text production and navigation events, as well as pauses, are automatically created. The tool also supports PE by pre-filling the target view. Both a horizontal and vertical text orientation between source and target can be configured. If alignments between source and target words are provided (by external tools), Translog-II can also visualize the translation progress by plotting the source against the target over time with fixations and typed characters.



Killer nurse receives four life sentences

Hospital nurse Colin Norris was imprisoned for life today for the killing of four of his patients. 32 year old Norris from Glasgow killed the four women in 2002 by giving them large amounts of sleeping medicine. Yesterday, he was found guilty of four counts of murder following a long trial. He was given four life sentences, one for each of the killings. He will have to serve at least 30 years. Police officer Chris Gregg said that Norris had been acting strangely around the hospital. Only the awareness of other hospital staff put a stop to him and to the killings. The police have learned that the motive for the killings was that Norris disliked working with old people. All of his victims were old weak women with heart problems. All of them could be considered a burden to hospital staff.

Mordersygeplejer modtager fire domme på livstid
Sygeplejeren Colin Norris fik en fængselsdom på livstid i dag for at slå

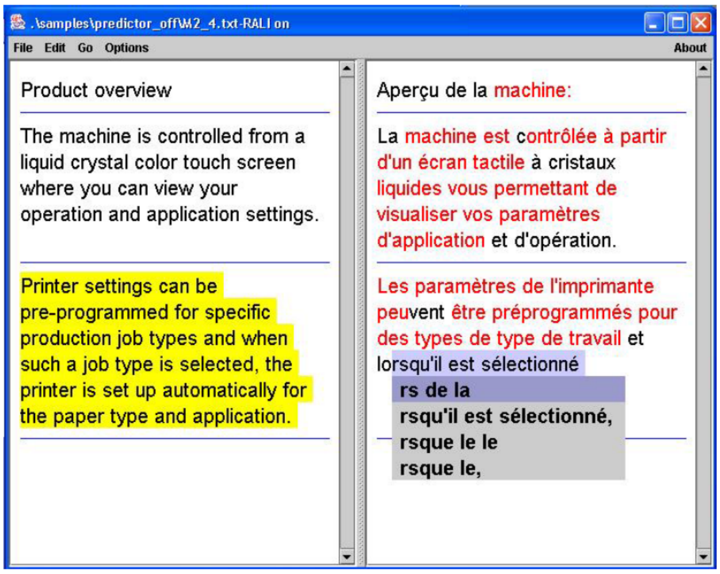
Figure 2.7: Replay mode in Translog-II, showing gaze positions for the right and left eye in red and green, and fixations as blue circles (taken from Carl (2012)).

TransType

TransType (Langlais and Lapalme, 2002) is another early project trying to improve CAT by offering a form of IMT, namely suggesting how an already started translation probably continues. The idea originated from the TransTalk project (Dymetman et al., 1994), which aimed to improve Automatic Speech Recognition (ASR) in translation by leveraging the translation model and language model. TransType follows a similar idea by using those models in a weighted fashion to create completion proposals from the vocabulary. These real-time proposals are supposed to save time by having the user type less. Up to 7 best completions are presented in a dropdown-like menu appearing at the cursor position. The user can select from this list using the mouse or page up and down keys.

Calculations assuming an ideal user who minimizes keystrokes as much as possible using the proposals show that most keystrokes can indeed be saved. A practical evaluation with users, however, showed that only 1 out of 10 participants was actually faster with TransType. A reason could be the limited time of the study for learning to optimally leverage the proposals. Another likely explanation is that users do not always watch the screen and therefore do not notice many of the proposals. Participants also stated that they would like to receive proposals beyond the single next word to make the TransType concept more usable. Interestingly, in a second study, almost all of 9 translators thought they are faster with TransType, even though productivity went down by 17%. The authors concluded that completions of less than four letters should not be displayed, as inspecting and accepting them takes longer than manual completion. Nevertheless, all except one participant were enthusiastic about the concept of TransType. A downside, as argued by some participants, might be that TransType induces a literal mode of translation, as one considers one word at a time.

TransType2 (Esteban et al., 2004) (see Figure 2.8) offers several iterative improvements, but follows the same approach of real-time completions based on translation and language models, which the authors see as a paradigm "between fully automatic MT and translation memory", combining the strength of MT with the competence of the human. If the prediction is good enough it may reduce the need to consult other CAT tools like dictionaries, term banks or TM. In TransType2, which is also built upon TransTalk, a microphone can also be used for dictation.



Product overview

The machine is controlled from a liquid crystal color touch screen where you can view your operation and application settings.

Printer settings can be pre-programmed for specific production job types and when such a job type is selected, the printer is set up automatically for the paper type and application.

Aperçu de la machine:

La machine est contrôlée à partir d'un écran tactile à cristaux liquides vous permettant de visualiser vos paramètres d'application et d'opération.

Les paramètres de l'imprimante peuvent être préprogrammés pour des types de type de travail et lorsqu'il est sélectionné

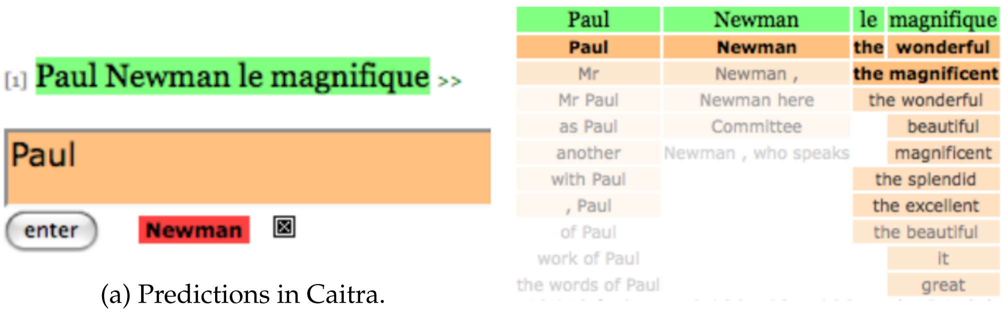
rs de la
rsqu'il est sélectionné,
rsque le le
rsque le,

Figure 2.8: TransType2 providing real-time suggestions for completion in a dropdown, and accepted suggestions in red (taken from Esteban et al. (2004)).

Caitra

Caitra (Koehn, 2009a,b; Koehn and Haddow, 2009) acts as a testbed for human MT interaction by providing 3 kinds of editing modes: (1) suggestions for sentence completion, (2) word and phrase options, and (3) normal PE.

In the prediction mode (see Figure 2.9a), the MT makes suggestions for sentence completion for the next word or phrase, and updates based on user input. Users can accept predictions using the TAB key, or just type anything else to receive new predictions. Only a few words are suggested to avoid overloading translators' reading capacity. This is implemented by matching the user input string against the MT decoding search graph which is pre-calculated and stored in a database.



[1] Paul Newman le magnifique >>

Paul

(a) Predictions in Caitra.

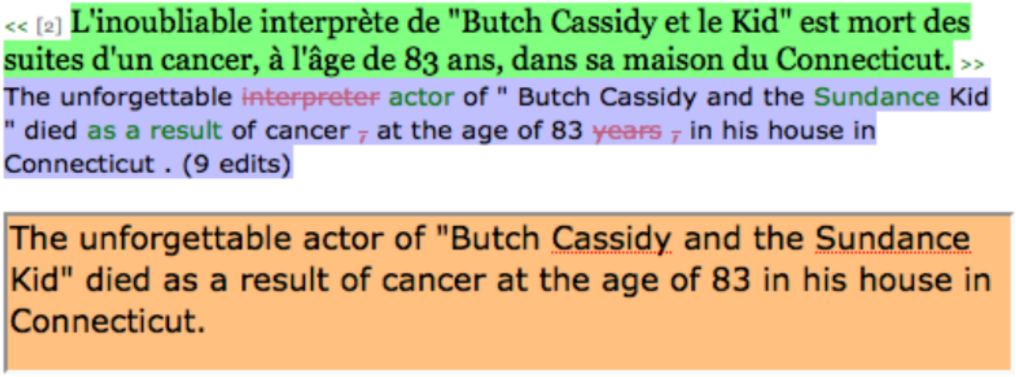
Paul	Newman	le	magnifique
Paul	Newman	the	wonderful
Mr	Newman ,	the magnificent	
Mr Paul	Newman here	the wonderful	
as Paul	Committee		beautiful
another	Newman , who speaks		magnificent
with Paul		the splendid	
, Paul		the excellent	
of Paul		the beautiful	
work of Paul			it
the words of Paul			great

(b) Options in Caitra.

Figure 2.9: Predictions and options in Caitra (taken from Koehn (2009b)).

In the options mode (see Figure 2.9b), the phrase translation tables from PBSMT are leveraged: "the most likely word and phrase translation are displayed alongside the input words, ranked and color-coded by their probability" (Koehn, 2009a). This aims to aid novice translators with unknown words and enhance the creativity of advanced users by providing alternatives to their active vocabulary.

In the PE mode (see Figure 2.10), the target is simply populated with MT output, which users can edit. As an additional feature, Caitra shows user changes above the editing field to possibly alert if content was dropped or added.



<< [2] L'inoubliable interprète de "Butch Cassidy et le Kid" est mort des suites d'un cancer, à l'âge de 83 ans, dans sa maison du Connecticut. >>

The unforgettable ~~interpreter~~ actor of " Butch Cassidy and the Sundance Kid " died as a result of cancer ~~,~~ at the age of 83 ~~years~~ ~~,~~ in his house in Connecticut . (9 edits)

The unforgettable actor of "Butch Cassidy and the Sundance Kid" died as a result of cancer at the age of 83 in his house in Connecticut.

Figure 2.10: PE in Caitra (taken from Koehn (2009b)).

For analysis of user activity, Caitra offers a graphical representation of user interaction with pauses, typing, deleting, accepting proposals, etc. A study involving 10 paid non-professional translators explored the five conditions (i) *unassisted*, (ii)

PE, (iii) *options*, (iv) *predictions*, and (v) *options combined with predictions*. Time and typing were recorded, and human judges were leveraged to assess translation quality. In general, participants were faster and achieved higher quality with assistance, however, the individual results vary: Using *PE*, 8 out of 10 participants were faster and achieved higher quality, using *options* 4 out of 10, and using either *predictions* or *predictions combined with options* 6 out of 10 performed better. Interestingly, the subjective feedback ranked the three modes almost inversely: the predictions combined with options were perceived best, followed by options, then predictions, and last PE. A simple time comparison to the unassisted condition reveals that options lead to a gain of only 16%, predictions sped up the process by 27%, the combination of the two lead to 25%, and PE to 39% performance gain. This shows that IMT is not necessarily superior to classical PE.

wikiBABEL

wikiBABEL (Kumaran et al., 2008) is a collaborative framework for community-based translation of content from a stable version in one language to other languages. As the name suggests, it targets pages like Wikipedia. To support the translation process, wikiBABEL integrates MT and other linguistic tools like bilingual dictionaries, but also focuses on supporting community-wide collaboration. Participating users register with demographic information and the quantity and quality of their contributions is tracked. The user interface shows the source and target next to each other in the original format, i.e., looking like the Wiki-webpage (see Figure 2.11). Initially, the target is filled with MT output. On mouse-over the alignment between source and target is visualized. The user can choose any sentence, click on it, and an editing box appears where changes can be made. The changes made by one user now become the new text version, so subsequent editors can edit the previous human translation. Through a rollback mechanism, they can also go back further and decide to edit an earlier version.

PET

PET (Post-Editing Tool) (Aziz et al., 2012) (see Figure 2.12) is a customizable CAT environment for PE any MT that is provided in XML format, or for translation from scratch, which focuses on collecting information about the PE process. It records the time, logs edits and keystrokes at the sentence level, and calculates the edit distance between MT and its final PE version. After each segment, PET can be configured to show an assessment window where, e.g., subjective ratings on the amount of PE required or the accuracy of the MT, can be captured. As it is intended to be used in a variety on translation and PE studies, it offers a variety of customization options like defining the text to be displayed in the assessment window, the amount of segments displayed at a time, or whether segments can be edited multiple times.

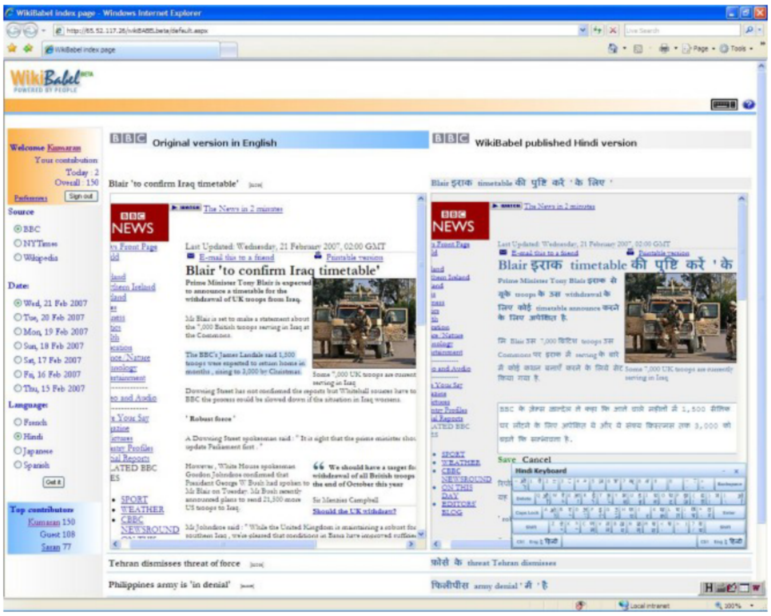

WikiBabel index page - Windows Internet Explorer
WikiBabel
Original version in English
WikiBabel published Hindi version
Blair 'to confirm Iraq timetable'
Welcome Kumaran
Source
BBC
NYTimes
Wikipedia
Date:
Language:
French
Hindi
Japanese
Spanish
Top contributors
BBC NEWS
Save Cancel
Hindi Keyboard
Tehran dismisses threat of force
Philippines army is 'in denial'

Figure 2.11: The interface of Wikibabel (taken from Kumaran et al. (2008)).

MateCAT

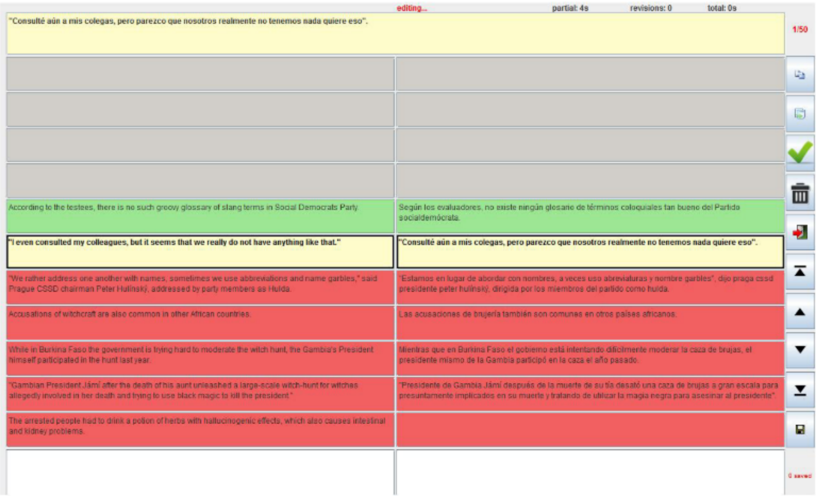
MateCAT[6] (Federico et al., 2014) is a web-based CAT tool offering several linguistic tools, including TMs, TBs, concordancers, and MT (see Figure 2.13). MT and TM proposals are provided below the editing field alongside their match score or confidence score, respectively. Concordance and glossary functionality is provided in separate tabs. MateCAT was particularly designed with the goal of investigating MT PE in mind. It provides an API to Moses or Google Translate for SMT, has project management features to create and assign translation jobs, and has an "intuitive web interface that enables the collaboration of multiple users on the same project" (Federico et al., 2014). The log files capture edits, timings, but also which translation suggestion (TM or MT) was used for PE.

CASMACAT

CASMACAT (Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation)[7] (Alabau et al. (2013a,b, 2014a), see Figure 2.14) is an open source web-based workbench for PE of MT, interactive translation prediction (ITP), visualization of word alignments, providing logging and replay, and exploring eye-tracking and e-pens. CASMACAT and MateCAT were developed in close collaboration, where MateCAT focuses on conventional CAT and CASMACAT on the user interactions.

[6] `https://www.matecat.com/`

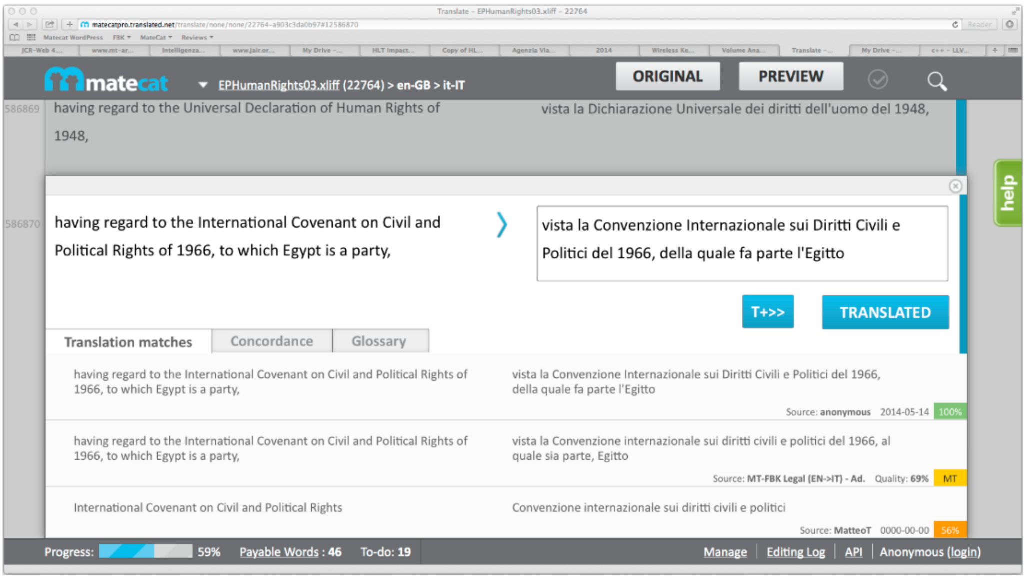
[7] `https://github.com/casmacat`



"Consulté aûn a mis colegas, pero parezco que nosotros realmente no tenemos nada quiere eso".

According to the testees, there is no such groovy glossary of slang terms in Social Democrats Party.	Según los evaluadores, no existe ningún glosario de términos coloquiales tan bueno del Partido socialdemócrata.
"I even consulted my colleagues, but it seems that we really do not have anything like that."	"Consulté aûn a mis colegas, pero parezco que nosotros realmente no tenemos nada quiere eso".
"We rather address one another with names, sometimes we use abbreviations and name garbles," said Prague CSSD chairman Peter Hulínský, addressed by party members as Hulda.	"Estamos en lugar de abordar con nombres, a veces uso abreviaturas y nombre garbles", dijo praga cssd presidente peter hulínský, dirigida por los miembros del partido como hulda.
Accusations of witchcraft are also common in other African countries.	Las acusaciones de brujería también son comunes en otros países africanos.
While in Burkina Faso the government is trying hard to moderate the witch hunt, the Gambia's President himself participated in the hunt last year.	Mientras que en Burkina Faso el gobierno está intentando difícilmente moderar la caza de brujas, el presidente mismo de la Gambia participó en la caza el año pasado.
"Gambian President Jámí after the death of his aunt unleashed a large-scale witch-hunt for witches allegedly involved in her death and trying to use black magic to kill the president."	"Presidente de Gambia Jámí después de la muerte de su tía desató una caza de brujas a gran escala para presuntamente implicados en su muerte y tratando de utilizar la magia negra para asesinar al presidente".
The arrested people had to drink a potion of herbs with hallucinogenic effects, which also causes intestinal and kidney problems.	

Figure 2.12: The interface of PET (taken from Aziz et al. (2012)).



Translate – EPHumanRights03.xliff – 22764
matecatpro.translated.net/translate/none/none/22764-a903c3da0b97#12586870
matecat
EPHumanRights03.xliff (22764) > en-GB > it-IT
ORIGINAL
PREVIEW
help
T+>>
TRANSLATED
Translation matches
Concordance
Glossary
Progress:
59%
Payable Words : 46
To-do: 19
Manage
Editing Log
API
Anonymous (login)

586869 having regard to the Universal Declaration of Human Rights of 1948, — vista la Dichiarazione Universale dei diritti dell'uomo del 1948,

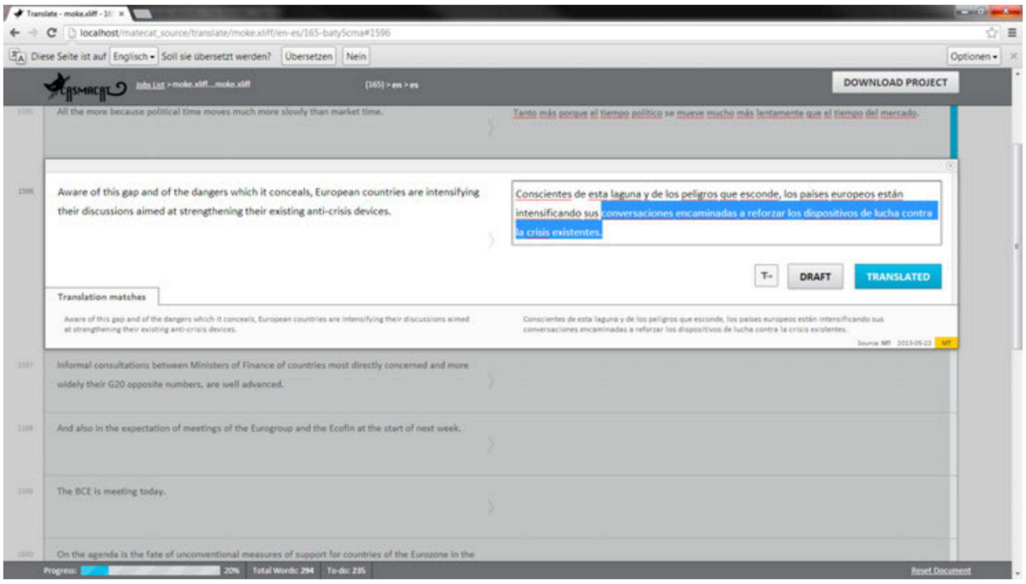
586870 having regard to the International Covenant on Civil and Political Rights of 1966, to which Egypt is a party, — vista la Convenzione Internazionale sui Diritti Civili e Politici del 1966, della quale fa parte l'Egitto

having regard to the International Covenant on Civil and Political Rights of 1966, to which Egypt is a party,	vista la Convenzione Internazionale sui Diritti Civili e Politici del 1966, della quale fa parte l'Egitto	Source: anonymous 2014-05-14 100%
having regard to the International Covenant on Civil and Political Rights of 1966, to which Egypt is a party,	vista la Convenzione internazionale sui diritti civili e politici del 1966, al quale sia parte, Egitto	Source: MT-FBK Legal (EN->IT) - Ad. Quality: 69% MT
International Covenant on Civil and Political Rights	Convenzione internazionale sui diritti civili e politici	Source: MatteoT 0000-00-00 56%

Figure 2.13: The MateCAT interface (taken from Federico et al. (2014)).

Documents to be translated are uploaded in XLIFF format. The MT server then sends the search graph alongside a list of best translations and word alignments to the client. The editing is done segment by segment, with the current segment being enlarged. As an additional support, word alignments are shown for the cursor position and for the mouse.

Overall, CASMACAT supports different editing modes: In the *normal PE* mode, up to three suggestions from MT or TM are provided to offer more options to the user. Using *Intelligent Autocompletion*, MT predictions are made after every keystroke (Barrachina et al., 2009), where the prediction for the remainder of


Translate - moke.xliff - 16
localhost/matecat_source/translate/moke.xliff/en-es/165-baty5cma#1596
Diese Seite ist auf Englisch Soll sie übersetzt werden? Übersetzen Nein
Optionen
Jobs List > moke.xliff...moke.xliff
(165) > en > es
DOWNLOAD PROJECT
T+
DRAFT
TRANSLATED
Translation matches
Source: MT 2013-05-22 MT
Progress: 20% Total Words: 294 To-do: 235
Reset Document

1595 All the more because political time moves much more slowly than market time. | Tanto más porque el tiempo político se mueve mucho más lentamente que el tiempo del mercado.

1596 Aware of this gap and of the dangers which it conceals, European countries are intensifying their discussions aimed at strengthening their existing anti-crisis devices. | Conscientes de esta laguna y de los peligros que esconde, los países europeos están intensificando sus conversaciones encaminadas a reforzar los dispositivos de lucha contra la crisis existentes.

Aware of this gap and of the dangers which it conceals, European countries are intensifying their discussions aimed at strengthening their existing anti-crisis devices. | Conscientes de esta laguna y de los peligros que esconde, los países europeos están intensificando sus conversaciones encaminadas a reforzar los dispositivos de lucha contra la crisis existentes.

1597 Informal consultations between Ministers of Finance of countries most directly concerned and more widely their G20 opposite numbers, are well advanced.

1598 And also in the expectation of meetings of the Eurogroup and the Ecofin at the start of next week.

1599 The BCE is meeting today.

1600 On the agenda is the fate of unconventional measures of support for countries of the Eurozone in the

Figure 2.14: The CASMACAT interface (taken from Alabau et al. (2013a)).

the sentence is placed to the right of the cursor. As providing and updating full predictions on every keystroke was shown to be cognitively demanding (Alabau et al., 2012), CASMACAT only shows words until the next potentially errorneous word according to confidence measures. Using the TAB key, the translator can view alternative predictions. As confidence measures, two thresholds are used: one for words likely to be incorrect (based on precision), and one for words that are dubious (based on recall). An investigation of intelligent autocompletion in comparison to traditional PE (Sanchis-Trilles et al., 2014) with nine freelance translators shows that this feature indeed slightly reduces the required keystrokes, however, is a bit slower than conventional PE. In terms of quality, the two conditions were comparable. Alves et al. (2016) also compared CASMACAT's intelligent autocompletion feature to normal PE. Based on the data of 16 participants, they find that, contrary to their expectations, normal PE is significantly faster and requires less keyboard input than translating with intelligent autocompletion. Recently, a neural version of interactive translation prediction was evaluated within CASMACAT (Knowles et al., 2019), were PE was slightly slower than interactive translation (on average 4.79 vs. 4.56 seconds per token), however, the differrence was not significant.

CASMACAT's main interface is focusing on mouse and keyboard as input modalities. However, motivated by the way handwriting is used in reviewing of human-translated text (Alabau and Leiva, 2014), CASMACAT also integrates e-pens for translation. When activating the e-pen functionality, the source is displayed above the target instead of next to it, to maximize the drawing area horizontally. The drawing is handled through MinGestures (Leiva et al., 2013), an interactive text editing approach supporting very simplistic gestures, namely lines of certain orientation and direction, combined with a handwriting recognition (when the strokes are not recognized as a drawing gesture). Unfortunately, no evaluation comparing the e-pen feature to normal editing with a keyboard was published.

Finally, an eye tracker was integrated to log the users' gaze behavior. For this, a browser plugin was developed. Note however, that it was not used for interaction, but only as a data capturing technique.

HandyCAT

HandyCAT (Hokamp, 2015; Hokamp and Liu, 2015) is an open-source CAT tool providing TM, MT, a concordancer and a glossary. It was developed within a component-centric design framework for translation interfaces, where each component can be tuned and optimized individually. An interesting view in this framework is the definition of "translation resources as functions which transform text sequences in one language to text sequences in another language", thus, comprising for example MT and TM (Hokamp, 2015). The author Chris Hokamp uses this definition to propose a two-dimensional "Translation Resource Continuum", where one dimension ranges between match-based and generation based approaches, and the other dimension between outputting a word and a whole sequence, as shown in Figure 2.15. In this continuum, a glossary and NMT are on completely opposite ends, where the former is purely match based and generates a single word, whereas the latter is a generative approach resulting in a sequence. Framing all translation resources within such a continuum, and allowing the separation into graphical elements and data services in HandyCAT, enables a quick and easy integration and optimization of individual components. HandyCAT was used for some further experiments, implementing QE and interactive PE[8].

CATaLog

CATaLog (Nayek et al., 2015) is a TM-based CAT tool focusing on color coding similarities between source sentences and TM matches to guide translators by showing parts likely requiring PE. To provide this color coding, alignment plays a crucial role: (1) the alignment between the actual source and the source of TM segments, and (2) the alignment between the TM match source and the TM match target. These two alignments are combined to figure out relevant parts of the TM match with respect to the input source sentence. The TER metric (Snover et al., 2006) is used to analyze the similarity between the input source sentence and the TM match sources and to produce alignments. Since in PE deletions can be achieved more quickly than insertions, substitutions, or reordering, the authors modify TER to weigh deletions as less severe problems than other editing

[8] https://www.youtube.com/watch?v=Abijz71Lz8Y

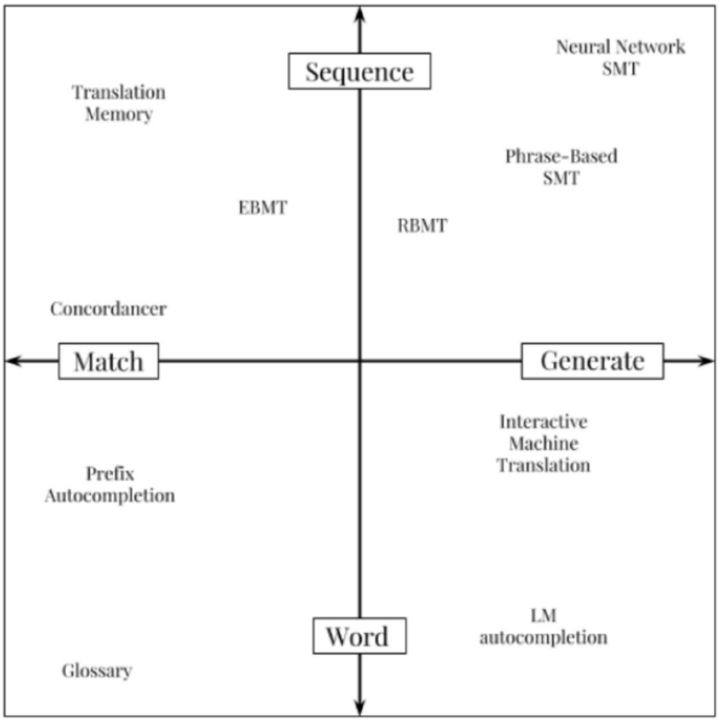
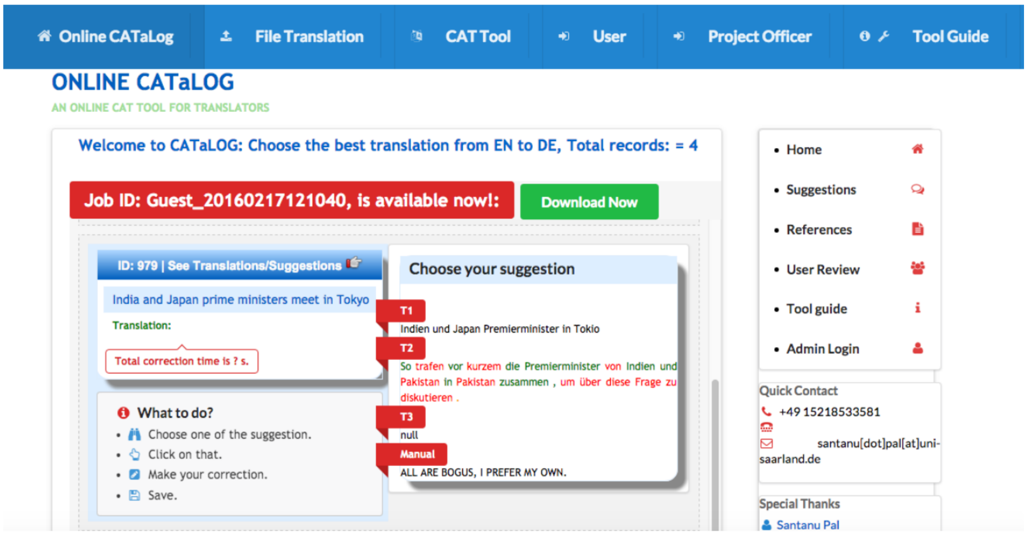


Figure 2.15: Chris Hokamp's translation resource continuum, where the x-axis ranges between match-based and generative approaches, whereas the y-axis defines the output length (taken from Hokamp (2015)).

operations. Color coding of the source sentences shows which parts of the source input are covered by a match and which are not. Color coding the target helps translators to choose among 5 provided TM matches and guides translators' editing process.

CATaLog Online (Pal et al., 2016c,e) is a web-based and extended version of CATaLog. It offers PE of TM, MT, and Automatic Post-Editing (APE) (see Figure 2.16), and offers logs of user behavior. The TM match is color coded, and upon selection of a TM match, the source is also color coded as in CATaLog (see Figure 2.17). Alignments between all kinds of translation proposals are provided: for TM, MT, APE, and human translations. Logs for deletion, insertion, substitution, and reordering are provided on character-level. CATaLog Online also supports teams of translators and project managers in organizing translation jobs. As it shows several translation aids simultaneously, it can be used to compare different MT and TM engines, where translators act as voters.

Vela et al. (2019) analyze CATaLog Online in two regards. First, CATaLog Online was compared to MateCAT in terms of PE performance in a between-subjects study with 16 translation students, showing that MateCAT was faster. Second, the quality of the 3 types of proposals (TM, MT, APE) was analyzed. Using the same participants, the authors found that the selection between MT and APE



Welcome to CATaLOG: Choose the best translation from EN to DE, Total records: = 4

Job ID: Guest_20160217121040, is available now!:

ID: 979 | See Translations/Suggestions

India and Japan prime ministers meet in Tokyo

Translation:

Total correction time is ? s.

What to do?

- Choose one of the suggestion.
- Click on that.
- Make your correction.
- Save.

Choose your suggestion

T1 Indien und Japan Premierminister in Tokio

T2 So trafen vor kurzem die Premierminister von Indien und Pakistan in Pakistan zusammen , um über diese Frage zu diskutieren .

T3 null

Manual ALL ARE BOGUS, I PREFER MY OWN.

Quick Contact
+49 15218533581
santanu[dot]pal[at]uni-saarland.de

Special Thanks
Santanu Pal

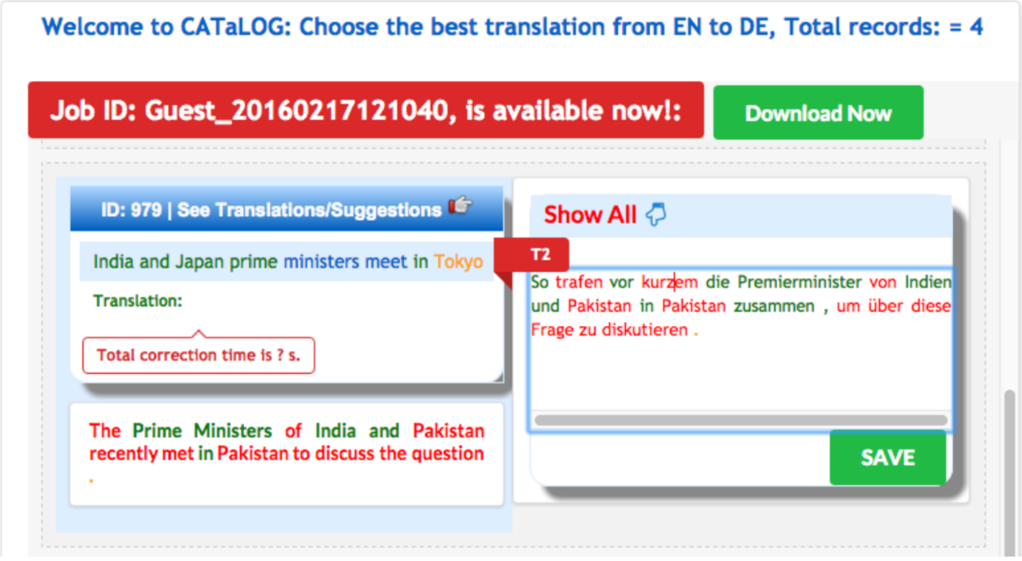
Figure 2.16: The interface of CATaLog Online provides TM, MT, and APE suggestions (taken from Pal et al. (2016e)).

was based on chance as the output is highly similar. Therefore they repeated the experiment with three professional translators choosing between MT, TM, and translation from scratch for 200 sentences. As already outlined earlier, in 80% of the cases, the participants chose MT PE, showing the importance of the concept. However, the agreement to decide for PE TM, PE MT, or translation from scratch was low, showing that there are no obvious cues that all translators use to decide which translation methodology to use. In terms of subjective feedback, CATaLog Online received positive comments on the TM color coding, the MT quality, and the arrangement of suggestions. However, the lack of spell checking, keyboard shortcuts, concordancer, as well as the overloaded interface were criticized.

Kanjingo and its successor

'Kanjingo' (O'Brien et al., 2014; Moorkens et al., 2016; Torres-Hostench et al., 2017) is an iPhone app for PE MT via touch and speech input. In the editor view (see Figure 2.18), the tool provides a machine translated target text on vertically stacked tiles, where "the user may click and drag each word tile to change word order, delete words by clicking the red '-' button, or add words using the blue '+' button" to either type with the iPhone's virtual keyboard or use ASR (Moorkens et al., 2016). The top of the screen always shows the corresponding source text.

Participants provided mostly positive feedback, especially on reordering words with touch drag and drop. An interesting feature request was a better support for moving groups of words. PE with voice worked better than participants expected, voice was preferred when translating from scratch due to low MT quality, but the iPhone keyboard was preferred when only small changes were required. Overall


Welcome to CATaLOG: Choose the best translation from EN to DE, Total records: = 4

Job ID: Guest_20160217121040, is available now!:

ID: 979 | See Translations/Suggestions

India and Japan prime ministers meet in Tokyo

Translation:

Total correction time is ? s.

The Prime Ministers of India and Pakistan recently met in Pakistan to discuss the question .

So trafen vor kurzem die Premierminister von Indien und Pakistan in Pakistan zusammen , um über diese Frage zu diskutieren .

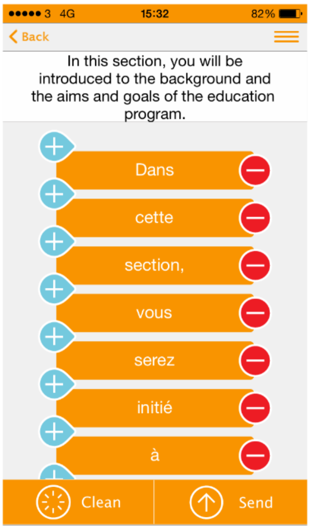
Figure 2.17: Color-coding for TM matches in CATaLog Online (taken from Pal et al. (2016e)).

the authors found that PE MT is feasible and realistic with a mobile app, although it cannot replace desktop-based PE. An example use case for the app was seen in not-for-profit projects. Missed features include the possibility to see all segments simultaneously, to integrate spell checking, or copy and paste functionality. The authors hypothesize that productivity would be slower, but the quality would be similar to desktop-based PE, however, no comparative study was conducted.

A very related CAT tool by partially the same authors was presented for the desktop setting, again exploring a combination of touch and speech, however, for translation from scratch, translation using TM, and translation using MT (Teixeira et al., 2019). Their prototype offers two main editing views: one for mouse, keyboard, and speech (see Figure 2.19, using Google ASR), and a tile view for touch (see Figure 2.20), where each word is a tile that can be dragged around for reordering, or moved on a trash bin for deletion. Besides the main editing views, functionality for comments, a lexicon, as well as global find and replace were implemented.

In their two studies of the iteratively refined prototype with a total of 18 participants, touch input received poor feedback mainly due to two factors: First, their tile view made reading more complicated. Second, touch insertions were rather complex to achieve within their implementation (click on previous tile, then at end of that word, adding a space and then the new word). These limitations were a result of the prototype focusing on touch for reordering, whereas, only few reorderings were required during PE in their study.

In contrast, integrating dictation functionality using speech was shown to be quite useful and even preferred to mouse and keyboard by half of the participants.



In this section, you will be introduced to the background and the aims and goals of the education program.

Dans

cette

section,

vous

serez

initié

à

Figure 2.18: The Kanjingo mobile application for PE MT (taken from Moorkens et al. (2016)).

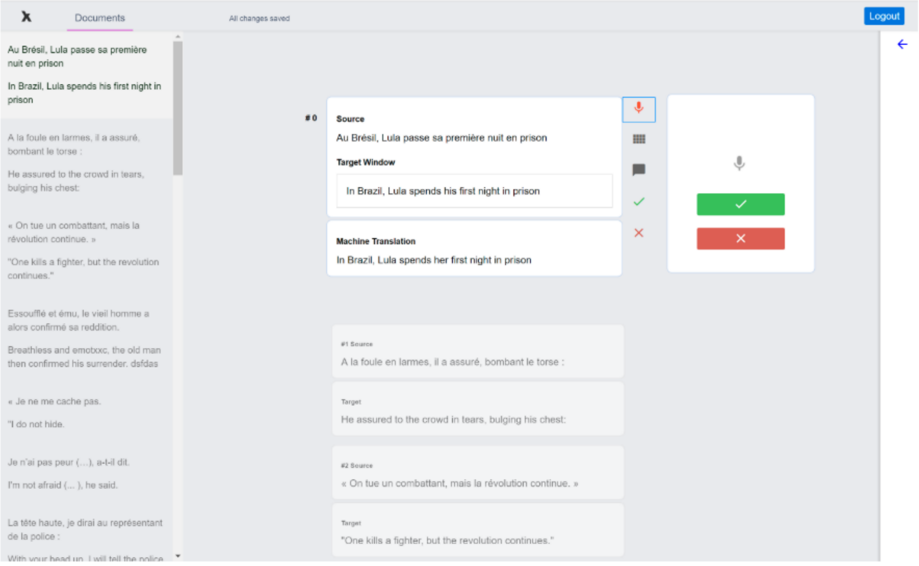
Only dictation and no voice commands were used in the study since the interface developers found that commands require too much time.

Interestingly, mouse and keyboard were slowest in their study, dictation the fastest, and touch in between. However, a quality analysis showed that the quality with mouse and keyboard was higher, indicating that the participants did not correct the MT output to the same level of quality. As in other CAT studies, participants found the lack of spell checking particularly annoying.

Intellingo/SCATE

The SCATE (Smart Computer-Aided Translation Environment) project deals with many aspects of CAT, ranging from improved fuzzy matching, parallel treebanks, integrations of TM with MT, QE, terminology extraction from comparable corpora, to the use of ASR in the translation process. The most interesting aspect for this dissertation is the CAT tool called Intellingo that was developed and tested as part of the project (Coppers et al., 2018; Vandeghinste et al., 2019). The overarching goal of Intellingo is to explore the concept of intelligibility to provide justifications for the commonly used translation aids like TM, MT, TB, or auto-completion. The authors argue that most translation aids are represented as black box systems, i.e., showing outputs of different quality without providing any justification for them.

Figure 2.21 shows the Intellingo interface and its features. As can be seen, the source is located above a text entry box for the target. Results from a hybrid MT (combining MT and TM) are shown below, using bold font for the parts stemming from TM to provide intelligibility for the output. TM is on the bottom



Au Brésil, Lula passe sa première nuit en prison

In Brazil, Lula spends his first night in prison

A la foule en larmes, il a assuré, bombant le torse :

He assured to the crowd in tears, bulging his chest:

« On tue un combattant, mais la révolution continue. »

"One kills a fighter, but the revolution continues."

Essoufflé et ému, le vieil homme a alors confirmé sa reddition.

Breathless and emotxxc, the old man then confirmed his surrender. dsfdas

« Je ne me cache pas.

"I do not hide.

Je n'ai pas peur (...), a-t-il dit.

I'm not afraid (...), he said.

La tête haute, je dirai au représentant de la police :

With your head up, I will tell the police

0

Source

Au Brésil, Lula passe sa première nuit en prison

Target Window

In Brazil, Lula spends his first night in prison

Machine Translation

In Brazil, Lula spends her first night in prison

#1 Source

A la foule en larmes, il a assuré, bombant le torse :

Target

He assured to the crowd in tears, bulging his chest:

#2 Source

« On tue un combattant, mais la révolution continue. »

Target

"One kills a fighter, but the revolution continues."

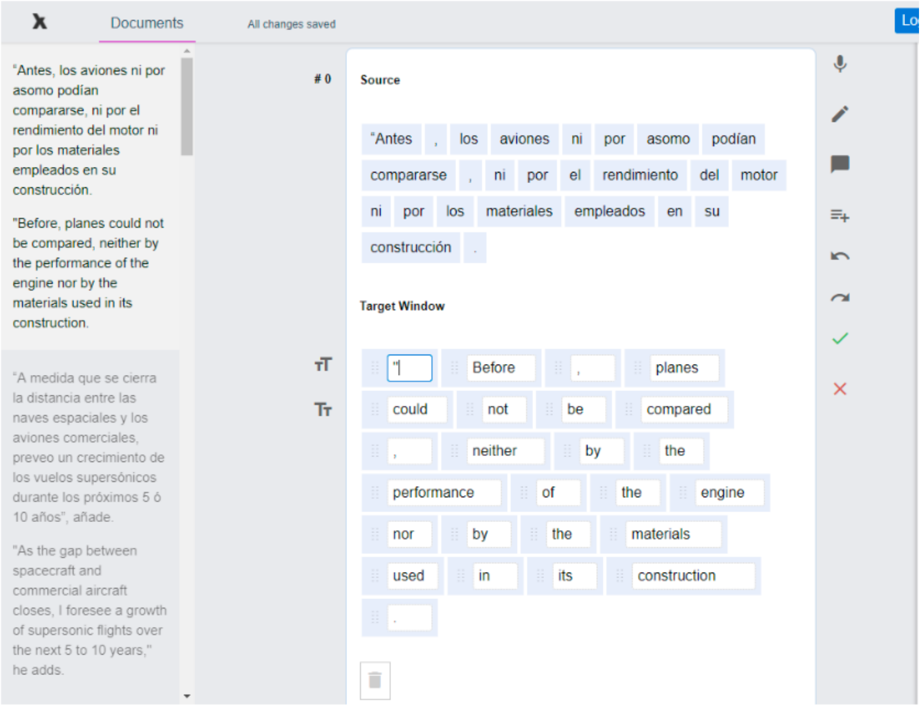
Figure 2.19: Teixara et al.'s multi-modal CAT tool: The main view for mouse, keyboard, and speech-based PE (taken from Teixeira et al. (2019)).

right of the MT, with match scores, an icon indicating the match strategy, and whether a match was used for the hybird MT, as some of the many intelligibility features. Translation alternatives are shown to the left of the MT, with icons to intelligibly indicate their source (TM, TB, MT). Occurrences of these alternatives are also shown in blue in the TM on the right, to indicate the context that they have been used in. Auto-suggestion is offered and using the alternatives and their highlights in the TM and MT, the user again receives explanations where this suggestion is coming from and why it has been made.

For the study, the authors compared an intelligible and non-intelligible version of the CAT environment. They could not find a significant effect on the user experience or time to translate, but argue that "translators can better assess translation suggestions without a negative" (albeit also not positive) impact (Coppers et al., 2018). Overall, the study shows that intelligibility is only preferred when the additional justification "benefits the translation process and is not part of the translator's readily available knowledge". Therefore, the authors advise to make the visibility of intelligbility information configurable.

IntelliCAT

The recently presented tool called IntelliCAT (Lee et al., 2021) combines sentence-level and word-level QE with interactive translation suggestions (conditioned on both the right and left context of the marked position), and word alignments to highlight the link between the source and target and to apply the source formatting on the target text (see Figure 2.22).



"Antes, los aviones ni por asomo podían compararse, ni por el rendimiento del motor ni por los materiales empleados en su construcción.

"Before, planes could not be compared, neither by the performance of the engine nor by the materials used in its construction.

"A medida que se cierra la distancia entre las naves espaciales y los aviones comerciales, preveo un crecimiento de los vuelos supersónicos durante los próximos 5 ó 10 años", añade.

"As the gap between spacecraft and commercial aircraft closes, I foresee a growth of supersonic flights over the next 5 to 10 years," he adds.

"Antes , los aviones ni por asomo podían compararse , ni por el rendimiento del motor ni por los materiales empleados en su construcción .

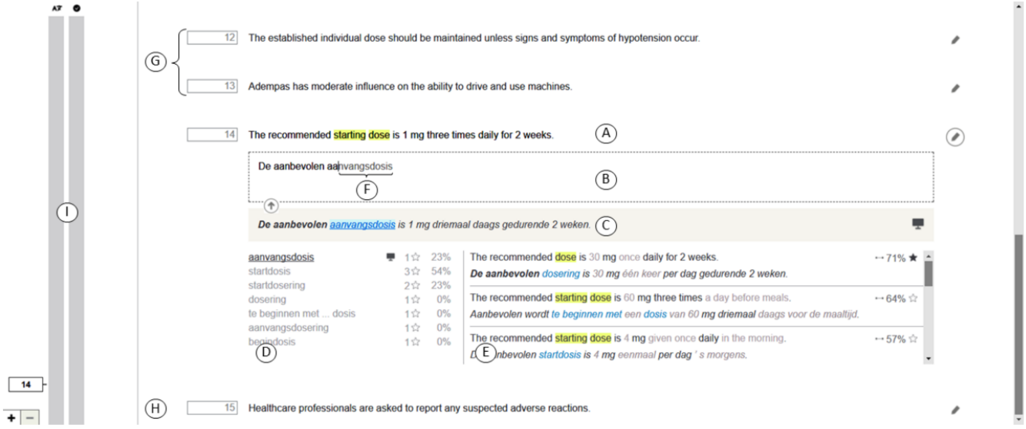
" Before , planes could not be compared , neither by the performance of the engine nor by the materials used in its construction .

Figure 2.20: Teixara et al.'s multi-modal CAT tool: The tile view for touch-based PE (taken from Teixeira et al. (2019)).

A data analysis showed that applying the top-1 suggestion on the errors highlighted by the word-level QE, which can be applied fully automatically, improves translation quality by -2.33 TER or +1.56 BLEU. Manually selecting among the top-5 suggestions can even lead to quality gains of -6.01 TER or +6.15 BLEU. An additional user study with nine participants showed that a version of the tool with QE, translation suggestions, and word alignments reduced the average editing time compared to normal PE (with text styling) from 688 to 555.66 seconds, showing that the additionally implemented features can boost productivity. However, a quality analysis of the final text was not conducted.

Commercial Tools

Apart from these tools originating in acedemia, there are naturally many commercial tools in day-to-day use by professional translators. Since only few studies exist and the features of these tools are to a large extend overlapping, we will only briefly give an overview of some of the available tools, while acknowledging that the list is incomplete.



12 The established individual dose should be maintained unless signs and symptoms of hypotension occur.

13 Adempas has moderate influence on the ability to drive and use machines.

14 The recommended starting dose is 1 mg three times daily for 2 weeks.

De aanbevolen aanvangsdosis

De aanbevolen aanvangsdosis is 1 mg driemaal daags gedurende 2 weken.

aanvangsdosis	1☆	23%
startdosis	3☆	54%
startdosering	2☆	23%
dosering	1☆	0%
te beginnen met ... dosis	1☆	0%
aanvangsdosering	1☆	0%
begindosis	1☆	0%

The recommended dose is 30 mg once daily for 2 weeks. ↔71% ★
De aanbevolen dosering is 30 mg één keer per dag gedurende 2 weken.

The recommended starting dose is 60 mg three times a day before meals. ↔64% ☆
Aanbevolen wordt te beginnen met een dosis van 60 mg driemaal daags voor de maaltijd.

The recommended starting dose is 4 mg given once daily in the morning. ↔57% ☆
De aanbevolen startdosis is 4 mg eenmaal per dag ' s morgens.

15 Healthcare professionals are asked to report any suspected adverse reactions.

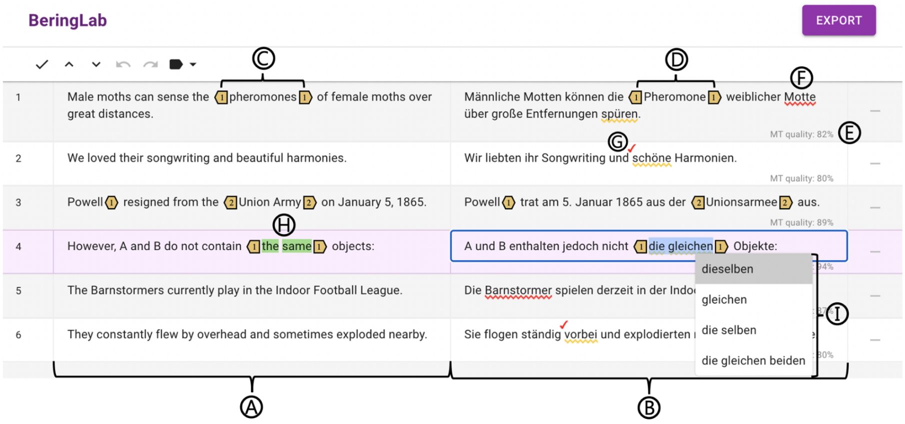
Figure 2.21: The Intellingo interface, with (A) the source, (B) the target entry box, (C) MT, (D) alternatives, (E) TM, (F) auto-completion, (G) and (H) previous and next segments, and (I) a progress bar (taken from Coppers et al. (2018)). Hovering a translation suggestion highlights its source equivalent in yellow and target equivalent in blue within all proposals.

SDL Trados Studio SDL Trados Studio[9] is the most frequently used translation environment (Moorkens and O'Brien, 2017). It has a flexible layout, i.e., allows users to reconfigure the tools to their own needs. According to SDL's own website, more than 270000 users translate within SDL Trados worldwide. SDL offers all well-known translation aids, including e.g., access to multiple MT systems (including a company-developed one), TM, QA, dictionaries, auto-suggestion, alignment. The large variety of features has also been commented negatively as it makes the tool slightly complex to use for new users (Vieira and Specia, 2011). Nowadays, Trados also combines desktop with cloud-based tools to enable working on various devices. Furthermore, it offers lots of of project management functionality, which are particularly relevant to coordinate larger projects.

memoQ memoQ[10] (see Figure 2.2) is another commercial CAT tool providing TM, MT, terminology management, QA, predictive typing (also known as auto-suggest), alignment, and other common features. It is mainly desktop-based, but also offers online work through Webtrans. Interestingly, memoQ nowadays also offers an integration for speech dictation: For this, users can download an iPhone app, dictate text which is transcribed by Apple's speech recognition, and then sent to the main memoQ interface on the computer.

[9] https://www.sdl.com/de/software-and-services/translation-software/sdl-trados-studio/

[10] https://www.memoq.com/



1	Male moths can sense the ⟨1⟩pheromones⟨1⟩ of female moths over great distances.	Männliche Motten können die ⟨1⟩Pheromone⟨1⟩ weiblicher Motte über große Entfernungen spüren. MT quality: 82%
2	We loved their songwriting and beautiful harmonies.	Wir liebten ihr Songwriting und ✓ schöne Harmonien. MT quality: 80%
3	Powell⟨1⟩ resigned from the ⟨2⟩Union Army⟨2⟩ on January 5, 1865.	Powell⟨1⟩ trat am 5. Januar 1865 aus der ⟨2⟩Unionsarmee⟨2⟩ aus. MT quality: 89%
4	However, A and B do not contain ⟨1⟩the same⟨1⟩ objects:	A und B enthalten jedoch nicht ⟨1⟩die gleichen⟨1⟩ Objekte: 94%
5	The Barnstormers currently play in the Indoor Football League.	Die Barnstormer spielen derzeit in der Indo
6	They constantly flew by overhead and sometimes exploded nearby.	Sie flogen ständig ✓ vorbei und explodierten 80%

Figure 2.22: The IntelliCAT interface, with (A) the source, (B) the target entry box, (C) formatting tags showing the source style, (D) automatically created formatting tags on the target text, (E) sentence-level QE scores, (F) and (G) word-level QE mapped to yellow or red highlighting of potential errors or a checkmark for omissions, (H) highlighting of the corresponding source position based on alignments, and (I) up to five translation alternatives (taken from Lee et al. (2021)).

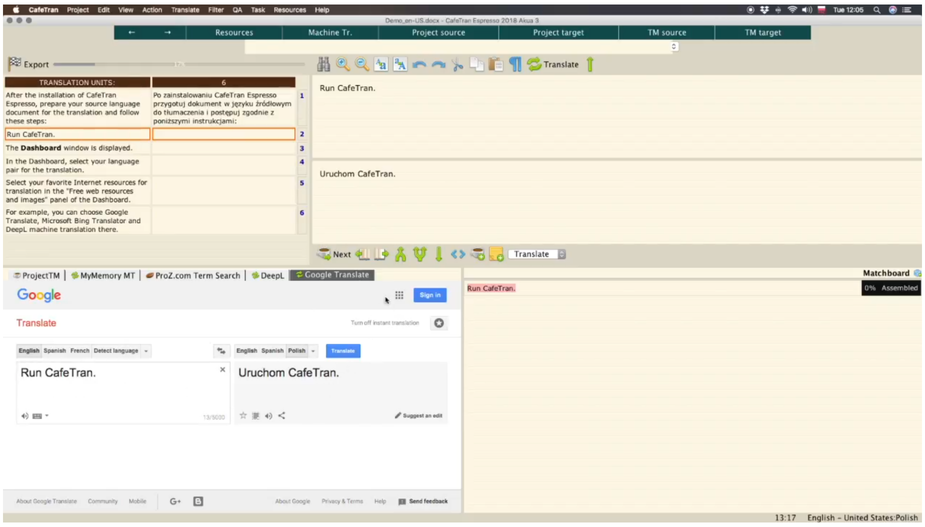
XTM Cloud XTM Cloud[11] focuses on TM technology to speed up the translation process. An interesting feature is the translation in context, showing a live preview of the translated content in the original format, which is especially useful for tasks like website translation. As most commercial tools, XTM also offers many aids for project management and workflow definition. While XTM does not provide its own MT engine, they offer connectors to many well-known MT providers by inserting API keys. Since the tool frames itself as a "Translation Management System", which offers lots of functionality like a wide range of file formats, and the ability to work with tags (for text formatting), it offers good preconditions to properly extract the text for a (third-party) MT, and afterwards QA tools can be used to ensure certain quality constraints.

Across Similarly, Across[12] is offering MT (provided by Systran), QA, TM, and terminology management, a preview of the target document in correct formatting, and a normal two-column segment-based CAT editor. Naturally, it also offers project management and workflow management functionality.

Wordbee As a professional web-based CAT tool, Wordbee offers lots of functionality for project management, receipt creation etc. However, it also supports collaboration on translation projects, TM, MT, QA, auto-suggest, and integration of content storage. Similar to Across, Wordbee also offers context-based translation, e.g., for translating websites.

[11] `https://xtm.cloud`
[12] `https://www.across.net/`

CafeTran CafeTran Espresso[13] is a freelancer oriented CAT tool, that is interoperable with all other major CAT environments. It supports a wide range of file formats, offers TM and MT integration, QA, auto completion, and a customizable interface. Google Translate, DeepL, and other common online resources are integrated by directly loading the corresponding website in the bottom left quarter of the interface, and having functionality to paste text from the main CAT tool into these websites and from there back to the main CAT tool. A screenshot of the interface with Google Translate integration can be found in Figure 2.23.



Demo_en-US.docx - CafeTran Espresso 2018 Akua 3

TRANSLATION UNITS:	6	
After the installation of CafeTran Espresso, prepare your source language document for the translation and follow these steps:	Po zainstalowaniu CafeTran Espresso przygotuj dokument w języku źródłowym do tłumaczenia i postępuj zgodnie z poniższymi instrukcjami:	1
Run CafeTran.		2
The **Dashboard** window is displayed.		3
In the Dashboard, select your language pair for the translation.		4
Select your favorite Internet resources for translation in the "Free web resources and images" panel of the Dashboard.		5
For example, you can choose Google Translate, Microsoft Bing Translator and DeepL machine translation there.		6

Run CafeTran.

Uruchom CafeTran.

Google Translate: English → Polish

Run CafeTran.

Uruchom CafeTran.

Matchboard: Run CafeTran. — 0% Assembled

13:17 English – United States:Polish

Figure 2.23: The CafeTran interface with Google Translate integration (taken from the "First Project with Resources Tutorial" on the CafeTran website[14]).

Déjà Vu Déjà Vu[15] dates back to 1993 and claims to be one of the most user-friendly tools. Naturally it supports TM (as the name suggests) and MT technologies. What is interesting is the voice command functionality using Dragon ASR. Live preview of Microsoft Office documents is also available. For TM functionality, type-ahead or how they call it "AutoWrite" functionality is implemented. Using fuzzy match repair, unmatched parts can be automatically fixed using the TB, statistical patterns in the databases, or MT.

OmegaT OmegaT (Welsh and Prior, 2014)[16] is a Java-based open source CAT tool financed through donations. It supports many file formats, including Microsoft Office, Open Office, and HTML pages. To integrate TM technology, the main editing view is separated into an editing pane, a fuzzy match pane and a

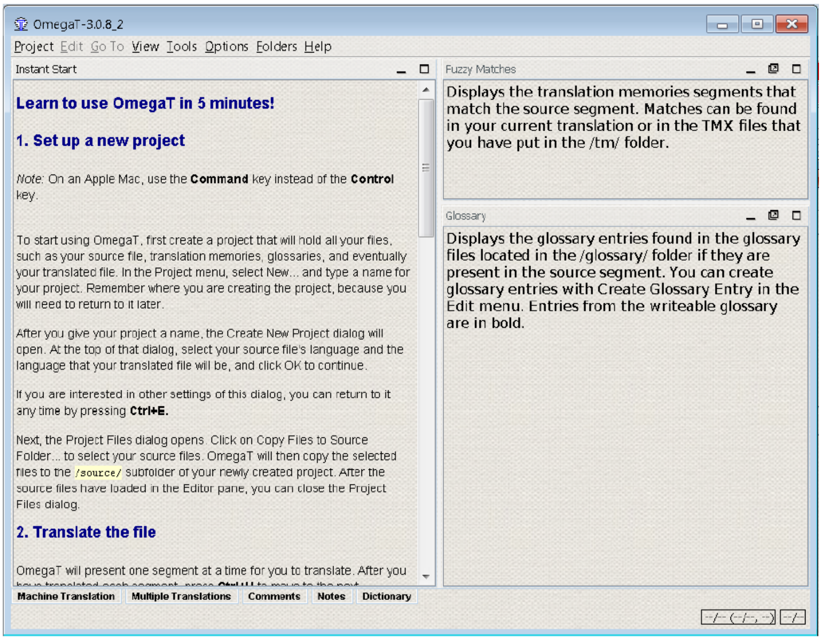
[13] https://www.cafetran.com/

[14] https://cafetran.freshdesk.com/support/solutions/articles/6000108186-first-project-with-resources

[15] https://atril.com/

[16] https://omegat.org/

glossary pane. In separate tabs, the user can view MT, further translations, notes, comments, and a dictionary. Figure 2.24 shows the interface. While translators are still expected to translate one segment at a time within OmegaT, the CAT tool does not follow the traditional approach of showing source and target side by side or one above the other. Instead, all untranslated text is shown in the source language, all already translated text in the target language, and only the currently edited segment is shown in both languages. To allow easier editing of tagged content, e.g., HTML pages, OmegaT by default copies the source into the target, so that the translator can keep the tags and translate the text within them. Fuzzy matches can also be pasted into the target using hotkeys.



Learn to use OmegaT in 5 minutes!

1. Set up a new project

Note: On an Apple Mac, use the **Command** key instead of the **Control** key.

To start using OmegaT, first create a project that will hold all your files, such as your source file, translation memories, glossaries, and eventually your translated file. In the Project menu, select New... and type a name for your project. Remember where you are creating the project, because you will need to return to it later.

After you give your project a name, the Create New Project dialog will open. At the top of that dialog, select your source file's language and the language that your translated file will be, and click OK to continue.

If you are interested in other settings of this dialog, you can return to it any time by pressing **Ctrl+E.**

Next, the Project Files dialog opens. Click on Copy Files to Source Folder... to select your source files. OmegaT will then copy the selected files to the `/source/` subfolder of your newly created project. After the source files have loaded in the Editor pane, you can close the Project Files dialog.

2. Translate the file

OmegaT will present one segment at a time for you to translate. After you

Displays the translation memories segments that match the source segment. Matches can be found in your current translation or in the TMX files that you have put in the /tm/ folder.

Displays the glossary entries found in the glossary files located in the /glossary/ folder if they are present in the source segment. You can create glossary entries with Create Glossary Entry in the Edit menu. Entries from the writeable glossary are in bold.

Figure 2.24: The OmegaT interface (taken from Welsh and Prior (2014)).

As an open source tool, it has been extended for other needs: iOmegaT for example integrated timing measurements to be able to compare translation speed gained by using MT PE in comparison to traditional translation.

Wordfast The original version of Wordfast[17], called Wordfast Classic, did not have an own UI, but instead was developed as a macro-based plugin for Micorosft Word. While translating, the tool shows each sentence in source and target language with a delimiter between them. It also offers TM functionality and keyboard shortcuts for editing segments one after the other. In the end, the document is cleaned up to contain only one language again. If no TM match is available, MT can also be used.

[17] https://www.wordfast.net/

Wordfast Pro[18] in contrast is a standalone Java-based project with the common 2 column spreadsheet-like layout, editing segment after segment outside of its original document. It offers a target preview in the correct format and supports multilingual translation projects (translating into several target languages at the same time). Wordfast Pro focuses on TM technology, where multiple TMs can be integrated and prioritized. However, MT is also integrated for segments where no good matches are retrieved from TM. The QA functionality checks spelling, grammar, terminology consistency, numbers, and punctuation marks.

Finally, Wordfast Anywhere is a browser based version of Wordfast, allowing translators to work from anywhere.

Lilt Lilt[19] is a commercial tool built upon the studies by Green et al. (2014b), which investigated a translator user interface focusing on IMT offering many assistive features like source word lookup, source coverage, auto-completion suggestions, and full completion proposals. Translators using such interactive proposals were shown to be slightly slower, however, produce slightly higher quality translations (Green et al., 2014a,b).

Lilt is based upon these publications and aims for efficient human machine collaboration for translation through interactive MT/TM proposals. It thus does not follow the classical PE pattern, where the machine acts first and the human second to clean up the machine output. Instead, the target field is initially empty, and the best MT proposal or TM match is shown below. When the user types, this single proposal is continuously updated based on the partial translation. In contrast to other commercial CAT tools, it therefore puts MT and TM at the center of translation and interactively adapts to user actions. Lilt has a vertical source target layout with a more prominent current segment, which appears less overloaded compared to most well-known CAT tools (Daems and Macken, 2019). As additional features, Lilt offers training the MT on the provided TM, pasting source or machine proposal into the editing field for classical PE, find and replace, QA, and a glossary with concordance functionality (Daems and Macken, 2019). It further provides access to style guides, TBs, and offers in-document comments for collaborative work, and many project management features.

Google Translate, DeepL, & Systran While Google Translate[20], DeepL[21], Systran[22], and similar tools cannot be considered fully-fledged CAT tools, translators and non-translators alike use them. Translators working within professional CAT environments can either just attach these MT services in their environment or use them as non-professionals would, i.e., by simply going to the respective

[18] https://www.wordfast.com/products/wordfast_pro_5

[19] https://lilt.com/

[20] https://translate.google.com/

[21] https://www.deepl.com/translator

[22] https://translate.systran.net/translationTools/

website to use the MT for gaining inspiration. These web-based interfaces have the source on the left and the target on the right. There is no segmentation, as the tools are frequently used for small texts of one or few sentences. After pasting or typing text in the source field, the language is automatically determined (if not specified in advance) and the MT kicks in to populate the target text field with the best MT proposal.

Differences in detail exist between the tools: Google gives more than one translation, and shows a dictionary and corpora lookup on double click. DeepL offers not only standard PE but also shows alternatives when clicking on a target word (see Figure 2.5). The selection further triggers the company's bilingual concordancer Linguee and displays its results below the target text. Users can also maintain a glossary to pre-define how certain terms should be translated and automatically inflects them. Furthermore, the user can choose between a formal and non-formal translation style (in the pro version only). Systran goes a step further in this regard and offers multiple proposals from different MT engines with distinct styles (e.g., generic, finance, IT). Apart from this, the Systram website is also kept rather simple, providing only the two editing fields as well as dictionary lookup.

Distribution among Translators According to a web-based survey by Coppers et al. (2018) (181 respondents, 72.38% freelancers), the most commonly used CAT tool is SDL Trados, followed by memoQ, CafeTran, and XTM. Furthermore, participants of the main study by Coppers et al. (2018) (26 translators) were using Trados, MemoQ, Across, CafeTran and Wordfast in daily live. In contrast, Moorkens and O'Brien (2013; 2014) report on a survey with 231 complete responses by translators, where by far the most common tool was SDL Trados, followed by Microsoft Word, and then Wordfast. According to a more recent survey by the same authors (Moorkens and O'Brien, 2017), 63% use more than one CAT tool, 72% of freelancers use Trados, while company translators use Trados in 84% of the cases. Interestingly, 38% still use Microsoft Word for PE, even though it cannot be considered an actual CAT tool.

2.3.3 User Interface Needs

Even though PE is becoming more widespread, most CAT interfaces still look very similar to interfaces used for translation from scratch and simply integrate MT as an additional feature. Therefore, O'Brien and Moorkens (2014) argue that CAT tools are not properly designed for PE of MT; it remains an interesting research question how the interfaces can be adapted to further facilitate this changed task. Another problem is the **limited involvement of end-users in the design process** of CAT tools (Kenny, 2017), or, as Lagoudaki (2006) puts it, that translators are usually "invited to provide feedback on an almost finished product with limited possibilities for change". O'Brien et al. (2010) argue that "UI design has [...] been driven by the needs of the translation client and not by the needs of the translator, the ultimate user of the software in question". Overall, an assessment of user interface needs "highlights a lack of HCI input in translation tool development and design, and would suggest a real need for input from HCI experts" (Moorkens and O'Brien, 2017). Similarly, the MT research community mostly focuses on developing supporting systems and their evaluation (Moorkens and O'Brien, 2017), instead of designing and building whole CAT tools together with users from the bottom up. E.g., most TM research was motivated by "technical improvement of the TM system and not how the TM system can best meet the needs of its users" (Lagoudaki, 2009b), a similar trend to what we now see with MT, which improves constantly but lacks research on a better integration in the translation process. Moorkens and O'Brien (2017) thus argue that "systems' usability and end-users' demands seem to have been only of subordinate interest" and that user-centric design has not been common practice.

A study by Koskinen and Ruokonen (2017) explores the **emotions** expressed by translators towards the technology they use, e.g., MT and TM, but also "standard tools" like Google search. For this, they ask participants to write letters to the technology and cluster them as love letters, ambivalent letters, and break-up letters. The results show that technology is indeed a central and mostly positive aspect of translators' work. Furthermore, the results do not support the notion of translators being averse to technology and further show no evidence of a generation gap. On the contrary, translation students were actually the ones with the most skepticism towards technology, probably due to their lack of experience. What translators do dislike, however, is non-functioning technology or poor usability, especially where efficiency and productivity are important. Thus, they are probably willing to adopt new software if it makes their work more efficient.

Various surveys and field studies investigated the **usefulness of different CAT features**: Semi-structured interviews with 9 translators revealed that the most important aspects of CAT environments are the ease of use, followed by speed of performance and TM as well as terminology management (Coppers et al., 2018). Half of Moorkens and O'Brien's 231 survey participants reported being dissatisfied with their tools, often due to performance or layout and compatibility issues (Moorkens and O'Brien, 2013, 2017; O'Brien and Moorkens, 2014). Other participants are however entirely happy with their tools, pointing to features like auto-propagation, quality assurance, and concordance search. As feature requests, their participants mostly want better display of meta-data, like the source of TM matches and the project they belong to, or wished for improved glossaries. 70% asked for seeing their performance (e.g., words per hour), but only for themselves and in such a way that it can be turned off, while 48% want to see dynamic reporting of their earnings (Moorkens and O'Brien, 2017). The follow-up interviews in Moorkens and O'Brien (2017) revealed more detailed ideas, like a participant dreaming of a feature to change the word order (for English-Portuguese), while another wanted to drag and drop words around. As discussed in Part II, our multi-modal CAT environment supports exactly that.

Half of the 231 respondents in Moorkens and O'Brien (2017) are unhappy with the default **layout**, coloring, and display of mark-up. 66% state that they rather customize their tools than using the default editor set-up, which, however, contradicts the online survey results by Coppers et al. (2018) stating that customization is rarely done. Many participants also mentioned performance issues, poor layout or visibility, outdated UIs, too many updates, formatting problems, as well as high learning curves for CAT tools, which might be a result of 57% requesting a clean and uncluttered UI. Vieira and Specia (2011) found that translators value on-the-fly highlighting of word alignment between source and target. Furthermore, translators want all information immediately when typing without navigation or shifting between areas (Lagoudaki, 2009a).

Keyboard shortcuts are considered to increase productivity for 80% of Moorkens and O'Brien's (2017) survey participants, where 40% claim to use them very often, while 29% use them often. 62% would like to have an additional shortcut for changing the capitalization. Many other keyboard shortcuts are also discussed in the study, showing the overall trend of a strong mouse and keyboard focus within translation tools.

Interestingly most negative comments in Moorkens and O'Brien (2013; 2014; 2017) were made with regards to **TM** tools, even though the technology is from 1990s and should have overcome initial issues: 56% report that they like TM and 75% believe it to preserve consistency and increase productivity.

Overall, the participants in Moorkens and O'Brien (2013; 2014; 2017) were critical regarding **MT**, where 56% said that MT was still problematic as it was "still in baby shoes" or that it is "just horrible", however, one should take these comments with care, as MT has improved tremendously since this study. Similarly, only 18% liked MT and only 30% responded that it supports their work. There were also many mentions of being unhappy with either the MT quality or the MT support within the CAT tool. In terms of visualization, 70% would like to see MT confidence measures as a percentage, while 25% would prefer color coding.

Comparing MT and TM, 88% of survey participants would still want to see both TM and MT, even in cases where the MT confidence score is higher than a fuzzy match score from TM (Moorkens and O'Brien, 2017). Interestingly, for cases where no TM is available, 80% would like to see the MT directly inserted into the target editor, showing that PE MT is important even though they mainly disliked MT. The threshold for when to use MT over TM is surprisingly low, for many participants a 65% fuzzy match is still preferred over MT, although research shows that it requires more effort to PE such matches than MT (Guerberof, 2008). For integration of TM and MT, methods like copy & paste or drag-and-drop were requested in Vandeghinste et al. (2019).

71% of participants in Moorkens and O'Brien (2013; 2014; 2017) further requested **dynamic changes of MT** proposals based on edits, which is especially important as participants do not like repeatedly fixing the same MT error. We will discuss approaches to quickly learn from post-edits in Part IV. 48% of survey participants further were positive regarding IMT which proposes the remainder of the sentence and adapts to changes. Vandeghinste et al. (2019) agree, stating that translators prefer IMT to classical PE, however, time measures showed that it was slower (Green et al., 2014b).

2.3.4 Multi-Modal Computer-Aided Translation

In traditional translation from scratch, which mostly encompasses text production in the form of writing, focusing on mouse and keyboard as input modalities seems like a suitable approach. However, the availability of high-quality MT, or even multiple MT proposals from different engines, and the resulting switch to PE results in major changes in translation workflows (Zaretskaya and Seghiri, 2018). In particular, PE changes the interaction pattern: While traditional translation consists of gisting, drafting, and revising phases (Carl et al., 2010), these phases are interleaved in PE, leading to a significantly reduced amount of mouse and keyboard events (Green et al., 2013). The task thus changes from mostly text production to comparing and adapting MT and TM proposals, or put differently, from control to supervision. Based on this change, we hypothesize that translation environments need not only be changed on a visualization and tool level, but that stronger changes in terms of using modalities different than the current standard mouse and keyboard could facilitate operations that are more important in PE, such as reordering or replacements. We thus propose to explore multi-modal input for PE. A very general definition of multi-modal input is provided in the Handbook of Multimodal-Multisensor Interfaces by Oviatt et al. (2017):

> "Multimodal input involves user input and processing of two or more modalities [...]. The input may involve recognition-based technologies (e.g., speech, gesture), simpler discrete input (e.g., keyboard, touch), or sensor-based information (e.g., acceleration, pressure) [...]. They are far more flexible and expressively powerful than past keyboard-and-mouse interfaces, which are limited to discrete input."

In such multi-modal interfaces, modality fusion (Lalanne et al., 2009) guarantees that modalities can be used in parallel or sequentially without losing information during the interaction. While the use of modalities other than mouse and keyboard within CAT tools opens up lots of potential for further research, some studies already investigated different input modalities, which we will discuss in the following sections.

Speech Input

Dictating translations dates back to the time when secretaries transcribed the dictaphone content on a typewriter (Theologitis, 1998). Nowadays automatic transcription is gaining popularity, where in 2016 around 15% of translators stated to use ASR technology in their work (European Commission representation in the UK and the Institute of Translation and Interpreting, 2016). An early approach to integrate ASR into the translation process is the TransTalk project (Brousseau et al., 1995; Dymetman et al., 1994). It incorporates MT output in the speech recognition process to increase the recognition performance, especially for lexical choice. Thus, the MT is not shown to the translator for PE, but only used system internally. In the TransType2 interface (see Figure 2.8), which also builds upon TransTalk, a microphone can be used for dictation. Similarly, Khadivi and Ney (2008) explore different approaches like N-best rescoring and word graph rescoring to enhance ASR recognition accuracy in a CAT setting, where apart from the target language speech one can rely on the source language text. The same idea was used within the SCATE project (Vandeghinste et al., 2019) (see subsection 2.3.2) where the accuracy of ASR was boosted "by making use of the extra information present in the translation model and by adapting the language model to the current domain or topic". ASR usually does not yield any punctuation, so SCATE further integrated punctuation prediction on the ASR output to make it more suitable for translation interfaces.

SEECAT (Martinez et al., 2014) integrates speech input into the CASMACAT workbench (Alabau et al., 2013a) (see Figure 2.14) for PE, allowing the user to press a button and start dictating text. The text is then placed at the cursor position. If a text range is selected, the dictation replaces the text span. Similar to TransTalk and SCATE, the MT hypothesis is used to rescore the ASR hypothesis, thereby aiming to improve ASR accuracy. A pilot study with two participants showed that PE with both ASR and typing was faster than PE or translation from scratch using only typing or only ASR. In a follow-up study with 10 professional translators, PE with typing-only and PE with typing and ASR were compared: 6 participants were faster when using ASR, while 4 were not. All reported that ASR is a promising feature, but that it might take time to learn. The authors thus argue that speech recognition combined with typing could boost productivity and that dictation might especially be useful when more text requires changes.

Mesa-Lao (2014) surveyed PE trainees before and after introducing speech technology to CAT. The results suggest positive views on ASR, where participants stated that "they would consider adopting [it] as an input method". The use of ASR turned out easier and more effective than previously thought, involving less fatigue. However, participants would "use ASR as a complement rather than a substitute" to classical methods.

Zapata et al. (2017) investigates effects on productivity and translation experience of four conditions: Translation Dictation (TD; which means dictate, then transcribe, then PE the transcription); PE in dictation mode (which means dictating approved sentences into a dictaphone, followed by manual transcription); TD with ASR; PE with ASR. All texts were printed-out and presented in hard copy, thus, no real CAT interface was used. In terms of time, TD with manual transcription was slower than TD with ASR in 3 out of 4 cases, as the transcription time counted for this analysis. For PE, the use of ASR was faster than manual transcription because revision of ASR is faster than the transcription process. No clear trend between translation dictation and PE dictation (both with manual transcription) was found. However, there was a big mismatch between what was most productive and what felt most productive: Most participants felt that they were faster when using ASR for translation from scratch than when using ASR for PE, while the contrary was true in actual time measures. Thus, the small-scale study showed that the use of ASR in comparison to manual transcription speeds up both translation from scratch and PE, even though the revision phases take longer in the ASR conditions.

Dragsted et al. (2011) compared written translations to sight translation (reading the source and speaking the translation in the target language) with and without ASR. In their study with translation and interpreting students, written translation was slowest, sight translation without manual transcription the fastest, and sight translation with ASR in the middle but closer to written than to sight translation. The quality, as judged by 3 evaluators, was highest with 3.2/5 for written, 2.8 for ASR, and 2.7 for sight translation. The majority of speech recognition problems was caused by the student's mispronounciations, as they translated into their second language.

Studies with the Kanjingo app (O'Brien et al., 2014; Torres-Hostench et al., 2017) (see Figure 2.18) show that PE with voice worked better than participants expected, that voice was preferred to the iPhone's virtual keyboard when translating from scratch due to a low MT quality, but that the keyboard was preferred for small changes.

Teixeira et al. (2019) (see subsection 2.3.2) showed that dictation functionality was quite good and its combination with mouse and keyboard even preferred to mouse and keyboard alone by 4 of the 8 participating professional translators. One should note that only dictation and no voice commands were used in the study since the interface developers found that commands require too much time, and therefore only added a dictation functionality. As an ASR engine, Google Voice was used. In their study, mouse and keyboard took longest, followed by touch, while dictation was the fastest. However, a quality evaluation showed that participants did not correct to the same level of quality using the different modalities, indicating that they traded-off quality for time.

After several decades of research showing that in some circumstances voice input can be beneficial to the translation process, professional CAT tools like memoQ and MateCat are now also integrating ASR: memoQ does so in combination with an iPhone App that uses Apple's speech services and focuses mainly on operating the interface with commands (undo/redo/copy source to target/copy translation suggestion to target, etc.) as well as dictation[23]. MateCat uses the Google ASR through the browser, however, it can only insert dictated text at the end of the target, does not allow placing text at different cursor positions or replacing words with it, and does not support dictating capitalization or punctuation marks.

[23] https://blog.memoq.com/hey-memoq-frequently-asked-questions

In contrast to these works, our CAT tool presented in Part II supports speech commands to edit MT output directly during PE. Furthermore, it supports simplified speech commands when text or positions are already defined through a selection process based on other interaction modalities, thereby combining the strengths of exact positioning of mouse/touch/pen with the speed of uttering speech commands.

Pen Input

Compared to speech input, pen input has received way less scientific attention. An early approach by Vo and Waibel (1993) combines 8 hand-drawn gestures (e.g., circle, opening parenthesis) with a speech recognizer detecting 11 keywords (e.g., delete, word, line) for text manipulation. Even though the set of recognized entities is limited, the combination of speech and pen offers a vast amount of flexibility: For example, a user can encircle a word and say "delete", or mark a paragraph, speak "move", and mark the destination. The vision paper by Alabau and Casacuberta (2012) proposes to rely solely on handwriting with e-pens for PE sentences with few errors in place, showcases symbols that are common for proofreading, and discusses how these could be used for insertions, deletions, transpositions, etc. Within the CASMACAT tool (Alabau et al. (2013a,b), see Figure 2.14) users can activate the e-pen mode, where the target is placed below the source to maximize the width available for drawing. CASMACAT relies on MinGestures (Leiva et al. (2013), see Figure 2.25), an interactive text editing approach supporting straight lines of different angles and directions as gestures. Everything not recognized as a drawing gesture is sent to a server for handwriting recognition. However, no formal evaluation of the e-pen feature within CASMACAT was conducted. A follow-up paper (Alabau et al., 2014b) discusses how the accuracy of handwriting recognition could be improved by constraining the search based on the MT hypothesis, the source, or the already translated sentence, which is a similar approach to the ASR recognition improvements discussed in TransTalk (Brousseau et al., 1995; Dymetman et al., 1994).

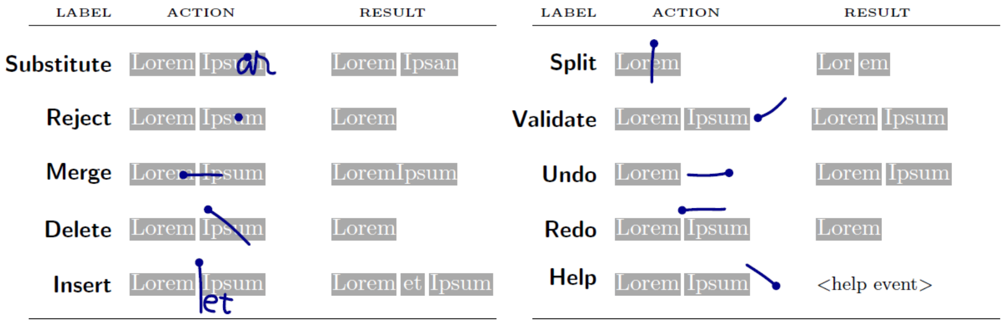


Figure 2.25: MinGestures (Leiva et al., 2013) for interactive text editing in CASMACAT (taken from Alabau et al. (2013b)).

Zapata (2016) proposes interactive translation dictation (ITD) which involves interacting with voice, touch, and pen enabled devices, such as touch screen computers or tablets. However, the system that was informally evaluated consisted simply of a touchscreen computer for translation within Microsoft Word and a (unconnected) tablet as a second screen for research. So while in theory participants could use touch and stylus, the interfaces were not designed for this purpose and no actual CAT tool was used. This could also explain why participants were slower with the ITD approach than with mouse and keyboard. Also, the study did not involve professional translators and used plain translation without any TM or MT. Only a browser was available to look up terms etc., which limits the functionality of touch and speech.

Our CAT interface, presented in Part II, supports handwriting with a stylus, including options to delete and add characters or words, as well as gestures for creating sufficient space to write into. Furthermore, words can be reordered with the pen through simple drag and drop. All pen functionality is evaluated in comparison to other input modalities to draw conclusions about its usefulness.

Touch

As discussed above, Zapata's (2016) study of voice and touch enabled devices also allowed finger touch input. However, no formal evaluation was conducted, and no proper touch input inside a CAT tool was used. Instead his work simply relied on the operating system's virtual keyboard on a tablet.

The Kanjingo mobile application (O'Brien et al., 2014; Torres-Hostench et al., 2017) (see Figure 2.18) allowed touch input for PE, where all words were presented as vertically aligned tiles, that could be reordered through drag and drop. Touch editing was possible on a tile level: Clicking the '-' removed words, while clicking '+' allowed entering words with the iPhone virtual keyboard, which could also be used to edit the text inside a tile. Especially the reordering functionality with touch was liked by the study participants. A notable feature request was a better support for moving groups of words.

The desktop-based CAT tool by Teixeira and partly the same authors as Kanjingo (Teixeira et al. (2019), see Figure 2.20) offered a similar touch mode with words in a tile view that can be reordered by drag and drop. Due to the additional space in the desktop setting, these tiles were not stacked vertically but placed one after the other as in normal text. However, as discussed before, reading was perceived as more complicated in the tile view and touch insertions, which required a click on the previous tile, a cursor placement at the end of that tile's word, entering a space and then the new word, also received poor feedback. For deletion, users could drag tiles on a bin icon. The tool also offered an on-screen virtual keyboard for touch interaction, which interestingly was perceived as good by a participant.

Our explorations, presented in Part II also allow reordering words as in Teixeira et al. (2019), however, without showing them in a tile view to enhance readability.

Furthermore and identical to our pen implementation, finger touch can be used for handwriting, including drawing gestures for deletion, thereby supporting all PE operations directly with the finger without any virtual keyboard.

Gesture Input

Hand gestures are a very intuitive and natural way of interaction (Ortega and Nigay, 2009; Sharma and Verma, 2015) and are thus investigated as an input modality in a variety of contexts. Suitable hand gestures depend on the application type and context (Nielsen et al., 2003; Wachs et al., 2011; Weichert et al., 2013), thus, there is no universally appropriate gesture set. Gestures must be easy to learn and memorize and metaphorically meaningful (Wachs et al., 2011; Weichert et al., 2013). A gesture lexicon should therefore be concise and executing the gesture should be comfortable to avoid muscle tension especially over long periods (Wachs et al., 2011).

To the best of our knowledge, no study or prototype has explored gesture interaction for PE or translation in general. Instead, most research focused on topics like remote controlling interfaces, e.g., TVs or cars, whereas only few papers work on document editing with hand gestures. One example is the idea presented in a patent by Rives et al. (2014) to use gestures to perform the operations cut, copy, paste, select, undo, and delete to edit a document. Furthermore, Ortega and Nigay (2009) found that using finger pointing to replace the mouse significantly reduces the switching time compared to keyboard and mouse (almost to zero) which suggests that an exploration of gestures for PE might be relevant.

Eye Input

Within the CAT tool literature, in contrast to human translation research, eye tracking received only minor attention: Translog-II (Carl, 2012) focused on capturing translation sessions, including gaze tracking, and allowed to replay these sessions with some analytical plots. Similarly, the SEECAT (Martinez et al., 2014) tool, based upon CASMACAT (Alabau et al., 2013a), logs eye tracking data, however, only to collect data for process-oriented translation research.

To the best of our knowledge, eye tracking has to date only been used to study the translation and PE process, but never as an input modality to control the translation environment. We nevertheless hypothesize that the combination of eye input with typing or speech could be a valuable extension for CAT tools, and present an implementation thereof in chapter 5.

Interesting works outside of the translation domain that might be suitable for CAT tools include the ReType approach by Sindhwani et al. (2019). ReType tackles the problem of context switching between keyboard and mouse, which is known to not only cause productivity loss but also leads to RSI (Repetitive Strain Injury) (Dernoncourt, 2014). ReType is a "gaze-assisted positioning technique combining keyboard with gaze input based on a new patching metaphor": The user types a correct prefix (which can be just a single character), then the text which should be replaced, then a correct postfix. Through the pre- and postfix, the algorithm identifies matching strings, and when the user looks at one of the matches and presses 'enter' the change is applied. The approach can also be used for positioning the cursor (write exact match, look at it, 'enter'), or text selection (by positioning at beginning, followed by a special key, then positioning at the end). One of the main results is that in terms of motor times, ReType is significantly faster than the mouse. Furthermore, 22 out of 24 participants in the study preferred ReType as they had to use the mouse a lot less due to the eye-based placement.

Another approach that could be interesting for the task of PE is the Gazemarks approach by Kern et al. (2010), addressing the problem of attention switching. While the paper did not consider the translation task at all, it might well be applied in this context: In PE, the translator needs to frequently switch the attention between source and target of the current segment, but also look back and forth to understand the context. Gazemarks could support this by memorizing the last fixation in each attention area (source, target, context), and highlight this position to make attention switches faster and less cognitively demanding by offloading the effort of memorizing where they left of. In their original study, they were able to show faster completion times in a simple visual search task. We thus also integrated the Gazemarks idea into our own CAT environment, as will be presented in chapter 5.

Naturally, one can also imagine gaze input to be useful for other tasks within CAT tools, e.g., for automatic scrolling or researching terms and phrases.

2.3.5 Summary

This section provided an overview of CAT environments. First, we looked at the individual features that are common among many CAT tools, e.g., alignments, TM, MT, IMT, or QE. Then we looked at the most relevant CAT tools from academia and gave an overview of commercial tools. We saw that while academic tools have explored particular features in detail, they are usually much more narrow in support tools offered. In contrast, the commercial tools are very broad, supporting a wide range of tools, at the cost of complexity of the interface and lagging behind on newest academic findings. Afterwards, we discussed studies exploring user interface needs, that highlighted the positive and negative aspects of current CAT tools, and showed that the tools were often designed without proper involvement of translators, thus, driven by advancements in the latest support tools rather than by a proper integration into the translation process. This can also be attributed to the rapid advances in MT and NLP research and the comparatively few studies on the HCI aspects of translation technologies. The advances on the tool side often lead to clients or LSPs demanding the latest features, without pursuing detailed integration studies with human translators.

Finally, we focused on this thesis' main research question, namely the use of modalities other than mouse and keyboard for the PE process, discussed why such approaches might become more and more relevant with better MT outputs, and which modalities and modality combinations have already been explored in the literature. We start Part II with a well-structured analysis of a considerable variety of input modalities for PE within CAT environments (see chapter 3), which we use to guide subsequent development and testing of our multi-modal CAT environment (see chapters 4 and 5).

2.4 The Cognitive Dimension of Post-Editing

Naturally, PE not only changes the required interactions, but also the mental processes before and during editing. The translator needs to continuously scan the MT proposal and source text, potentially combined with TM output, for mistakes and chunks that can be reused, meanwhile thinking about the surrounding context, the target audience, perform corrections to individual parts, and make a plan to assemble an overall translation. Thus, PE can be seen as a highly demanding task. In this section, we will first introduce cognitive load theory, then talk about measures that have been used in the literature to estimate CL. Among these measures, some have already been explored within the translation domain, which we will cover in the following section. Then we will look at the user perspective: are users more concerned about some sensors than others for estimating CL? Finally, we will discuss a domain where CL is not only measured but used for adaptations, since we want to propose the concept of CAT tools that can react to their users' perceived CL, and thereby avoid cognitive overload.

This section is most relevant for our work on RQ2, which is presented in Part III.

2.4.1 Cognitive Load Theory

Cognitive Load (CL) theory (Paas and 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). The theory assumes that humans have "a limited working memory which interacts with an unlimited long-term memory" (Paas et al., 2003). CL theory comes from the educational context, where learning is defined as the creation of schemas in the working memory, which then go over into long-term memory where they can be reused without additional strain (Leppink, 2017). In cases of cognitive overload, that means when working memory capacity is overloaded, learning is impeded, or in other contexts, errors increase (Paas et al., 2003).

CL theory distinguishes three types of CL (Leppink, 2017; Paas et al., 2003):

- *intrinsic CL*, which is the difficulty of the task itself, like a simple arithmetic addition compared to solving an integral equation. The lack of automation or cognitive schemas about the content, which leads to more elements needing processing in working memory, as well as the interactivity between elements, lead to intrinsic CL.
- *extraneous CL*, which is load stemming from other factors that do not contribute to the actual task, e.g., dividing attention between information sources, or the load introduced by a bad design of learning materials. Thus, extraneous CL should be minimized in favor of freeing up working memory for intrinsic or germane CL. Note, however, that a representation minimizing extraneous CL for novices might be unsuitable for experts and therefore lead to extraneous CL for experts.
- *germane CL*, which is the load induced by the construction of learning schemas, so the load by a deliberate engagement in cognitive processes beneficial to learning, like self-explanation of content.

Usually the total amount of CL is measured as it is not clear how the three can be distinguished (Leppink, 2017). Furthermore, a variety of researchers have modified the model to a dual model that only distinguishes intrinsic and extraneous CL and consider germane CL as a part of intrinsic CL (Leppink, 2017).

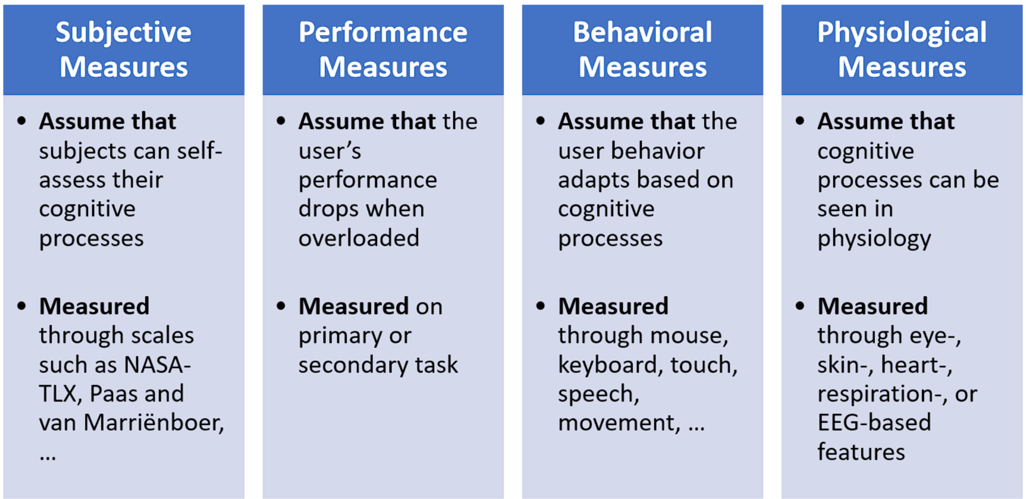
CL can also be described across the time domain: Paas et al. (2003) distinguish instantaneous load, peak load, average load, accumulated load, and overall load. Here, instantaneous load is the most immediate measure and reflects the load at every moment in a task, while peak load is the highest instantaneous load, average load is the average instantaneous load, and accumulated load is the total amount of load. Last, overall load represents the individual's perception of mental effort, that is, the experienced load.

2.4.2 General Cognitive Load Measures

While CL theory comes from psychology, the measurement of CL has especially been studied in the field of HCI. The approaches can be roughly divided into four categories (Chen et al., 2016): subjective measures, performance measures, physiological measures, and behavioral measures, as depicted in Figure 2.26.

Subjective Measures

Subjective measures are based on the assumption that subjects can self-assess and report their cognitive processes after or during performing a task (Paas and Van Merriënboer, 1994). Introspection and reporting of CL was shown to be sensitive to small differences and reliable (Paas et al., 2003). Naturally, asking users within or after tasks does not capture real-time changes in CL (Moissa et al., 2019). Thus, it is mostly used to capture average or overall load (Chen



Subjective Measures	Performance Measures	Behavioral Measures	Physiological Measures
• **Assume that** subjects can self-assess their cognitive processes	• **Assume that** the user's performance drops when overloaded	• **Assume that** the user behavior adapts based on cognitive processes	• **Assume that** cognitive processes can be seen in physiology
• **Measured** through scales such as NASA-TLX, Paas and van Marriënboer, …	• **Measured** on primary or secondary task	• **Measured** through mouse, keyboard, touch, speech, movement, …	• **Measured** through eye-, skin-, heart-, respiration-, or EEG-based features

Figure 2.26: Overview of cognitive load measures.

et al., 2016). As scales, Likert scales of different ranges (5, 6, 7, 9, 10) have been proposed (Moissa et al., 2019). Such introspection is often used as a ground truth to evaluate how well CL can be assessed by other means, such as physiological measurements. Furthermore, Van Gog et al. (2012) have shown that timing and frequency of effort ratings affect rating results and that asking users after each task (of limited duration) is preferable.

Performance Measures

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). In the learning context, such measures could be grades, percentage of correctly answered questions, time spent on a task, etc. (Paas et al., 2003). Within translation, it could be the translation/PE time or the translation quality. Another measuring approach frequently deployed in various domains is a dual-task (Chen et al., 2016), where the subject is asked to perform a secondary, dissimilar task additionally to the primary one. The performance on this secondary task is then assumed to change as different loads are induced by the primary task. However, this measuring approach can only be used in controlled experiments and is not feasible for real-world use cases.

In a performance-oriented society, one might all too easily trust in performance measures, even though they too have some drawback. For example, humans can maintain their performance under rising CL over a period of time (Hockey, 1997). This however, comes at the cost of additional strain and fatigue. A problem with time measurements, as pointed out by Moissa et al. (2019), is that it does not allow to distinguish internal and external conditions, e.g., whether the user is tired or not or in a noisy environment. Furthermore, quality and time can often be traded against each other, especially in post-editing. Finally, quality metrics are not easily applicable in every domain: e.g., in translation there is no fully reliable quality estimation without a reference translation or human ratings. In PE, researchers have often used the MT quality as a proxy for PE effort, which naturally is often related, but cannot be considered equal to CL: Consider e.g., very bad MT proposals that are still very easy to PE due to the simplicity of the segments, or the opposite, a very high MT quality where spotting the error can remain difficult and induce a high CL.

Physiological Measures

A lot of research has been conducted on physiological measurements, which assume that human cognitive processes can be seen in the human physiology (Kramer, 1991). Demberg and Sayeed (2016) explain the origins of physiological measures: When the load rises, the brain stem and in particular the locus caeruleus are signaled that more processing resources are needed. This area then releases the neuro-transmitter norepinephrine, which enhances information processing, while also affecting the pupil muscle, the heart rate, and the skin conductance as a side effect. An advantage of physiological measures is that these can be measured at high frequency and thus can capture variation of CL over time (Chen et al., 2016). This section provides an overview of the measures:

Eye tracking is frequently used for physiological CL measurements: Due to an effect called task-evoked pupillary response, the pupil diameter increases with higher CL (Beatty, 1982; Iqbal et al., 2004; O'Brien, 2006a). However, the pupil size is impacted through pupillary reflexes to changing light conditions by a magnitude more than by the task difficulty (Pfleging et al., 2016). An approach to detect CL based on pupil diameter independent of light conditions is called Index of Cognitive Activity (ICA) that measures the frequency of rapid dilations instead of the overall diamter (Marshall, 2002; Demberg and Sayeed, 2016). As eye tracking has been frequently used in the translation domain for CL studies, we will provide more detail in section 2.4.3.

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). Here, the increased level of sweat lets skin conductance rise as task difficulty is increased. Yamakoshi et al. (2008) further explore remote measurements of the skin temperature using thermal cameras by comparing the truncal and peripheral skin temperatures, which are impacted to a different degree by a change in CL (e.g., the nose temperature drops).

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), which is a measure of "the oscillation of the interval between consecutive heartbeats" (Rowe et al., 1998), have been shown to correlate with CL. Here, as CL increases, the heart rate goes up, while HRV goes down. Combinations of such measures with respiration measurements have also been explored: Chanel et al. (2008) combined several measurements to optimize the task difficulty in the gaming context.

Research on so-called **brain-computer interfaces** shows how brain activity can be used to adapt software systems to the user's cognitive state, e.g., by detecting emotional stress using Electroencephalography (EEG) (Hosseini and Khalilzadeh, 2010) or with less invasive functional Magnetic Resonance Imaging (fMRI) (Solovey et al., 2012). EEG is for example measured in the power spectrum, where alpha frequencies are reduced and theta frequencies increased as CL rises (Klimesch, 1999).

Behavioral Measures

Last, behavioral measures can be extracted from user activity while performing a task (Chen et al., 2016). These include eye features, mouse, pen, or keyboard input, speech and linguistic features, changes in gait or facial expressions, etc.

In the context of PE (without additional input modalities), mouse and keyboard input-based features, like mouse movements and trajectories in the temporal and spatial domain are especially interesting, as these were shown to correlate with CL (Arshad et al., 2013).

Other input modalities allow further behavioral features to be engineered, e.g., pen input allows the measure of angle-based, space-based, centroidal, temporal, pressure-based, trajectory-based, and other features (Prange et al., 2018).

For speech interfaces, the length of silent and filled pauses, as well as the response latency increase with rising CL (Chen et al., 2016; Khawaja et al., 2007). The authors also found that among the linguistic features explored, positive emotions, agreement words, first-person and third-person singular pronouns decreased as CL increased.

Apart from the physiological features captured with eye trackers, these devices can also be used to detect behavioral changes: the blink frequency and duration decrease with increasing load (Van Orden et al., 2001). Furthermore, Chen and Epps (2013) as well as Stuyven et al. (2000) showed that fixations and saccades (the rapid movement between fixations) can also be used for CL predictions, where the fixation frequency and duration, as well as the saccade distance increase as the tasks get more complex.

Furthermore, recent improvements in computer vision using deep learning can be used to automatically extract emotions from videos (Kahou et al., 2016), which could be an indication of cognitive processing as well. However, simple features like the head pose have also been shown to correlate to CL when learning (Asteriadis et al., 2009).

Multi-Modal Measures

Multi-modal approaches, investigating a variety of sensors simultaneously, have also been presented: As an example, Guhe et al. (2005) combine heart rate, GSR, skin temperature, blinking, gaze and pupil-related measures, as well as head movement, mouth openness, and click pressure to estimate the workload during a so-called N-back task. In N-back tasks, letters or numbers appear one after the other on a screen, thereby forming a sequence. Participants are asked to press a button whenever a letter is identical to the letter N steps earlier, thus, the difficulty can be easily increased by switching to a higher N task (e.g., from a 2-back to a 3-back task). In general, several studies showed that multi-modal approaches increase robustness of CL estimation as the individual modalities can compensate and complement each other (Chen and Epps, 2013).

2.4.3 Cognitive Load Measurement in Translation

While the previous section provided a high-level overview of existing CL measures, this section goes into more depth for those measures that were already explored in the translation domain. Overall, translation can be seen as a horizontal process, where translation simultaneously happens during comprehension (as opposed to a vertical process where comprehension happens first followed by reformulating, Macizo and Bajo (2006)). The increased use of working memory resources (likely due to the simultaneous activation of both languages), makes translation a particularly interesting domain for CL studies. Krings (2001) divided PE effort into three types: temporal, cognitive, and technical effort. In PE, cognitive effort involves the cognitive processes required, such as spotting errors and planning corrections. Technical effort means the effort for typing, copy and paste, etc. to transform the MT output into the final translation.

Time

Temporal effort on the other hand is simply the overall time required for PE, which Krings (2001) sees as a very important indicator of PE effort as it contains both cognitive and temporal effort. A frequently used performance measure therefore is the translation throughput or PE duration.

Subjective Measures

Krings (2001) further 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. Vieira (2014, 2016) instead used the 9-point scale proposed by Paas and Van Merriënboer (1994) for the translator to post-hoc judge the mental effort of PE a segment, which is a simple 9-point scale, ranging from “very, very low mental effort” to “very, very high mental effort”.

Pauses in Typing

O’Brien (2005) explored correlating extended pauses in typing behavior to potentially difficult source text features. In a follow-up analysis (O’Brien, 2006b) she analyzed this hypothesis on source segments comprising characteristics that are known to induce different effort, but concluded that “while pauses provide some indication of cognitive processing, supplementary methods are required to give a fuller picture”. A problem was the use of the pause ratio (PR) (the total pause time in a segment divided by the total time in a segment), where lots of short pauses might not have a big impact on the total pause time but often occurred during cognitively challenging tasks.

Lacruz and Shreve (2012; 2014) therefore analyzed clusters of shorter pauses instead of examining long pauses. Their metrics called Average Pause Ratio (APR) (“the average time per pause in the segment divided by the average time per word in the segment”) and Pause to Word Ratio (PWR) (the number of pauses in a segment divided by the number of words in a segment) could be correlated to technical effort (the required mouse and keyboard actions) measured as HTER, arguing that “it is likely that in many situations technical effort and cognitive effort will be related”. A direct correlation analysis in comparison to CL in the psychological sense was not performed.

Mellinger (2014) focused on cognitive effort when using TM by correlating keystroke logs and pause metrics to a subjective measure of translation quality on a 5-point Likert scale. He found that the cognitive effort in terms of APR was highest when translating from scratch, followed by fuzzy matches and finally exact matches.

Eye Tracking

An early comparative study between PE and manual translation using Translog by Carl et al. (2011) showed that in translation, fixation counts are generally higher than in reading with more regressions (i.e., backward movements of the eyes happening during reading to revisit misidentified words). More fixations were found with more complex texts. Gaze times were used as simple indicators of cognitive effort, where more fixations and longer gaze times were found on the target text in PE. Therefore the authors argue that there is more cognitive effort on the target side. In contrast, more effort for reading and understanding the source is required in manual translation. However, the authors do not argue whether overall effort is higher in translation from scratch or PE.

O’Brien (2006a) proposed pupil dilation as a measure of CL, as this feature has been shown to correlate with CL in other domains. She focused on correlations between percentage change in pupil dilation and different translation aids. Translation from scratch required the highest CL (as per pupil diameter), exact matches the least effort, and MT PE similar to fuzzy matches between 80 and 90%. A general drawback of the pupil diameter is its delayed reaction to the increase of CL and dependence on very static light conditions.

Doherty et al. (2010) also explored eye tracking as a means for MT evaluation by measuring different features while reading MT output and correlated it with human evaluation, thus, measuring MT quality for reading. They found that gaze time and fixation count correlate with human evaluation of MT quality; however, fixation duration and pupil dilation were less reliable in their study. While these correlations are quite interesting, they do not capture CL but MT quality. Furthermore, the focus is on a reader instead of a post-editor, thus, not targeted towards post-editing but direct usage of MT output.

The ICA, a pupil dilation-based measure, has also been correlated to linguistic processing difficulty in a variety of experiments (e.g., investigating grammatical gender mismatch or semantic anomalies, Demberg and Sayeed (2016)). As discussed above, the ICA separates changes in pupil size resulting from light reflex (which are larger and slower) from the rapid and small dilations that are driven by CL, and well reflected linguistic processing difficulty in the experiments.

Multi-Modal Cognitive Load Measurements

Koglin (2015) uses eye tracking (total fixation duration) and keylogging data to investigate the difference between translating metaphors manually or through PE. The results of a between-subjects study suggest that cognitive effort is lower when PE: PE was significantly faster, having significantly less insertions, but no differences in deletions. She also replicated the findings by Carl et al. (2011) that fixations are longer on the target for PE and on the source for manual translation. Furthermore, total pause duration was lower, albeit this can simply be attributed to the twice as long translation times in manual editing. Thus, due to the increased speed in PE, less pauses and insertions were made since a draft text already exists. The authors argue that the fewer pauses indicate a lower CL, however, I contend that this only applies to accumulated load (due to the shorter time) but not necessarily to average, peak, or experienced load.

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.

Kruger et al. (2016; 2018) explore a combination of eye tracking (fixation counts and durations), EEG, and self-reported psychometrics, to investigate the impact of different movie subtitles on CL. Fixation count provided an indication of extraneous load, while fixation duration hinted at the depth of processing.

In contrast to these quality-, time-, and expectation-based measures, Vieira (2014) uses a psychology-motivated definition of CL. Using an eye tracker, 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 (based on linguistic features) and their corresponding METEOR ratings.

In a follow-up work, Vieira (2016) argued that instead of comparing the number of editing operations, the cognitive effort would be a "more decisive indicator of the overall effort expended by post-editors". Therefore, he analyzed 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. He stated the possibility 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". Overall, CL can be considered "an inherently subjective variable that depends on how individuals cope with variation in the demands of a task".

The disseration by Daems (2016) presents a comparative analysis between PE and manual human translation for English to Dutch texts with both professional and student translators. A combination of speed analysis, quality analysis, keystroke logging, eye tracking (fixation duration), and surveys for perceived quality, speed, and usefulness were used. Overall there were less fixations in PE than human translation which is why Daems argues PE would be less demanding. Again, given that the tasks differ strongly, I believe that a simple fixation count is too simplistic to draw conclusions about overall effort. She also found more attention on the target and less attention on the source for PE compared to manual translation. Furthermore, fewer external resources were consulted in PE, which could come from the already given proposal by the MT.

What leads to effort?

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 longer PE times.

Temnikova (2010) compared MT that was simplified using controlled language rules with normal MT in terms of errors. She 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. The MT results were then considered to induce a specific cognitive effort depending on the error types contained; however, an analysis of which CL these errors induce on editors based on CL measures was not performed.

Daems (2016) further found that “the more technical effort indicators (number of production units, pause ratio, and APR, HTER) are mostly impacted by grammatical errors (grammar, structure, word order), whereas the more cognitive effort indicators (fixations and time) are influenced most by coherence and other meaning shifts”.

Koponen (2012) studied the relationship between cognitive and technical post-editing effort by comparing edit distances to human judgments of segments on a 5-point scale 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 (2012) further found that her study participants found sentences with reordering more cognitively demanding, and changing a word form least cognitively demanding. Furthermore, the analysis showed that the factor most affecting the perception is the sentence length, with long segments being perceived as requiring much more effort, even when the number of edits is small.

Most of these works however, again measure quality/expected percentage that needs editing and implicitly assume 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”. They used the classification of error types in terms of CL proposed by Temnikova (2010), ranking error categories by the authors’ expectation of CL. A link was found between the types of errors within segments and the time required for PE. In a recent survey, Koponen (2016) concludes that determining the amount of PE effort is still not fully solved and that “accurate measurement of the actual effort would be important as it has implications not only for productivity, but also for the working conditions of the post-editors”. Lacruz and Shreve (2014) correlate different error types, classified into mechanical and transfer errors, to PWR, HTER, and to user ratings on a 5-point scale 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.

Vandeghinste et al. (2019) also have an MT error taxonomy, where monolingual errors are fluency errors, and errors requiring both languages are accuracy errors. Both classes again have several subclasses. Very interesting and related is that they use these MT errors to predict temporal PE effort, i.e., PE time. For this, a study with two Master’s students found that PE-time “can be estimated with high accuracy, provided that the types of errors in the MT output are known” (Vandeghinste et al., 2019). Furthermore, the authors found that accuracy-mistranslations and fluency-grammar errors are the two main error categories impacting PE time. Not all but only relevant error categories were required to predict PE time, which can also be seen as a sentence-level QE system. This automatic error detection for predicting PE time is achieved using RNNs.

To summarize, these works provide insight into which features of a MT output lead to longer PE times or worse subjective quality ratings, but a direct link to CL in the psychological sense was not shown, but only assumed to exist.

Challenges of the translation and PE domains

The translation and in particular the PE domain pose 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 as in the often-used artificial tasks (e.g., N-back or arithmetic tasks). Also, the frequently used performance measure in a dual-task design is impractical for such a real-world task since the focus should remain on the PE task without distraction. This makes performance-based measures less feasible for this domain. Other performance measures, 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.

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, while speech and linguistic based measures (Khawaja et al., 2014) are only possible in multi-modal CAT environments.

Last, any sensors should act in the background and 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.

The inter-translator differences, which make performance measures difficult, could, however, be captured well by subjective measures. Such subjective measures, however, interrupt the PE process and cost time. An interesting research direction with only few publications so far is therefore to use physiological and behavioral measures proposed in HCI to estimate the subjective CL while PE. For example, one could integrate gaze data, heart features, or skin resistance to detect text parts that are hard to correct and automatically propose alternatives to the post-editor. This is what we investigate in Part III, where we first analyze which adaptations might be suitable for PE, and then present a multi-modal CL estimation framework which we explore in several studies.

2.4.4 User Acceptance of Cognitive Load Measurement

We have seen that there is a large variety of approaches to estimate CL, and some of them have been explored in the translation domain. Apart from investigating how suitable sensors are for CL measurement, it is important to understand whether users would accept being tracked by these measures, or if they have privacy concerns. This section reviews literature on the user acceptance of different CL measurement approaches and builds the basis for chapter 9.

Fensli et al. (2008) analyzed how well patients accept wearable sensors in the medical context. While medical information and the healthcare context are perceived as particularly sensitive, many health apps capturing similar data show poor information privacy practices (Schomakers et al., 2018). Perez and Zeadally (2018) split privacy issues and solutions for consumer wearables into three areas: context privacy, bystander privacy, and external data-sharing privacy. Relevant for our use case are context privacy and external data-sharing privacy, which include users' fears, data disclosure, etc. Privacy concerns when wearing a sensor suit were seen as most critical in the context of conversation and commuting; collecting stress information, temporal and spatial data, as well as sharing the data with the general public, increases these concerns further (Raij et al., 2011).

Lehto and Lehto (2017) find that participants do not perceive the numerical information collected by wearables as sensitive; however, health records including written information are considered very private. Motti and Caine (2015) show that users have different concerns based on the type of data collected, the sensor used, and the purpose of the wearable: microphones and cameras pose the most privacy concerns followed by GPS, while heart rate monitors or activity trackers are seen as less problematic. According to Vitak et al. (2018) and Zimmer et al. (2020), most fitness tracker users only express minimal privacy concerns and show only an average level of concern if their data were compromised. A survey on privacy concerns in wearable devices is provided in Datta et al. (2018) and anonymization techniques to solve some of the privacy issues are discussed. Users also tend to underestimate or ignore potential risks, e.g., the lack of a keyboard makes users assume the collected data cannot be sensitive (Lowens et al., 2017). In a study with college students (Udoh and Alkharashi, 2016), users assumed that the producer of the wearable would take appropriate measures against privacy issues and therefore felt safe.

The willingness to share personal data is also linked with the trust in the security (Acquisti et al., 2013) and the storage location (Lidynia et al., 2017) of fitness tracker data. A taxonomy developed on privacy risks for consumer health wearables (Becker et al., 2017) reveals that these risks refer to the perceived data sensitivity, data variety, and tracking activity. In general, however, users often share private information even when they claim to be concerned about privacy (Williams, 2018), which can be seen as an attitude-behavior gap.

A methodologically interesting work by Acquisti et al. (2013) asked participants for the amount of money necessary to share otherwise private data, and the amount of money participants would pay to make otherwise public data private. While results are not relevant for CL estimation since they focus on the shopping scenario, we will use this approach in chapter 9 to study which CL measurement approaches is most useful in practical applications.

2.4.5 Adaptations to Measured Cognitive Load

Since CL theory stems from educational psychology, many proposals of CL-aware systems have also been made in the **learning context**: Kuo et al. (2007) propose the idea of a context-aware learning system that considers factors like facial expressions, human voice, or body temperature. Recommendations of learning content based on ontologies about the learner and the content, as well as behavioral, positional, temporal, and physiological data, have also been proposed (Pernas et al., 2014; Yu et al., 2007). Furthermore, dynamic user interface adaptations (Ghiani et al., 2015) and adaptive visualizations (Chen, 2016), driven by physiological parameters, were suggested to support learning. The concept of affective e-learning, which uses emotion feedback to improve the learning experience, was proposed in Shen et al. (2008). The work showed in a feasibility study that biosensors can be utilized for this purpose. A review of affective computing in education can be found in Wu et al. (2016a), which highlights the essential role that positive emotion has on comprehension performance. Bahreini et al. (2016) investigate emotion recognition using webcams and microphones to better respond to the affective states of students, as human teachers would in traditional learning. Similarly, Ishimaru et al. (2017) link eye tracking data, including fixations and pupil diameter as well as thermography, to surveys about cognitive states of high school students when studying a digital physics book. Based on this, they propose to provide individualized information to enhance learning abilities. Leony et al. (2012) showed that such adaptations can affect cognitive processes like memorization and decision making. Sensor data was also linked to a subjective measure of flow (Léger et al., 2010), stress detection (Rodrigues et al., 2013), and motivation (Bauer et al., 2018) for the case of e-learning.

A framework for learning analytics based on wearable devices "to capture learner's physical actions and accordingly infer learning context" was proposed by Lu et al. (2017). As a first use-case they focus on traditional learning and implement student engagement detection for the classroom based on arm movement to intervene when engagement is low, or to provide incentives when it is high. Moissa et al. (2019) review the literature on measuring students' effort and propose to use wearables to estimate cognitive load and predict the CL level for future learning tasks. Students could also be alerted when they should take a break or move to less challenging tasks as detected through wearables.

Apart from these works focusing mainly on conceptual design or correlations, several works go a step further and investigate the feasibility of adaptations to CL by training **predictive models**: For arithmetic calculations, Borys et al. (2017) trained trinary classification models (low CL, high CL, without task) based on brain and eye activity, reaching up to 73% accuracy with a KNN classifier. Similarly, a Ridge regression model based on EEG data during arithmetic operations was able to determine the task complexity with comparatively low error (Spüler et al., 2016). The resulting model was then used in a second study to adaptively propose tasks to learn arithmetic additions in the octal number system

(Walter et al., 2017), where the resulting learning success was comparable to a learning system adapting to the amounts of errors made by the learner. Similarly, Galán and Beal (2012) use SVMs to predict the success or the failure of students solving math problems based on a combination of attention and workload signals from EEG sensors, achieving between 57 and 87% accuracy depending on the exercise difficulty. Instead of EEG data, Mock et al. (2016) use touchscreen interactions to classify CL for children solving math problems, with an average binary classification accuracy of 90.67% using SVMs.

Similar to the above ideas on CL adaptations in the learning context, chapter 6 presents our discussions with professional translators how CAT tools could react to estimated CL to improve the PE process. Chapter 7, section 8.1, and section 8.2 then explore how well CL can be measured by a multi-modal frame during PE, and section 8.3 uses the same setting in an e-learning task, to explore the generalization of our multi-modal CL estimation approach.

2.4.6 Summary

The remuneration for PE are often established through productivity tests considering the performance improvements compared to traditional translation from scratch, however, this might not capture the full effort involved and could lead to frustration (Cumbreno and Aranberri, 2019). This section therefore discussed how the user's cognitive state can be modeled and presented a variety of works, both from other domains and from the translation domain, that used a variety of sensing approaches to estimate CL. We have then discussed privacy concerns that users might have towards such estimations, and finally talked about approaches to use the CL measurements to create cognition-aware systems that adapt to the user to avoid cognitive overload as well as boredom.

Compared to the literature on learning, no works on cognition-aware translation environments have been published, which is why we discuss this idea with professional translators in chapter 6. Furthermore, many of the CL estimation techniques presented above have not yet been explored in the context of translation, thus, we present such a multi-modal CL estimation framework in chapter 7 and explore it in three studies (chapter 8). Finally, we also look at the user acceptance of different sensors for CL estimation in chapter 9.

Combining the previously proposed CAT tool supporting multi-modal input, with such multi-modal sensing approaches, would overall lead to a multi-modal multi-sensor interface. "Multimodal interfaces also support improved cognition and performance because they enable users to self-manage and minimize their own cognitive load." Here, "sensors and input modalities can be coupled within an interface - including that either can be used intentionally in the 'foreground' or they can serve in the 'background' for transparent adaptation that minimizes interface complexity and users' cognitive load" (Oviatt et al., 2017).

2.5 Automatic Post-Editing

Even if a human translator can efficiently translate using all the above tools and improved interfaces, she still does not want to correct repetitive mistakes of a MT again and again, but pursue an intellcutally rewarding activity (O'Brien and Moorkens, 2014).

The most intuitive adaptation to post-edits would be to include the correction with corresponding source segment as a new training sample and retrain the MT system on the fly. While in theory, the underlying machine learning algorithms should learn from the corrections, this approach has two fundamental disadvantages: (1) training times for modern NMT approaches are way too long to make frequent retraining viable in real-world translation workflows; (2) given the millions of sentence pairs that MT systems are trained on, it is very unlikely that a few additional samples change the model weights enough to guarantee that it is able to avoid the mistakes in future. Thus, research has focused on approaches to learn from post-edits, e.g., through online model adaptation in SMT by adding new rules to the translation model, adapting the lanugae model, and parameters (Denkowski et al., 2014), or through active learning based on source-side information (Dara et al., 2014).

Another approach to tackle this issue is Automatic Post-Editing (APE), which incrementally adapts MT to post-edits and thereby aims to automatically correct errors made by MT systems before performing actual human post-editing (Knight and Chander, 1994). As depicted in Figure 2.27, APE can be viewed as a 2nd-stage MT system, translating predictable error patterns in MT output to their corresponding corrections. Knight and Chander (1994) see this as an alternative to having improvements right inside the MT system, where they would become a part of the black box. In theory, APE systems can also be independent of particular MT systems, thus, making them "portable across MT systems" accomplishing "their tasks without reference to the internal algorithms" (Knight and Chander, 1994). Upon availability of human-corrected post-edited data, APE can thus adapt any black-box (1st-stage) MT engine without incremental training or full re-training to improve the overall translation quality (Pal, 2018).

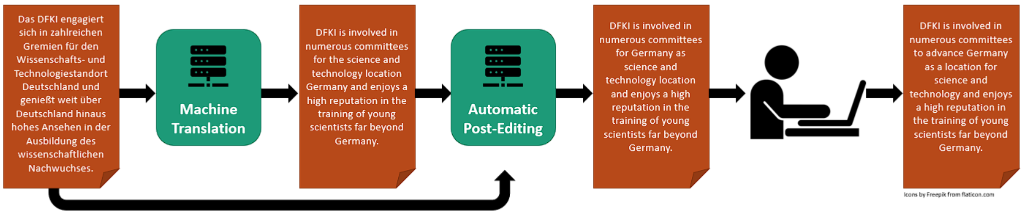

Das DFKI engagiert sich in zahlreichen Gremien für den Wissenschafts- und Technologiestandort Deutschland und genießt weit über Deutschland hinaus hohes Ansehen in der Ausbildung des wissenschaftlichen Nachwuchses.
Machine Translation
DFKI is involved in numerous committees for the science and technology location Germany and enjoys a high reputation in the training of young scientists far beyond Germany.
Automatic Post-Editing
DFKI is involved in numerous committees for Germany as science and technology location and enjoys a high reputation in the training of young scientists far beyond Germany.
DFKI is involved in numerous committees to advance Germany as a location for science and technology and enjoys a high reputation in the training of young scientists far beyond Germany.
Icons by Freepik from flaticon.com

Figure 2.27: Automatic post-editing.

Already back in 1994, Knight and Chander (1994) distinguished two types of APE called *adaptive* and *general*: *Adaptive* approaches learn from human post-editors (or their captured post-edits as a corpus) and "begin to emulate what the human is doing", however, noting that the types of errors are different for any MT system or domain. In contrast, *general* APE approaches deal with certain linguistic criteria that should be corrected on any MT output or domain, e.g., selecting the correct article. Reflecting on the current state-of-the-art in APE (see below), we can see that modern approaches focus on adaptive APE, probably because MT research itself is currently also highly data and less rule-driven.

The aim of such adaptive APE is to reduce the translators' workload and increase their productivity (Pal et al., 2016a; Parra Escartín and Arcedillo, 2015a,b). Depending on the training data, it can not only learn corrections, but also individual stylistic preferences of the translator or domain, e.g., decisions to paraphrase certain sentences. The large first-stage MT system is thus kept as is, while the second-stage APE model is optimized to learn from post-edits. An interesting key difference between APE and MT as pointed out by Bojar et al. (2015) is that MT must translate each word, whereas APE can decide to keep words untouched.

2.5.1 WMT Shared Task on Automatic Post-Editing

Since 2015, the Workshop/Conference on Machine Translation (WMT) hosts a shared task on APE (Bojar et al., 2015, 2016, 2017; Chatterjee et al., 2018, 2019, 2020; Akhbardeh et al., 2021). In a shared task, different teams compete on the same provided datasets, which allows for structured analyses of the different approaches and thereby pushes the state-of-the-art in APE. As the works presented in Part IV were also tested on WMT data, it is important to provide an understanding of this shared task on APE.

In this task, participants are asked to automatically correct "errors produced by an unknown machine translation system" (Bojar et al., 2015). For this, the task organizers prepared triples of source data (src), machine translation output (mt), and human post-edits thereof (pe). The latter (pe) are naturally only accessible for training, whereas the final goal is to produce outputs similar to pe automatically for provided src, mt tuples in the test set (or at inference time). The shared task is thus framed as a black-box scenario, where the internal workings of the MT system are unknown and the triples are all that is known. Thus, APE approaches cannot interfere in the decoding process. Instead, knowledge from the triples must be gained to automate repetitive parts of the PE process. The task organizers further chose to use independent segments which are thus not coherent. While this prevents APE systems from considering context, it allows randomly splitting segments into train, development and test sets, which offers a higher error repetitiveness between those sets.

Evaluation

The evaluation of an APE system's performance can be achieved by comparing the hypothesis pe with the reference pe output, either through automatic measures or manual human evaluation. In WMT 2015, TER was the only metric used, both case-sensitive and case-insensitive. In 2016, both TER and BLEU were used in case-sensitive mode, which was especially important for the chosen target language (German). Furthermore, this second round added human evaluation based on ranking up to 5 anonymous APE system outputs. The human evaluation matched well the automatic scores, showing their reliability for the APE task. In 2017, apart from again using TER and BLEU, the manual evaluation was changed towards using direct assessment (Graham et al., 2017) (both with crowd-workers on Amazon Mechanical Turk and translation students). To compare APE outputs to unmodified mt and human-corrected pe, these were also included in the evaluation and rated by the evaluators. To reduce annotation effort, a preprocessing step checked if multiple systems produced the same output, which happens quite frequently given that APE models can choose to do nothing or only few modifications. Interestingly, the quality control of the ratings showed that students are more reliable annotators, which all passed the test in contrast to only 54% of the crowd-workers. In 2018, the same 3 metrics were used (TER, BLEU, direct assessment), however, the manual evaluation was this time conducted by professional translators and proficient translation students. Furthermore, WMT 2018 additionally explored TER and BLEU not only by comparing the hypothesis to pe, but also by comparing it to external references, or against both pe and external references. This helps exploring whether APE systems perform edits different than pe that might still be valid, i.e., whether mt is transformed more towards pe or some other reference. In 2019 and 2020, the analysis was done mostly in line with the 2018 approach. In 2021, the approach was again comparable, however, no comparison to external references was conducted.

Baselines

These evaluation approaches are suitable to compare different APE models; however, baselines further help understanding whether the proposed models really extend the state-of-the art. The standard baseline for automatic evaluations is to simply compare the provided mt with the reference pe, thus, calculating the quality of the first-stage MT system. This can also be seen as an APE system that leaves the mt unmodified and merely copies it as a hypothesis pe. Calculating TER/BLEU between this raw MT baseline and pe is thus identical to HTER/HBLEU, as it measures the similarity between mt output and its corresponding post-edits (pe). In WMT 2015, 2016, and 2017 also another baseline was used, namely a re-implementation of the statistical APE method proposed by Simard et al. (2007a). In WMT 2018, this statistical APE baseline was dropped, as it was not competitive with the new neural APE approaches.

Language Pairs

As language pairs, the tasks dealt with English-Spanish in 2015, followed by English-German in 2016, 2017, 2018, 2019, 2020, and 2021. Additionally, WMT 2016 explored German-English, WMT 2019 added English-Russian, and WMT 2020 and 2021 used English-Chinese as additional sub-tasks.

Data Considerations

Every year, the task organizers explore why their dataset was well or badly chosen to be successfull in APE, which in turn shows where APE already performs well and where its limits are. **WMT 2015** used data from the news domain, however, based on the results, the task organizers found that the variability in such general domains makes learning tough because there are fewer patterns that can be learned from the training data and transferred on the test data. Back then, they also believed that data repetitiveness is an important factor, which intuitively makes sense but turned out to be less important in later rounds of the shared task. Crowd-based post-edits were less consistent in terms of editing patterns and amount of editing compared to post-edits from professional translators. Therefore, all following shared tasks used post-edits from professional translators as data sources.

Apart from using professional post-edits, **WMT 2016** switched to data from the information technology (IT) domain, which is more restricted and thus features more repetitiveness and a smaller vocabulary. In **2017**, for English-German again data from the IT domain was used while for German-English the Pharmaceutical domain was chosen. The German-English task turned out much more challenging due to higher quality MT, having 45% compared to 14% of perfect (TER=0) translations in the test set. This calls for much more conservative approaches that carry less risk of deteriorating the mt. Indeed, the proposed APE approaches worked poorly on this dataset, even though the German-English dataset was more than twice as large (25k vs 11k). However, the English-German data was more repetitive than the German-English data.

Since MT research switched from PBSMT approaches to NMT approaches, **WMT 2018** offered two sub-tasks, one for PBSMT (reusing data from 2016 and 2017) and one for NMT (using new data). Due to higher MT quality, the NMT task can be considered much more challenging, with 25.2% vs. 15% of test sentences that can be considered perfect (TER=0).

In **2019** only NMT data was used, as PBSMT was already too outdated to draw useful conclusions. Apart from the English-German NMT set, an additional English-Russian NMT set was released, that proved to be of higher quality, making this sub-task more challenging. An interesting finding of the data analysis was that the repetition rate, originally thought to be very important for APE, is only of marginal relevance compared to the MT quality, as the English-Russian data had high repetitions and highest quality, with no submission beating the simple *mt* baseline. A reason could be that much less can be learned when the MT output is already near-perfect. While for English-German 25.2% of segments were already perfect (TER=0), for English-Russian 61.4% made the task considerably more difficult as many automatic edits would worsen the results.

Instead of running APE on domain-adapted MT output as before, **WMT 2020** explored how well APE works to fix errors in a non domain-adapted neural MT using Wikipedia text. This is particularly interesting, as one of the main motivations for APE is its ability to adapt a first-stage MT system towards gathered data. The comparably lower MT quality gives APE more room for improvements and shows more errors to learn from in the data, even though the repetition rate was the lowest investigated so far. Furthermore, the lower quality MT lead to much less sentences in the data that are near-perfect, therefore APE has less risk of damaging good MT output. The task organizers hypothesized that ideal conditions for APE would have a peak in error distribution in the post-editable section, thus, not perfect but neither too bad for post-editing, for instance bigger than 0 but smaller than 40 in TER, a threshold that humans consider post-editable and would not translate from scratch (Turchi et al., 2014). **WMT 2021** also focused on a non domain-adapted MT system, yielding a similar repetition rate as in 2020, high-quality outputs (18.05 TER as a baseline), and a TER distribution that is strongly skewed, with over 50% of sentences having a TER value below 10. The authors therefore consider it the hardest APE task so far.

Artificial Datasets and Data Size

Since the first round of the shared task, the usage of any additional training data was allowed, and Junczys-Dowmunt and Grundkiewicz (2016) indeed proposed an artificial dataset containing 4.5 million sentence pairs generated through backtranslation. This dataset was then officially advertised as an additional training resource since 2017.

Furthermore, the switch to neural APE requires more data, which lead to the introduction of another artificial dataset called eSCAPE (Negri et al., 2018b). It contains 14.5 million samples, 7.25 of which coming from NMT and PBSMT respectively, created by using the reference of MT training data as fake-*pe* and exploiting another NMT system to generate the *mt*.

By contrast, the amount of real data is rather small, especially compared to the data amount required for training full MT models: the sentence pairs for training ranged from 11k in 2015, 12k in 2016, 11k/25k in 2017 (for different domains), 13k in 2018, 15k in 2019, and 7k for 2020 and 2021 (Bojar et al., 2015, 2016, 2017; Chatterjee et al., 2018, 2019, 2020; Akhbardeh et al., 2021). Thus, given only a small amount of (real) data for the task, APE needs to be able to outperform the 1$^{\text{st}}$-stage MT system.

2.5.2 Approaches to Automatic Post-Editing

SMT vs. NMT

The field of APE covers a wide methodological range, but similar to MT, the approaches changed from SMT-based approaches (Chatterjee et al., 2017b; Lagarda et al., 2009; Pal et al., 2016d; Rosa et al., 2012; Simard et al., 2007a,b) towards neural APE, which was first proposed by Pal et al. (2016b) and Junczys-Dowmunt and Grundkiewicz (2016). We will therefore focus on neual APE in the remainder.

Single- vs. Multi-Source Approaches

Originally, neural APE was proposed for the single-source scenario which does not consider src, i.e., $mt \rightarrow pe$. Such a single-source APE system can be viewed as a mono-lingual editor correcting the MT output. Apart from single-source ($mt \rightarrow pe$) APE approaches, the topic can also be addressed as a multi-source ($\{src, mt\} \rightarrow pe$) task, which also considers the source and thus resembles the normal PE process. Since multi-source approaches can take advantage of the dependencies of translation errors in mt originating from src, thereby providing the context that mt was created in, these approaches have already been proposed for the statistical case (Béchara et al., 2011; Chatterjee et al., 2015) and later neural APE (Chatterjee et al., 2017a; Junczys-Dowmunt and Grundkiewicz, 2016; Libovický et al., 2016; Varis and Bojar, 2017), making it the de-facto standard in modern APE. A multi-source neural APE system can for example be configured by using a single encoder that encodes the concatenation of src and mt (Niehues et al., 2016) or by using two separate encoders for src and mt and passing the concatenation of both encoders' final states to the decoder (Libovický et al., 2016). Figure 2.28 compares single-source APE to two multi-source APE approaches, one with a single and one with two encoders. Nowadays, most works focus on the multi-source multi-encoder scenario, and the question how to best encode src and mt is actively researched. We also present two such architectures in Part IV.

Evolution of APE Approaches

This section provides a historic look at how the approaches changed by focusing on submissions to the shared task on APE at WMT:

WMT 2015 In WMT 2015 (Bojar et al., 2015), the methods were mostly statistical based on the approach by Simard et al. (2007a). However, the novelties included using pipelines of different modules (e.g., predicting if the mt or APE output is better instead of always applying APE). A rule-based approach was also explored, having e.g., rules for predicting the word case or verbal endings. However, none of the submitted runs was able to beat the baseline MT system, meaning that no approach was able to reliably learn automatic correction patterns. While all

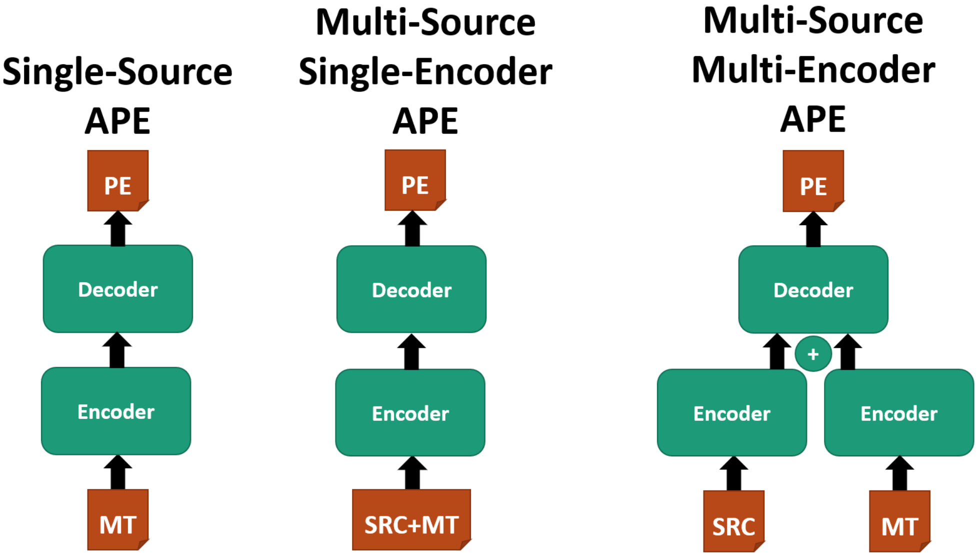

Single-Source APE
Multi-Source Single-Encoder APE
Multi-Source Multi-Encoder APE
PE
Decoder
Encoder
MT
SRC+MT
+
SRC

Figure 2.28: Single- vs. multi-source and encoder automatic post-editing.

approaches were worse than the baseline, the more conservative ones performed better than the more aggressively editing systems that risk performing wrong, redundant, or merely different edits than the single pe version.

WMT 2016 In WMT 2016 (Bojar et al., 2016), for the first time a neural approach participated: The winning system by Junczys-Dowmunt and Grundkiewicz (2016), that also proposed the artificially created large APE corpora, builds upon the work by Bahdanau et al. (2014) that leverages attention mechanisms and is trained using BPE (Sennrich et al., 2016) to avoid out-of-vocabulary words. They combine two single-source models, one from mt to pe and another from src to pe, by ensembling. Finally, string matching comparing the mt and the hypothesis pe is used to penalize words that are not part of mt to avoid over-correction. Another neural approach using attention (Bahdanau et al., 2014) was presented by Libovický et al. (2016), which leverages two separate bi-directional RNN encoders for src and mt working on one-hot vectors, where the RNN decoder uses a weighted combination of both encoder states as input. Interestingly, they work on transformed target sentences to focus on the post-edits: Instead of the actual pe string, they calculate the difference to mt and reformulate pe as a sequence of “keep”, “delete”, and inserts, where inserts use the word to be inserted. Apart from these neural approaches, automatic rule learning approaches and a variety of extensions to statistical approaches were explored. Different from 2015, in 2016 half of the approaches managed to beat the baseline. This can probably be attributed to the higher repetitiveness in the IT data compared to the more general news data, which lead to the approaches of 2016 also editing many more sentences than those of 2015. The organizers conclude that the top-ranked neural system (Junczys-Dowmunt and Grundkiewicz, 2016) was able to learn better corrections from the data, however, as they created a large artificial dataset that was unavailable to others, the results might also stem from this additional data. Especially evident was this winning approach's ability to learn applying shift operations.

WMT 2017 In WMT 2017 (Bojar et al., 2017), all submissions were based on neural networks, and mostly sticked to multi-source models leveraging both src and mt to predict pe. Furthermore, the artificial dataset from 2016 was widely adopted to feed the data hungry neural approaches. Improvements over the baseline were much higher for English-German than for German-English, where the mt was already much better and therefore harder to improve. The focus was on comparing single-source to multi-source approaches; comparing different attention mechanisms (soft attention vs. hard monotonic attention); exploring character-to-character vs. BPE-based neural networks; having a single encoder working on the combination of src and mt vs. a multi-encoder setup passing a weighted combination of both encoders to the decoder; incorporating features from word-level QE; ensembling and re-ranking based on the distance to the mt; or again focusing on predicting keep, delete, insert instead of the actual pe in the hope of achieving a better focus on the corrections and therefore making it easier to learn the identity function. In general, WMT 2017 clearly showed technological advances compared to 2016, where all systems significantly outperformed the do-nothing baseline, often by a large margin. Naturally, however, the gap to human pe was still large, showing that APE does not automate PE, but only reduces the gap between mt and pe and thereby human effort.

WMT 2018 While previous rounds have shown that especially neural APE approaches can be used to improve over PBSMT output, WMT 2018 additionally explored if neural approaches can also be used to improve NMT output. Furthermore, all submitted methods now built upon the Transformer architecture (Vaswani et al., 2017), which had previously achieved the new state-of-the-art in NMT (see subsection 2.1.4). The attention mechanisms in the Transformer allow to capture global dependencies between input and output, thus, between src, mt, and pe. The APE shared task on NMT proved to be much more challenging, which can be explained by its much higher mt quality, but also by less available training data (28k vs. 14k). While all participating teams experimented with the Transformer architecture, the details differed, ranging from learning a single model for the PBSMT and NMT task and informing the model about the data source through a separate token, to different approaches how to combine src and mt in a multi-encoder setup. As we also participated in WMT 2018 (see chapter 10), we present these approaches in more detail:

Our own submission to WMT 2018 (Pal et al., 2018) uses three self-attention-based encoders, two separate self-attention-based encoders to encode mt and src, followed by a self-attended joint encoder that attends over a combination of the two encoded sequences from mt and src and is used by the decoder for generating the post-edited sentence pe. This work will be explained in more detail in chapter 10.

Tebbifakhr et al. (2018), the NMT-subtask winner of WMT 2018 ($wmt18^{nmt}_{best}$), employed sequence-level loss functions in order to avoid exposure bias[24] during training and to be consistent with the automatic evaluation metrics. Shin and Lee (2018) propose that each encoder has its own self-attention and feed-forward layer to process each input separately. On the decoder side, they add two additional multi-head attention layers, one for $src \rightarrow mt$ and another for $src \rightarrow pe$. Thereafter, multi-head attention between the output of those attention layers helps the decoder to capture common words in mt which should remain in pe.

The WMT 2018 winning system for the PBSMT task by Junczys-Dowmunt and Grundkiewicz (2018) ($wmt18^{smt}_{best}$) also presented a Transformer-based multi-source APE architecture called dual-source Transformer. They use two encoders and stack an additional cross-attention component for $src \rightarrow pe$ above the previous cross-attention for $mt \rightarrow pe$. Comparing Shin and Lee (2018)'s approach with the winner system, there are only two differences in the architecture: (i) the cross-attention order of $src \rightarrow mt$ and $src \rightarrow pe$ in the decoder, and (ii) the winner system additionally shares parameters between two encoders. In a WMT research paper, Libovický et al. (2018) investigated four different input combination strategies for multi-encoder based Transformer architectures by varying the encoder-decoder attention: serial (computing encoder-decoder attention one by one), parallel (attends to each encoder separately and sums up context vectors), flat (concatenating both encoder states and attending to this flattened vector), and hierarchical (first attend to each encoder separately, then attend over the resulting contexts). They evaluated their methods on translation tasks with multiple source languages and showed that the models are able to use multiple sources and improve over single source baselines.

All approaches presented at WMT 2018 managed to drastically outperform the mt baseline for the PBSMT task, however, automatic improvements on top of the NMT system were only minor or even worse than the NMT baseline. Further analysis on the PBSMT data clearly showed that the APE models indeed learn the patterns between mt and pe and thus push mt in the direction of pe and not in the direction of other external references. However, the comparison to external references showed that acceptable over-corrections (that are correct but not necessary as they are not in pe) indeed do happen (at least in the PBSMT case). Furthermore, the higher NMT quality lead to APE systems being less aggressive than those for the PBSMT task. However, to achieve higher quality on the NMT task, APE models would need to become more aggressive. Further interesting insights were the types of edits done by the APE systems, where reordering was much more important for PBSMT than for NMT, which can be explained by the

[24] The exposure bias problem in sequence-to-sequence models is a result of using the ground truth sequence during training, which has a different data distribution than the previously generated tokens used for predicting the next token during inference.

higher fluency of NMT output. Instead, lexical choice becomes more important for NMT (reflected in the amount of substitutions done by APE). While direct assessment showed clusters of statistically different performing APE methods for the PBSMT sub-task, all NMT submissions were on the same level with the "do-nothing" baseline.

WMT 2019 The English-German data in 2019 (Chatterjee et al., 2019) was identical to that of 2018, allowing a fair comparison that showed improvements. For the more challenging English-Russian task, with a higher quality NMT as a basis, no system was able to improve the scores compared to the mt baseline. As approaches, again all teams used neural models, in most cases based on the Transformer architecture and using a multi-source approach (Lee et al., 2019; Pal et al., 2019; Xu et al., 2019). The approaches mostly differ in the way the Transformer was reused for APE, i.e., the way src and mt are encoded. Our own WMT 2019 submission (Pal et al., 2019) will be explained in more detail and with an extended analysis in chapter 11. Still, we want to review other well-performing approaches from WMT 2019.

The winner system by Lopes et al. (2019) ($wmt19^{nmt}_{best}$) uses a single pre-trained BERT (Devlin et al., 2019) encoder that receives both the source src and mt strings and applies a BERT-based encoder-decoder model. Additionally, they add a conservativeness penalty factor during beam decoding to avoid over-corrections.

In general, the results on English-German show the technological progress since 2018, however, the failed improvements over the English-Russian baseline show that APE does not always help, and, similar to MT, that APE has problems with morphologically rich languages like Russian, which leads to more sparse data that (in combination with the higher MT quality) made the task too challenging for the explored approaches. As in 2018, results (now the NMT results) showed that APE systems indeed correct towards pe and not towards independent reference translations, thus, showing APE's adaptation towards the style in the training material. The analysis of over-corrections by calculating TER/BLEU not only against pe but also against external references showed that the 2019 approaches do not have a tendency towards such over-corrections. Inspecting the submissions, they are all rather conservative and edit much less than required. Thus, the performance difference comes from precision on the edits and not from different levels of aggressiveness.

WMT 2020 In 2020, considerable improvements were achieved by APE on the non-domain adapted NMT output, showing the effectiveness of APE to improve MT quality towards a certain domain, even though this domain is not as repetitive as the IT domain. This shows APE's potential as a second-stage MT system performing domain or style adaption of a non-adapted MT system. Again most approaches were built upon the Transformer architecture in a multi-source fashion, either by concatenation src and mt or exploring multi-encoder approaches. Also the integration of the BERT language model, QE techniques, or ensembling of several approaches was analyzed. As in previous rounds of the task, creating synthetic corpora as additional training material was explored. Interestingly, the top-ranked system was ranked on the same level as human PE data, which might however have been influenced by the data and evaluation setting, thus, the task organizers by no means claim human parity even if the outcome suggests this. Nevertheless, it can be interpreted as further progress in APE. Note that the new (non-domain adapted) data source makes it hard to compare to previous rounds, and thus, does not show whether a well-optimized MT engine could also benefit from these APE approaches. One of the teams performed better than all others by a large margin. A notable aspect about their work was that they trained their Transformer-based model not only on concatenated src and mt but additionally added an auxiliary MT received from Google Translate, thus, yielding three inputs. This system modified more than 90% of the sentences while still achieving a precision of 0.69. In terms of applied changes, the 2020 approaches were not only more aggressive, but the lower quality data they operated on lead to many more structural changes (deletions, insertions, shifts) and much less lexical changes (substitutions) compared to previous rounds.

WMT 2021 Even though there was a similar number of data downloads in 2021 compared to the previous years (Akhbardeh et al., 2021), only 2 teams participated for the English-German, and no one for English-Chinese sub-task. The task organizers believe the reason for the lack of participants to be the especially challenging task. The systems by Amazon Prime Video relied on the fairseq NMT model, trained first on MT data from the news domain, then fine-tuned on APE data by concatenating src and mt as new src of the MT system, and using pe as "target" of the MT model. The other system proposed by the Netmarble AI Center instead focuses on multitask learning with a Transformer architecture, relying on data from tasks like speech recognition, named entity recognition, masked language modelling, and keep/translate classification to deal with data sparsity. In terms of automatic scores, all systems are roughly on par, with only one run submitted by Netmarble statistically outperforming the baseline. However, both teams significantly beat the baseline in terms of human evaluation, and Netmarble also significantly outperformed Amazon according to the human judgment. The finding that all systems were rated higher than the baseline in terms of human evaluation shows that APE is still a promising technology, however, it became harder to measure the improvements.

2.5.3 Summary

One negative aspect about PE frequently mentioned by translators is the need to correct repetitive mistakes (O'Brien and Moorkens, 2014) and the inability of the MT to properly learn from human corrections, which forms the basis of our third research question. This section presented APE as one approach to tackle this issue by adapting MT output based on a limited set of corrections. We have seen how the problem of APE has been addressed over the years by especially focusing on the WMT shared task, which nicely shows the transition from phrase-based to neural approaches, but also the evolution of evaluation schemes and data considerations. In our review, we particularly focused on the WMT 2018 and 2019 shared tasks, as we also explored two architectures working on this data, called the *Multi-Source Transformer* and *Transference* architectures, that we discuss in detail in chapters 10 and 11. One key take-away is that with the limited amount of real data available in the shared tasks, APE might not be able to improve highly domain optimized state-of-the-art NMT, but it is well suited to improve generic non domain-adapted models. As we will discuss in chapter 12, more data also allows to improve over specialized MT systems, and online APE can help avoid repetitive mistakes during PE.

2.6 Conclusion

We started our literature review with a brief history of MT, where we especially discussed that even though the quality is continuing to improve, there are few cases where the MT could be published without human intervention.

Thus, we presented the main paradigm of human-AI collaboration investigated in this dissertation: Post-Editing. Apart from discussing what PE is, we in particular focused on effort and time savings achieved through PE, quality improvements in terms of correctness, but also the loss in linguistic variety as well as translators' attitudes towards PE, which are by no means always positive.

We then looked at CAT tools, the translation environments used to perform translation and also PE. These are feature-rich, and a vast amount of CAT alternatives have been developed in academia and industry. Nevertheless, studies on user interface needs for CAT tools show that there is still room for improvement, especially because CAT tools were often designed with traditional translation in mind and only added MT as an additional feature, even though PE requires less text input but more text manipulation. A particularly related area of research is that of multi-modal CAT which we reviewed in depth. Part II, focusing on RQ1, contributes to this field by conducting a structured analysis of interaction modalities that might be suitable for PE. The results of this analysis are then transformed into a multi-modal CAT environment called Multi-Modal Post-Editing (MMPE), which is one of the main contributions of this dissertation.

Next, we looked at the cognitive dimension of PE, where we focused on cognitive load theory and approaches to measure CL. While a variety of measuring approaches have been proposed in research areas other than translation, many of them remain unexplored for PE. There are however, a variety of studies showing which sentence features are particularly hard to translate. Furthermore, we discussed the literature on privacy concerns regarding CL measures, as the mere possibility to estimate CL with a given sensor does not imply that users would accept their data being used for this purpose. Finally, we presented systems that directly react to measured CL to adapt the interface or task automatically. Part III builds upon these works to investigate RQ2, where we first interview translators to understand which CL adaptations might be useful for PE, then present a multi-modal CL estimation framework which we explore in three studies, and finally conduct a survey on the willingness to share data for this purpose.

Finally, we discussed Automatic Post-Editing as a tool to tackle the problem of correcting repetitive mistakes of the MT in PE. We introduced the WMT shared task on APE, which defines the problem and tracks progress in APE, and we discussed which approaches have been proposed over time. Our own APE models addressing RQ3 are presented in Part IV: first, we describe the Multi-Source Transformer, then the Transference architecture, and finally discuss how these could be integrated into CAT tools to reduce human PE effort by allowing domain-adaptation and translator-personalization.

Part II

Exploring Multi-Modal Interactions for Post-Editing of Machine Translation

As discussed in the literature review, PE of MT saves time and reduces errors compared to traditional translation from scratch (Green et al., 2013). Therefore, professional translators are gradually moving towards PE (Zaretskaya et al., 2015; Zaretskaya and Seghiri, 2018). The post-editing process differs significantly from traditional translation, which changes the interaction patterns to significantly less keyboard input but more navigational interactions (Carl et al., 2011; Green et al., 2013). The obvious question therefore is, whether mouse and keyboard are still the best interaction modalities for PE, or if other interaction modalities like handwriting or speech input might better support the correction of MT errors. Ideally, improved interaction possibilities could also enhance translators' perceptions of PE.

Thus, this part presents research towards a CAT environment that supports *explicit* input from a variety of interaction modalities. Chapter 3 presents an elicitation study (Vatavu and Wobbrock, 2015) with professional translators to integrate the target users early on in the design process. Based on these findings, chapter 4 presents MMPE, a multi-modal CAT interface allowing users to directly cross out or hand-write new text, drag and drop words for reordering, or use spoken commands to update the text in place. Apart from explaining the system capabilities, the chapter also discusses MMPE's advantages and disadvantages as captured in a controlled experiment with professional translators. Finally, chapter 5 summarizes and reflects upon improvements to MMPE based on these findings, and presents two additional studies conducted with the prototype: one focusing on mid-air hand gestures and the other on word-level QE for PE.

Part II is based on publications Herbig et al. (2019a), Herbig et al. (2020b), Herbig et al. (2020c), Herbig et al. (2020d), Jamara (2021), Jamara et al. (2021), Shenoy (2021), and Shenoy et al. (2021).

Chapter 3
Eliciting Multi-Modal Interactions for Post-Editing Machine Translation from Professional Translators

To involve target users early on and base the subsequent prototype development upon their visions, this chapter presents an elicitation study conducted with professional translators. For this study, we (a) propose a set of common PE operations (or referents) to figure out (b) which modalities translators find appropriate for which PE task, (c) what perceptions they have regarding modalities commonly used in HCI, and (d) how they envision an ideal translation environment setup for PE. We find that especially a digital pen, touch, speech, and a combination of pen and speech could support the different PE tasks well in a touch-friendly screen setup.

This chapter is based on publication Herbig et al. (2019a).

3.1 Evaluation Method

As an evaluation method, we chose an elicitation study (Vatavu and Wobbrock, 2015) paired with semi-structured interviews. We will first present the concept and literature on elicitation studies to properly introduce the key aspects. Then we will provide an overview of the different parts of our study, talk about the so-called referents (common operations), and modalities to achieve those referents with. The study has been approved by the university's ethical review board.

3.1.1 Elicitation Studies

Elicitation studies are a specific form of participatory design (Schuler and Namioka, 1993) and a common tool to design natural user interfaces. Such studies are often used in early stages of research for figuring out which interactions are most suitable for which tasks. Later stages then try to put the findings into a prototype that is again evaluated with users, to see how good the elicited interactions work in practice. Important aspects of such elicitation studies include leading participants away from technical thinking (Nielsen et al., 2003), making them assume that no recognition issues occur, and considering their behavior as always acceptable (Wobbrock et al., 2009). Furthermore, they should only be informed about the essential details of the task so as not to bias them towards existing approaches (Wobbrock et al., 2005). Instead, participants should be presented with so-called referents (i.e., common operations) and asked to propose actions to achieve these referents (Good et al., 1984). This approach has been shown to result in an increased immediate usage compared to highly iterated design approaches (Wobbrock et al., 2005). We use the formalization for elicitation studies by Vatavu and Wobbrock (2015), who define the **agreement rate** as the number of pairs of participants in agreement with each other, divided by the number of all possible pairs (Findlater et al., 2012). The introduced **coagreement rate** defines how much agreement two referents share, and a **significance test** for agreement rates is provided.

While most elicitation studies only explore the best interactions within a modality (e.g., for gesture input), Morris (2012) performed a multi-modal (speech + gesture) elicitation study. Here, users were allowed to suggest more than one interaction per referent which Vatavu and Wobbrock (2015)'s formulas do not support. For the analysis, Morris (2012) instead proposed the **max-consensus** (i.e., the percentage of participants proposing the most popular proposal) and the **consensus-distinct ratio** (i.e., the percent of distinct interactions for a given referent that achieved a predefined consensus threshold, here 2). Later, Morris et al. (2014) showed that elicitation studies are often biased by the user's experience with technology (called *legacy bias*), and discussed approaches against this bias: *production* (producing more proposals), *priming* (making them think about a specific technology before proposing), and *partners* (participating in a group).

In this chapter, an elicitation study with professional translators is used to determine which interaction modalities might be most suitable for which PE operation.

3.1.2 Study Overview

Initial Questionnaire

After providing informed consent to use the gathered data, participants are asked to fill in a general questionnaire capturing demographics as well as advantages and pain points of CAT tools. Similar to Wobbrock et al. (2009), we gather conceptual complexity ratings on a 5-point scale for all our referents to understand how difficult translators believe the distinct referents are without having proposed concrete actions yet, thus, capturing intuitive complexity.

Unbiased Elicitation

After this, we conduct a classical elicitation study similar to Vatavu and Wobbrock (2015). First, the general idea of a multi-modal CAT tool is introduced without biasing the participants. For this, we explain that "other interactions than the usual mouse and keyboard-based interactions" should be proposed, and that everything they come up with "could be perfectly recognized". Then, common operations (a.k.a. referents, see below) are presented and the participants are asked how they would perform this task. After each referent, they rate the *goodness* and perceived *ease of use* for the invented interaction (analog to Wobbrock et al. (2009) on a 7-point scale), and state on the same scale whether "the interaction I picked is a *good alternative* to the current mouse and keyboard approach". We intentionally specified that the interface could look whichever way the translators imagined it for the elicitation task, such as having multiple screens of arbitrary sizes and orientation, etc.

Biased Elicitation

After having talked about each referent without any prior bias (in the unbiased elicitation), we present all common interaction modalities (see below) and ask them again which modality (or combination thereof) they would use for which referent, but also to rate the different modalities on the same three scales as before and to discuss their decisions. Analogously to Morris (2012), we allow multiple proposals here, to support creativity. This second elicitation aims to avoid legacy bias, where the introduction to modalities can be seen as the priming strategy, while proposing multiple ideas is called production in Morris et al. (2014). Furthermore, this more guided process aims to counteract our participants' limited knowledge on interaction design.

Multi-Modal CAT Setup

Afterwards, we conduct a semi-structured interview to understand what the participants would imagine an ideal multi-modal CAT environment to look like, what kind of display devices would be located where, and how the interface parts would be arranged. We decided to put this setup discussion after the actual elicitation study, so that they can consider their proposed modalities and interactions when designing the interface.

General Notes on Methodology

The whole session is videotaped and participants receive a reimbursement for their time. The unbiased part of the experiment is necessary so that participants can think more broadly, which might lead to suggestions that are not within our list of modalities in the biased elicitation. The biased part ensures that subjects consider all suggested modalities, and provides more common ground for the participants. In both parts, we counter-balance the order of the referents using a balanced Latin square to avoid ordering effects.

3.1.3 Referents

The referents used in elicitation studies are an essential part, since the results are limited to this set. To find good referents, we look at different PE task classifications in the literature. Popovic et al. (2014) propose 5 PE operations: correcting word form, correcting word order, adding omission, deleting addition, and correcting lexical choice. Koponen (2012) additionally distinguishes between moving single words or groups and the distance of the movement. Temnikova (2010) further categorizes the addition or replacement of punctuation and the correction of mistranslated idiomatic expressions, and distinguishes between replacing a word with a different lexical item vs. with a different style synonym. Based on these works, which focused on investigating cognitive processes, we propose the classification depicted in Table 3.1 that we argue better captures the necessary operations from an interaction perspective.

Abbreviation	Name	Description
A	Addition	Missing word/punctuation that needs to be added/inserted
RO_s	Reorder single	Word order error that requires moving a *single item*
RO_g	Reorder group	Word order error that requires moving *multiple grouped items*
RP_s	Replace single	Incorrect word/punctuation that requires replacing with a *different item*
RP_p	Replace part	Word form error that requires replacing with a *different ending*
D_s	Delete single	Extra word/punctuation that requires deleting a *single item*
D_g	Delete group	Extra words/punctuations that requires deleting *multiple grouped items*

Table 3.1: Referents used for the elicitation study.

For each referent, we prepared a simple example that was presented to the participants orally, to provide a better understanding of the error concerned.

3.1.4 Modalities

The modalities introduced at the beginning of the biased elicitation are depicted in Table 3.2.

Abbreviation	Name	Description
MK	Mouse and Keyboard	Mouse and keyboard to be combined with other modalities
T	Touch	Finger-based touch screen input, including touch gestures
P	Pen	Digital pen/stylus used for touch input
G	Gestures	Mid-air hand gestures
S	Speech	Speech commands/dictation
E	Eye Tracking	Gaze positions, blinking, etc.
XY	Combinations	Combinations of X and Y, e.g., TS for Touch and Speech

Table 3.2: Modalities used for the biased elicitation study.

We explain each of these modalities to the subjects based on examples drawn from daily life (e.g., touch would be well-known from smartphones, or gestures from science fiction movies) and explain how they can be used (e.g., a pen to draw or for handwriting).

3.2 Evaluation Results and Discussion

In this section, we present the findings of each individual part of the study.

3.2.1 Participants and Conceptual Complexity

Overall 13 (female=9, male=4) professional translators participated in the experiment, 5 freelance and 8 in-house translators. Their ages ranged from 28 to 62 (mean=40.23, σ=9.11), with 2 to 34 years of professional experience (mean=13.65, σ=9.66) and a total of 39 language pairs (mean=3). For most participants the self-rated CAT knowledge was good (5 times) or very good (5 times). However, participants were less confident about their PE skills (6 neutral, 2 good, 5 very good), thereby matching well with the CAT usage surveys. Years of experience with CAT tools ranged from half a year to 18 years (mean=9.12, σ=5.23), where participants had used between 1 and 9 distinct CAT tools (mean=4.39, σ=2.18), most frequently using Trados Studio (13), Across (9), Transit (9), MemoQ (7), and XTM (7). Overall, participants are quite satisfied with their current CAT tools (mean=4.92, σ=1.04, on a 7-point scale). As most liked features, translators most often reported TM (9), terminology management (8), and concordance (7).

The ratings for the conceptual complexity of the referents on a 5-point scale (Wobbrock et al., 2009), with 5 being the most complex, are shown in Table 3.3. Overall, participants found reordering multiple words the most complex, followed by reordering a single word, deletion of multiple extra items, replacement of an item, addition of missing items, corrections of the word form, and last, deleting a single extra item. However, only the difference between reordering groups (RO_g) and deleting single items (D_s) is statistically significant ($p < 0.05$). The fact that reordering was rated as most complex is interesting because it intuitively is complex to perform with mouse and keyboard. In contrast, the typing tasks (addition and replace single/part) are perceived as less complex, probably since keyboards are well suited for this.

Referent	Average	σ
RO_g	4.08	0.86
RO_s	3.23	0.93
D_g	3.08	0.76
RP_s	2.92	0.64
A	2.69	1.18
RP_p	2.31	1.11
D_s	2.08	0.86

Table 3.3: Conceptual complexity as rated prior to eliciting interactions.

3.2.2 Unbiased Elicitation

Here we report the results of the initial, completely unbiased elicitation study, including agreement rates, co-agreement rates and proposed modalities.

Agreement Rates

We consider suggestions as equal if they consider the same modalities, i.e., different touch proposals are considered the same, while a touch and a pen proposal are considered distinct. The reason for this is that most proposals with the same modality could be supported in parallel, while the modalities have a direct impact on the way the setup should be designed. We found an average agreement rate for all referents of .282, which is comparable to the literature: Vatavu and Wobbrock (2015) found an average agreement rate of .261 (min=.108 with N=12 participants, max=.430 with N=14) in 18 elicitation studies, and calculated that 90% probability is already reached for an agreement rate of .374 for N=20. Since we have fewer participants, we recalculate the cumulative probability of our agreement rate for N=13, resulting in a cumulative probability of 67.3%, which is within the medium range [22.9%, 82%] proposed by Vatavu and Wobbrock (2015). A reason for the medium level agreement rate could be the interplay between a very restricted and well-known task (e.g., RP_s) and the flexibility of proposals, where similar interactions with different modalities (e.g., with pen vs. touch) were counted as distinct proposals.

The agreement rate and the corresponding cumulative probability with its interpretation per referent is shown in Table 3.4. The highest agreement was reached for replacing single items and can be interpreted as highly agreed upon, while all other agreement rates need to be interpreted as medium according to Vatavu and Wobbrock (2015). This suggests that replacement is intuitively solved with similar approaches, while less agreement was shared amongst the other referents; however, only the extreme differences between replacing single items and the deletions are significant ($p = 0.021$). Furthermore, Table 3.4 shows that all agreement rates are statistically significantly larger than 0.

Referent	**AR**	**CI$_{95\%}$**	**V$_{rd}$**	**P$_C$**	**Modalities**
A	.24	[.21,.49]	19*	.58 (m)	S(5), TS(4), MK(3), TpS
RO_s	**.21**	[.12,.54]	16*	.50 (m)	T(6), S(2), MK(2), P, ES, Tp
RO_g	.28	[.15,.59]	22*	.67 (m)	T(7), S(2), P, MK,MKS, Tp
RP_s	.46	[.24,.85]	36*	.88 (h)	S(9), T, TS, PS, TK
RP_p	.37	[.19,.72]	29*	.80 (m)	S(8), MK(2), T, TS, PS
D_s	**.21**	[.14,.47]	16*	.50 (m)	T(5), S(4), P, TS, MS, TpK
D_g	**.21**	[.14,.47]	16*	.50 (m)	T(5), S(4), P, TS, MK, TpK
All	.28	-	-	.67 (m)	-

Table 3.4: Agreement rate (AR), confidence intervals ($CI_{95\%}$), V_{rd} statistics against zero (* means $p = 0.001$), cumulative probability (P_C) and their interpretation (m=medium, h=high), as well as the proposed modalities (with number of proposal if >1) for all referents. The highest AR is shaded in cyan; the lowest ARs are marked in bold.

Co-agreement Rates

The co-agreement, i.e., the agreement shared between two referents, between reordering single and groups of items (.128 out of .205 = 62.4%) is lower than that between deleting single and groups of items (.167 out of .205 = 82.5%), suggesting that the two deleting referents are considered more similar than those for reordering. A very high co-agreement exists between the two replacements (.359 out of .372 = 96.5%), which suggests that replacing a single word or part of a word require similar interaction.

Proposed Modalities

Of the 91 total proposals, speech was most commonly suggested (34), followed by touch (25), mouse and/or keyboard (9), and touch combined with speech (8). Apart from this, several less frequent proposals were made (see below), of which 17 combined at least two modalities. More than half of all proposals

(48) involved speech, while 39 proposals contained touch. According to the subjective ratings on 7-point scales, participants thought their inventions were good (averages in range [5.4,6.6]), easy to use ([5.3,6.6]), and a good alternative to mouse and keyboard ([5.0,6.6]). Except for reordering groups (where it is equal), the majority proposal achieved higher rates on all three scales than the average among all proposals. While we did not test this for significance, it could indicate that participants choosing the majority class feel more confident about their proposal.

Speech

Speech was the majority proposal for the addition task (A) and replacement tasks (RP_s, RP_p), but was also proposed second most often for the deletions (D_s, D_g) (see Table 3.4). The suggestions were mostly trying to correct the mistake in place, e.g., "(add) X before/after Y"; however, restating the correct sentence was also suggested several times.

Participants appeared quite satisfied with their proposals, stating that speech becomes better the more changes are required, or that it "would reduce tiredness". For all tasks where speech was frequently proposed, it achieved average goodness ratings in the range [5.8,6.2], ease of use ratings within [6.0,6.4], and good alternative ratings within [5.6,6.0].

Touch

For reordering and deletion, touch was the majority suggestion: The idea was mainly to *select* the word(s) that need to be moved/deleted by simple tapping, encircling, or a long press followed by swiping over the words, and then *dragging* towards the final position/pressing a delete *button*. Other ideas were to reorder words by using several fingers simultaneously without prior selection, or to use touch gestures on top of the words to be deleted.

Participants again appeared enthusiastic regarding the use of touch, stating "I like this" or similar expressions. The importance of using a tilted screen and big buttons was also emphasized. The average ratings for the touch proposals were in the ranges [6.0,6.6], [5.9,6.6], and [5.3,6.6], for goodness, ease of use, and good alternative.

Touch and Speech

For the task of adding missing words (A), the combination of touch and speech was also proposed quite often for which it received average goodness ratings of 5.5, ease ratings of 5.3, and good alternative ratings of 5.5. The proposal was to place the finger at the correct position and verbally state "X" or "enter X" or "space X" (which shows the legacy bias of the keyboard).

Other Ideas

One participant liked the idea of having a **touchpad** for most referents while another frequently proposed the **digital pen**, both proposing it at times in combination with speech input. Infrequently, **eye tracking** and **gestures** or combinations thereof with other modalities were also proposed. One interesting idea from the proposals was to select words by touch, followed by a snapping gesture or swiping through the air, to make them disappear.

Some participants either were less creative or simply did not want to move away from the current **mouse and keyboard** approach. Overall, the classical methods were proposed 9 times and sometimes combined with speech.

Support tools were also often discussed: One participant suggested that when clicking on the space between two words for insertion, alternatives should appear to select from, while others asked for a list of different word forms or orderings when selecting a word. Similarly, it was proposed to integrate a thesaurus, where users can click on the word and see either synonyms, related words, or antonyms to better support stylistic corrections.

Discussion

The proposals indicate that speech and touch are by far the most relevant modalities. Overall, a medium level of agreement was reached among participants, with a lot more agreement for the replacement tasks than for the other referents. For reordering, touch was proposed most often; for replacement, speech; and for insertion and deletion, both (or a combination) were suggested. The high ratings also suggest that it is definitely worth investigating these modalities in practice. While still only a few multi-modal approaches were suggested (18.7%), this is already very high compared to Morris (2012) (3.1%). Even though we asked participants to propose modalities other than mouse and keyboard, a few participants were unable to come up with a different solution, which we see as a strong legacy bias (Morris et al., 2014). This can also be seen from the fact that most participants tended to propose *select first, then X* interactions known from the mouse and keyboard even for the new modalities.

3.2.3 Biased Elicitation

The second part of the elicitation study was conducted similar to Morris (2012) considering the priming and production strategies of Morris et al. (2014) to avoid legacy bias. In settings with multiple proposals per participant, agreement rates (Vatavu and Wobbrock, 2015) are not meaningful; instead, we use the metrics max-consensus and consensus-distinct ratio (Morris, 2012).

Participants proposed on average two interactions per referent, leading to 185 interactions overall. These can be clustered into 18 distinct modality combinations (compared to 13 in the unbiased study). The most proposed modalities are the pen (50), speech (31), touch (29), and pen combined with speech (17). This differs strongly from the unbiased elicitation, where the pen was only proposed four times while the percentage of touch and speech proposals was even higher. Table 3.5 and 3.6 summarizes the findings per referent. Here, the overall max-consensus and consensus-distinct ratios are again calculated on a modality level, while the ratios per modality distinguish the different proposals per modality, e.g., considering restating a whole sentence verbally as a distinct proposal from saying "replace X by Y", but considering the latter equal to "change X to Y". This allows to see both the consensus within and among modalities.

In contrast to the unbiased study, all participants came up with suggestions other than keyboard and mouse for every single referent and assigned high subjective ratings. This shows the importance of this second study phase to also elicit opinions from participants who are rather reluctant towards new approaches.

Referent	**Number (tot/dist)**	**MM% (tot/dist)**	**Common Proposals**
A	**21**/12	38.1/63.6	P(6), S(4), PS(3)
RO_s	29/7	**3.5/14.3**	P(10), T(8), E(4), S(3)
RO_g	27/**6**	3.7/16.7	P(11), T(8), S(4)
RP_s	25/10	48.0/60.0	S(8), PS(5), P(3)
RP_p	22/8	31.8/62.5	P(7), S(5), T(3)
D_s	33/14	48.5/71.4	P(7), T(5), S(4), PS(3)
D_g	28/14	57.1/71.4	P(6), T(3), PS(3)

Table 3.5: Proposals per referent in the biased elicitation study: (total and distinct) number of proposals, the percentage of multi-modal proposals (MM%), and modalities suggested ≥ 3 times. The highest scores are shaded in cyan, the lowest in bold text.

Referent	**ALL**		**P**		**T**		**S**		**PS**	
	$\mathbf{C_m}$	$\mathbf{C_d}$	$\mathbf{C_m}$	$\mathbf{C_d}$	$\mathbf{C_m}$	$\mathbf{C_d}$	$\mathbf{C_m}$	$\mathbf{C_d}$	$\mathbf{C_m}$	$\mathbf{C_d}$
A	**46.2**	**0.27**	83.3	0.50	-	-	75.0	**0.50**	100	1.00
RO_s	76.9	0.71	100	**0.33**	100	1.00	66.7	**0.50**	-	-
RO_g	84.6	0.66	90.9	0.50	87.5	**0.50**	**50.0**	1.00	-	-
RP_s	61.5	0.50	66.7	0.50	-	-	100	1.00	80.0	**0.50**
RP_p	53.9	0.62	100	1.00	66.7	**0.50**	80.0	**0.50**	-	-
D_s	53.9	0.57	85.7	0.50	**60.0**	1.00	100	1.00	100	1.00
D_g	**46.2**	0.57	100	1.00	66.7	**0.50**	-	-	**67.7**	**0.50**

Table 3.6: Proposals per referent in the biased elicitation study: the max-consensus (C_m) and consensus-distinct (C_d, threshold = 2) ratios for all and the most frequent modalities. The highest scores are shaded in cyan, the lowest in bold text.

Most proposals were given for deleting single items, the least for additions (total) and reordering groups (distinct). Compared to Morris (2012) (3.1%) and our unbiased study (18.7%), we see far more multi-modal proposals (mean=33.0%). Only the reordering referents received few multi-modal suggestions, probably due to the high agreement on pen and touch, which do not require a secondary interaction modality.

Inspecting the most common proposals, it seems that pen, speech, touch, and the combination of pen and speech are by far the most important ones. The highest max-consensus ratio (on modality level) was achieved for reordering with a pen. Regarding the consensus-distinct ratio, the most consensus was again achieved for the reordering referents, the lowest for additions. The overall average ratings among all proposals for all referents were 6.0, 5.9, and 5.8 out of 7 for goodness, ease of use, and good alternative, respectively, showing a high level of satisfaction of the participants towards their proposals.

Pen

We see a max-consensus ratio of more than 80% for pen for all but one referent, showing that participants would use the pen in a rather similar fashion. The consensus-distinct ratio ranges from 0.33 to 1.00, meaning that either all agreed, or there were one or two single alternative suggestions to the majority opinion. Most translators suggested to simply write at the correct position for additions and replacements (after strike-through), to select (e.g., by encircling or underlining) and drag for reordering, and to use a strike-through for deletions. One participant proposed the interactions in a proofreading style (e.g., using a missing sign or drawing arrows) while another translator suggested to use a button integrated into the pen (e.g., to pick up one or multiple words). In general, participants were quite enthusiastic about the pen, also stating that it would be good since it is more precise than touch. This can also be seen in the high average ratings, with a goodness of 6.1, an ease of use of 6.1, and alternative being 5.9.

Touch

Touch was commonly suggested for all referents except additions (A) and replacing single items (RP_s), probably because these two require the most generation of text. The max-consensus for touch is on average lower than that for pen, and the consensus-distinct ratio varies between 0.5 and 1.0; this taken together with the concrete touch proposals shows that participants agreed less on how to interact with touch. Most proposals again took the form of selection followed by some other action and the ratings were also very good with an average of 6.2 for goodness, 5.9 for ease of use, and 6.0 for alternative.

Speech

Speech was among the set of common suggestions for all referents except deleting groups (D_g). In this case the max-consensus ratio is also on average lower than that for pen, and the consensus-distinct ratio ranges from 0.5 to 1.0. Inspecting the data, we see that it all boils down to the two proposals *correct in place* (e.g., "Add X after Y") or *restate correct sentence*. One should note that except for deleting and reordering groups (D_g and RO_g), the option *correct in place* was always the one favored by our participants. This makes sense as for more complex sentences it might be simpler to reformulate the correct sentence, a fact that a participant also pointed out as an explanation. Speech ratings were a bit lower, but still good, having 5.8 for goodness, 5.9 for ease of use, and 5.5 for alternative.

Pen and Speech

The combination of pen and speech was proposed for all referents except for reordering groups (RO_g), with all proposers agreeing completely for additions (A), replacing parts (RP_p), and deleting single items (D_s). More distinct suggestions were provided for replacing single items (RP_s) and deleting groups (D_g). The suggestions were all very straightforward, either placing the pen at a specific position or selecting the important parts and then uttering a speech command. The combination of pen and speech received the highest subjective ratings: 6.2 for both goodness and ease, and 6.3 for alternative.

Other Ideas

Apart from these common modalities, many other suggestions were made, although most of them without consensus. Many suggestions combined **eye/touch/touchpad/pen with a keyboard or speech**. Interestingly, the keyboard was sometimes only integrated for the delete key, but it was also suggested by several participants who did not like handwriting. **Gestures** were occasionally paired with speech or with eye tracking, e.g., by looking at an item and then shaking the head or swiping through the air. It was also pointed out that gestures would require a large screen, could activate muscles, and might be easy to perform, but would initially require training. **Eye tracking** was also proposed several times, e.g., by picking up and dropping words by blinking or combined with a touch button or speech input. However, most participants were less optimistic regarding eye tracking, expecting it to require high concentration.

As in the unbiased elicitation, several ideas for **support tools** arose: again the idea of receiving a list of word forms, synonyms etc., which could be selected with any interaction modality, was proposed. Furthermore, eye tracking was suggested to mark the position the translator last looked at within the source and target text, to avoid getting lost when scrolling, similar to the Gazemarks approach (Kern et al., 2010).

In general, several participants argued for multiple approaches working simultaneously to avoid tiredness and to be able to rest the hands or the voice from time to time. Participants were often also confident that their suggestion could be faster than mouse and keyboard.

Discussion

We find that both reordering and deletion tasks would be best supported by pen and touch and to a lesser degree speech. For the insertion and replacement tasks that also require the generation of new text, pen, speech, or pen combined with speech appear to be most promising. In general, the pen was suggested very often by participants and based on the discussions, they really liked the idea of a digital pen for PE. Participants were often also confident that their suggestion could be faster than mouse and keyboard. Apart from reordering, there were lots of multi-modal suggestions, most commonly pen and speech, but also (albeit without consensus) a lot of eye/gesture + X approaches. Several participants also argued for multiple approaches working simultaneously to avoid tiredness and to be able to rest the hands or the voice from time to time. The legacy bias appeared lower in this biased elicitation, which can be seen in the vast amount of suggestions of modalities that people do not usually use in their daily lives, and the fact that no one came up only with mouse and keyboard, as was the case in the first unbiased elicitation study. Therefore, we argue that it is worth doing this two-step setup, as it provides insights into the overall user model but also into their thoughts on possible approaches.

3.2.4 Multi-Modal CAT Setup

Next, we discussed what the participants would imagine their CAT system to look like. Regarding screens, 9 participants preferred having only a single screen; however, we see a clear tendency towards **big screens**: only 3 participants (two of whom wanted more than 1 screen) argued for a normal screen size, while 7 requested a big, and 3 even a giant screen (e.g., flipchart-sized). 2 translators proposed editing on a touchscreen placed on the table combined with other tools above this editing area. In general, 7 participants argued for a **tiltable** screen for better adjustment of the viewing angle and improved touch interaction. This is in line with the feedback we received during the two elicitation tasks, where most participants that proposed a touch or pen interaction simultaneously argued for a tilted screen layout. Furthermore, 3 participants asked for a setup that allows one to work both in a seated and a standing position. Apart from this relatively straightforward setup, the idea to integrate hardware buttons or even interact with the feet was also proposed, e.g., to confirm segments.

Regarding the interface, the arrangement of source and target text was a widely discussed topic with most translators (9/11) arguing for a **horizontal layout**. The **integration of browser functionality** like an online corpus, synonyms, encyclopedia, dictionaries, forums, etc. into the interface was also mentioned 4 times. Displaying the text in the correct document format was another emerging topic: 3 proposed to see a preview of the target, 1 wanted to see the source, and another requested both.

Many different arrangements of common CAT features were proposed. Some discussed the importance of having everything relevant (TM, dictionary, etc.) on the same horizontal level as the current segment, while others also proposed more vertical arrangements. One participant argued for movable interface elements, which, given the amount of distinct layouts proposed, is probably the best and only good option, even if it remains to be seen if this customization would be used in practice (see Coppers et al. (2018)). 2 argued for **enlarging only the current segment**, allowing the user to view a lot of context and still see the current segment in a comfortable manner. This would also facilitate pen or touch interaction, as it offers more space.

Further interesting ideas were to enlarge words upon selection, and to read back text (text-to-speech) while reading with the eyes to detect errors more easily. One participant proposed to ambiently display images of words you are looking up or currently editing within the room (through a projection) while another participant involved in technical translation proposed to display 3D visualizations of the machines that the text is about.

3.2.5 Limitations

Since the whole study was based only on elicited ideas, all findings need to be verified on a working prototype. Only after such tests would it be possible to fairly compare the techniques, including all potential technical limitations.

3.3 Conclusion

Due to the changed task and interaction patterns compared to classical translation, we sought out to investigate the PE process more strongly from an interaction modality perspective. For a well-structured test, we first propose a set of common operations (referents) necessary for PE and find that translators believe reordering tasks to be the most complex. After having run an elicitation study using this set of referents, we find that a lot of agreement is shared between replacing single items and parts of items, and thus would reduce our proposed set to addition (A), reorder single/group (RO_s/RO_g), replacement (RP), and delete single/group (D_s/D_g) in the future. Note that this set of operations appears more suitable for NMT, which produces very fluent text having mostly correct word forms, but tends to make wrong lexical choices, as it prioritizes word-level manipulations over sub-word-level operations.

However, our initial, completely unbiased elicitation study, showed that participants mostly envision touch and speech modalities for these referents, while other modalities were only rarely suggested. We believe this to come from the limited background of the participants with interactive systems and a strong legacy bias. However, the participants' high subjective ratings indicate that they were quite satisfied with their proposals.

After having introduced the translators to a set of common modalities, their proposals changed, where now the digital pen was considered a favorite together with speech, and touch. Taken together the ratings and the strong agreement on these three modalities, we argue that one should move away from mouse and keyboard-only approaches and investigate such interactions in practice. Based on the statements that several modalities should work in parallel to avoid monotonicity and thereby possibly fatigue, we believe that these commonly proposed modalities should all be supported for every single referent. This would allow switching the interaction mode but avoid enforcing such modality shifts as it could irritate users if some modalities only work for a subset of referents. In contrast, other modalities like eye tracking or gestures appear less promising for this application area.

For the overarching goal of this part of the thesis, namely to investigate whether interaction modalities other than mouse and keyboard could well support the PE process, this initial chapter ensured a user-centered investigation and basic understanding of the potential of different modalities. Furthermore, the initial unbiased approach aimed to ensure that no important modalities are missed as might have happened in a purely developer-driven implementation. Finally, the qualitative findings in the various interviews throughout the two elicitation studies, as well as the semi-structured interview on CAT design allowed to grasp the mindset and user needs of professional translators, thereby helping us to make more suitable design decisions in the chapters ahead.

Chapter 4
MMPE: A Multi-Modal Interface using Handwriting, Touch Reordering, and Speech Commands for Post-Editing Machine Translation

Based on the findings elicited from professional translators, this chapter now presents MMPE, the first translation environment combining standard mouse & keyboard input with touch, pen, and speech interactions for PE of MT. We will first present the prototype, allowing users to directly cross out or hand-write new text, drag and drop words for reordering, or use spoken commands to update the text in place. All text manipulations are logged in an easily interpretable format (e.g., *replaceWord* with the old and new word) to facilitate translation process research. We evaluate this prototype in a study with 11 professional translators, showing that participants are enthusiastic about these alternatives, and suggesting that pen and touch are well suited for deletions and reorderings, whereas speech and multi-modal input is suitable for insertions and replacements.

This chapter is based on publications Herbig et al. (2020b) and Herbig et al. (2020c).

4.1 Prototype

This section presents the MMPE prototype, which combines pen, touch, and speech input with a traditional mouse and keyboard approach for PE of MT. The prototype is designed for professional translators in an office setting. A video demonstration is available at `https://youtu.be/tkJ9OWmDd0s`. MMPE specifically focuses on traditional PE, in contrast to translation from scratch or interactive machine translation.

4.1.1 Apparatus

On the software side, we decided to use Angular[25] for the frontend, and node.js[26] for the backend. The frontend, including all of the newly implemented modalities for text editing, is what the system currently focuses on. While this Angular frontend could be used in a browser on any device due to the usage of Bootstrap[27], we initially design for the following hardware to optimally support the implemented interactions:

As requested by the majority of translators in the elicitation study reported in chapter 3, we use a large tiltable touch & pen screen (see Figure 4.1), namely the Wacom Cintiq Pro 32 inch display. Together with the Flex Arm, the screen can be tilted and moved flat on the table (similar to how users use a tablet), can be vertically positioned like a normal screen, or be moved up in the air to work in a standing position. Wacom displays such as this one are rather expensive (ca. 3500 Euro[28]) but also known for very accurately recognizing pen input at different angles and pressure levels with little delay, since they target professional drawing, design, sketching, as well as image and video editing. We further use the Sennheiser PC 8 Headset (ca. 35 Euro[29]) for speech input, although informal tests showed that the ASR indeed does work well with most headsets and only suffers with microphones integrated in laptops. Last, mouse and keyboard are provided. The goal of this setup was to limit bias induced as much as possible, in order to get results on the modalities and not on a flawed hardware setup.

Since it is not the focus of this work, the backend is kept rather minimal: It allows saving and loading of projects from JSON files, can store log files, etc. Here, the project files simply contain an array of segments with source, target, as well as any MT or TM proposal that should initially be shown for PE.

4.1.2 Overall Layout

As can be seen in Figure 4.2 and as requested in our initial elicitation study, we implemented a horizontal source-target layout, through which the user can vertically scroll using the mouse wheel or a touch drag. Similar to well-known CAT tools like SDL Trados Studio, each segment's status (unedited, edited, confirmed) is visualized between source and target. On the far right, support tools are of-

[25] `https://angular.io/`

[26] `https://nodejs.org/en/`

[27] `https://getbootstrap.com/`

[28] At the time of writing.

[29] At the time of writing.

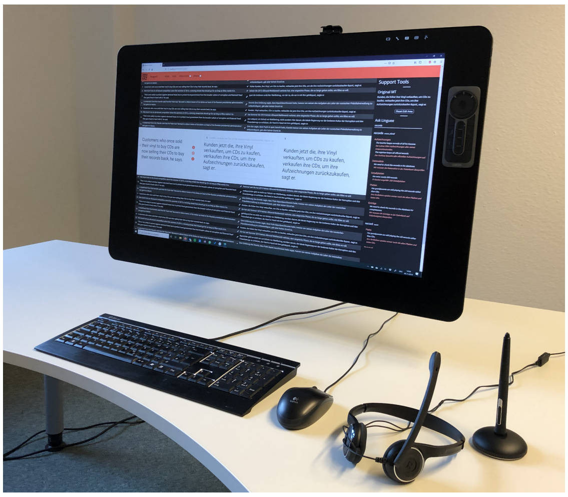
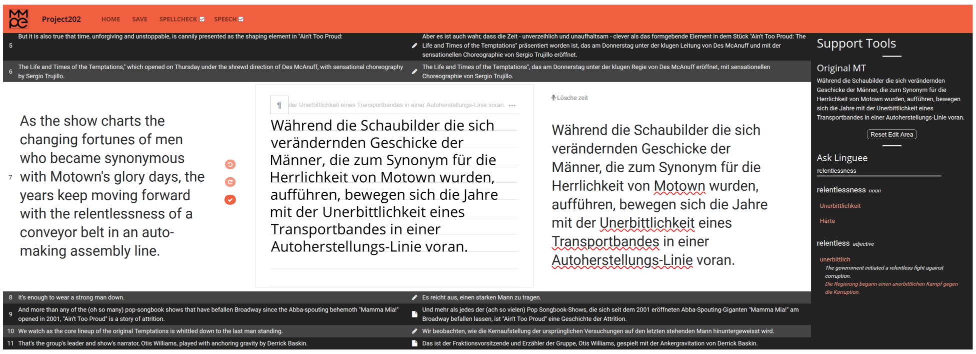


Figure 4.1: Apparatus of MMPE.

fered as requested in the semi-structured interview on the interface setup: (1) the unedited MT output, to which the user can always revert to during their editing using a button, and (2) a corpus combined with a dictionary: when entering a word or clicking/touching a word in the source view on the left, the Linguee[30] website is queried to show the word in context and display its primary and alternative translations. The top of the interface shows a toolbar where users can enable or disable speech recognition and spell checking, and save and load projects, or navigate to another project.



5 But it is also true that time, unforgiving and unstoppable, is cannily presented as the shaping element in "Ain't Too Proud: — Aber es ist auch wahr, dass die Zeit - unverzeihlich und unaufhaltsam - clever als das formgebende Element in dem Stück "Ain't Too Proud: The Life and Times of the Temptations" präsentiert worden ist, das am Donnerstag unter der klugen Leitung von Des McAnuff und mit der sensationellen Choreographie von Sergio Trujillo eröffnet.

6 The Life and Times of the Temptations," which opened on Thursday under the shrewd direction of Des McAnuff, with sensational choreography by Sergio Trujillo. — The Life and Times of the Temptations", das am Donnerstag unter der klugen Regie von Des McAnuff eröffnet, mit sensationellen Choreographie von Sergio Trujillo.

7 As the show charts the changing fortunes of men who became synonymous with Motown's glory days, the years keep moving forward with the relentlessness of a conveyor belt in an auto-making assembly line.

¶ der Unerbittlichkeit eines Transportbandes in einer Autoherstellungs-Linie voran. ...

Während die Schaubilder die sich verändernden Geschicke der Männer, die zum Synonym für die Herrlichkeit von Motown wurden, aufführen, bewegen sich die Jahre mit der Unerbittlichkeit eines Transportbandes in einer Autoherstellungs-Linie voran.

Während die Schaubilder die sich verändernden Geschicke der Männer, die zum Synonym für die Herrlichkeit von Motown wurden, aufführen, bewegen sich die Jahre mit der Unerbittlichkeit eines Transportbandes in einer Autoherstellungs-Linie voran.

Original MT: Während die Schaubilder die sich verändernden Geschicke der Männer, die zum Synonym für die Herrlichkeit von Motown wurden, aufführen, bewegen sich die Jahre mit der Unerbittlichkeit eines Transportbandes in einer Autoherstellungs-Linie voran.

Ask Linguee: relentlessness

relentlessness *noun*
Unerbittlichkeit
Härte

relentless *adjective*
unerbittlich
The government initiated a relentless fight against corruption.
Die Regierung begann einen unerbittlichen Kampf gegen die Korruption.

8 It's enough to wear a strong man down. — Es reicht aus, einen starken Mann zu tragen.

9 And more than any of the (oh so many) pop-songbook shows that have befallen Broadway since the Abba-spouting behemoth "Mamma Mia!" opened in 2001, "Ain't Too Proud" is a story of attrition. — Und mehr als jedes der (ach so vielen) Pop Songbook-Shows, die sich seit dem 2001 eröffneten Abba-Spouting-Giganten "Mamma Mia!" am Broadway befallen lassen, ist "Ain't Too Proud" eine Geschichte der Attrition.

10 We watch as the core lineup of the original Temptations is whittled down to the last man standing. — Wir beobachten, wie die Kernaufstellung der ursprünglichen Versuchungen auf den letzten stehenden Mann hinuntergeweisst wird.

11 That's the group's leader and show's narrator, Otis Williams, played with anchoring gravity by Derrick Baskin. — Das ist der Fraktionsvorsitzende und Erzähler der Gruppe, Otis Williams, gespielt mit der Ankergravitation von Derrick Baskin.

Figure 4.2: Screenshot of the MMPE interface.

The current segment is enlarged, thereby offering space for handwritten input and allowing the user to view a lot of context while still seeing the current segment in a comfortable manner (as requested in the interviews, see Figure 4.2). The view for the current segment is further divided into the source segment (left) and two editing planes for the target, one for handwriting and drawing gestures (middle), and one for touch deletion & reordering, as well as standard mouse

[30] `https://www.linguee.com/`

and keyboard input (right). Naturally, for mouse and keyboard all common navigation inputs and shortcuts work as expected from other software (e.g., ctrl+arrow keys or ctrl+c). Both editing planes initially show the MT proposal and synchronize on changes to either one. The reason for having two editing fields instead of only one is that some interactions are overloaded, e.g., a touch drag can both be interpreted as handwriting (middle) and reordering (right). As we will see in the evaluation, having two views next to each other was considered problematic by our participants, thus, the two views were later on redesigned into tabs as described in chapter 5. Undo and redo functionality, as well as confirming segments, are also implemented through buttons between the source and target texts, and can further be triggered through hotkeys. The target text is spell-checked, as a lack of this feature was criticized in Teixeira et al. (2019).

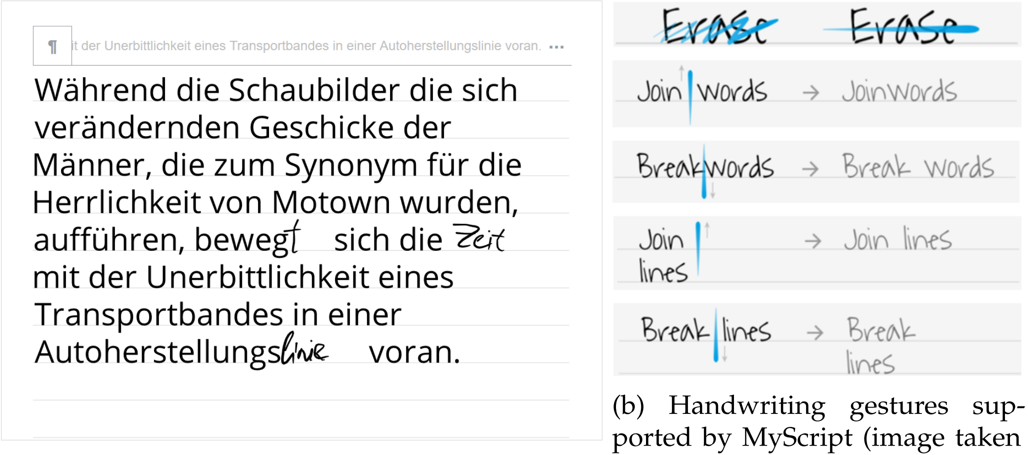
4.1.3 Handwriting

For handwriting recognition (see Figure 4.3a), we use the MyScript Interactive Ink SDK[31] which is one of the market leaders in handwriting recognition. Apart from merely recognizing the written input, it offers gestures[32] like strike-through or scribble for deletions, breaking a word into two (draw line from top to bottom), and joining words (draw line from bottom to top). For inserting words, one can directly write into empty space, or create such a space first by breaking the line (draw a long line from top to bottom), hand-writing the word, and joining the lines again (draw line-crossing line from top to bottom). While these gestures are similar to the ones requested in our elicitation study above, they are not identical. However, since they are the same as the ones used in the widespread MyScript Nebo note taking app[33], we decided to use these well designed and tested gestures instead of defining new ones ourselves. All changes are immediately interpreted, i.e., striking through a word deletes it immediately instead of showing it in a struck-through visualization. While reordering is not supported out of the box, as it is less common in normal note-taking scenarios, we allow reordering words with the pen on the right-hand editing view, as described in the next section. The editor further shows the recognized handwritten text immediately at the very top of the drawing view in a small gray font (see Figure 4.3a), where alternatives for the current recognition are offered when clicking on a recognized word. Since all changes from this drawing view are immediately synchronized into the right-hand view, the user can also see the recognized text there. While it is not necessary to convert text from the handwritten appearance into computer font, the user can do so using a small button at the top of the editor. Apart from using the pen, the user can use his/her finger or the mouse on this handwriting view.

[31] https://developer.myscript.com/

[32] https://developer.myscript.com/docs/concepts/editing-gestures/

[33] https://www.nebo.app/



¶ it der Unerbittlichkeit eines Transportbandes in einer Autoherstellungslinie voran. ...

Während die Schaubilder die sich verändernden Geschicke der Männer, die zum Synonym für die Herrlichkeit von Motown wurden, aufführen, bewegt sich die Zeit mit der Unerbittlichkeit eines Transportbandes in einer Autoherstellungslinie voran.

(a) Handwriting on left target view.

(b) Handwriting gestures supported by MyScript (image taken from the MyScript website[34]).

Figure 4.3: Handwriting in MMPE.

4.1.4 Touch Reordering & Deletion

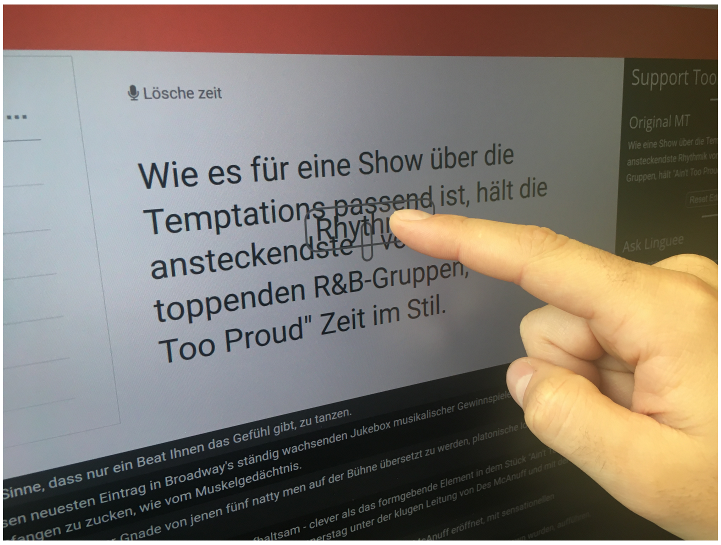
Further touch input (i.e., both finger and pen interaction) is supported on the right-hand editing view: The user can delete words by simply double-tapping them, or reorder them through a simple drag and drop procedure (see Figure 4.4). This procedure visualizes the picked-up word as well as the current drop position through a placeholder element. Spaces between words and punctuation marks are automatically fixed, i.e., double spaces at the pickup position and missing spaces at the drop position are corrected. This reordering functionality is strongly related to Teixeira et al. (2019); however, instead of having all words in a tiled view, which has been criticized to lead to bad readability (see Figure 2.20), only the currently dragged word is temporarily visualized as a tile to offer better readability. Furthermore, the cursor can be placed between words using a single tap, allowing the user to combine touch input with e.g., the speech or keyboard modalities (see below).

4.1.5 Speech Input

For speech recognition, we stream the audio recorded by the headset to IBM Watson[35] servers to receive a transcription, which is then analyzed in a command-based fashion. Thus, our speech module not only handles dictations as in Teixeira et al. (2019), but can correct mistakes in place.

The transcription itself is visualized at the top of the right target view (see Figure 4.4). As commands, the user has the option to "*insert*", "*delete*", "*replace*", and "*reorder*" words or subphrases. To specify the position if it is ambiguous,

[34] https://developer.myscript.com/docs/concepts/editing-gestures/
[35] https://www.ibm.com/cloud/watson-speech-to-text



Wie es für eine Show über die Temptations ~~passend~~ ist, hält die ansteckendste Rhyth ve toppenden R&B-Gruppen, Too Proud" Zeit im Stil.

Wie eine Show über die Tem ansteckendste Rhythmik vo Gruppen, hält "Ain't Too Prou

Sinne, dass nur ein Beat Ihnen das Gefühl gibt, zu tanzen.
sen neuesten Eintrag in Broadway's ständig wachsenden Jukebox musikalischer Gewinnspiele
fangen zu zucken, wie vom Muskelgedächtnis.
Gnade von jenen fünf natty men auf der Bühne übersetzt zu werden, platonische
haltsam - clever als das formgebende Element in dem Stück "Ain't
erstag unter der klugen Leitung von Des McAnuff und mit de
McAnuff eröffnet, mit sensationellen

Figure 4.4: Touch reordering on right target view in MMPE.

one can define anchors as in *"after"*/*"before"*/*"between"*, or define the occurrence of the token (*"first"*/*"second"*/*"last"*). A full example is *"insert A after second B"*, where A and B can be words or subphrases. In contrast to the other modalities, character-level commands are not supported, so instead of deleting an ending, one should replace the word. For the subsequent study, we set the speech recognition to German and also added a few synonyms for each of the four cases (e.g., 'lösche' and 'entferne' for 'delete') to allow for more flexibility. Apart from enabling and disabling speech in the navigation bar, the headset further provides a hardware button for muting, which turned out to be very useful in preventing erroneous recognitions when participants were simply talking to experimenters. Again, spaces between words and punctuation marks are automatically fixed upon changes. For the German language, nouns are automatically capitalized using the list of nouns from Wiktionary[36].

4.1.6 Multi-Modal Combinations

Oviatt (2003) argues that multi-modal combinations often work better than uni-modal systems. Therefore, MMPE further offers multi-modal combinations, i.e., pen/touch/mouse combined with speech. For this, a target word/position first needs to be specified by placing the cursor on or next to a word using the pen, finger touch, or the mouse/keyboard; alternatively, the word can be long-pressed with pen/touch. Afterwards, the user can use a voice command like *"delete"*, *"insert A"*, *"move after/before A/between A and B"*, or *"replace by A"* without needing to specify the position/word, thereby making the commands less complex.

[36] https://en.wiktionary.org/wiki/Category:German_noun_forms

4.1.7 Logging

We implemented extensive logging functionality: On the one hand, we log the concrete keystrokes, touched pixel coordinates, etc.; on the other hand, all UI interactions (like *segmentChange* or *undo/redo/confirm*) are stored, allowing us to analyze the translator's use of MMPE.

Most importantly, however, we also log all text manipulations on a higher level to simplify text editing analysis: For *insertions*, we log whether a single or multiple words were inserted, and add the actual words and their positions as well as the segment's content before and after the insertion to the log entry. *Deletions* are logged analogously, and for *reorderings*, we add the old and the new position of the moved words to the log entry. Last, for *replacements*, we log whether only a part of a word was replaced (i.e., changing the word form), whether the whole word was replaced (i.e., correcting the lexical choice), or whether a group of words was replaced. In all cases, the words before and after the change, as well as their positions and the overall segment text are specified in the log entry.

Furthermore, all log entries contain the modality that was used for the interaction, e.g., speech or pen, thereby allowing the analysis of which modality was used for which editing operation. All log entries with their timestamps are created within the Angular client and sent to the node.js server for storage in a JSON file.

4.2 Evaluation Method

The prototype was evaluated with professional translators. The study has been approved by the university's ethical review board. Freelance participants were paid their usual fee, while in-house translators participated during working hours. The data and analysis scripts can be found at `https://mmpe.dfki.de/data/ACL2020/`. We used EN-DE text, as our participants were German natives and we wanted to avoid ASR recognition errors as reported in Dragsted et al. (2011). In the following, "modalities" refers to Touch (T), Pen (P), Speech (S), Mouse & Keyboard (MK), and Multi-Modal combinations (MM), while "operations" refers to Insertions, Deletions, Replacements, and Reorderings.

The study took approximately 2 hours per participant and involved three separate stages. First, participants filled in a questionnaire capturing demographics as well as information on CAT usage. In stage two, participants received an explanation of all of the prototype's features and then had time to explore the prototype on their own and become familiar with the interface. Finally, stage three included the main experiment, which was a guided test of all implemented features combined with Likert scale ratings and interviews, as described in detail below. In the end, a final unstructured interview to capture high-level feedback on the MMPE prototype was conducted.

4.2.1 Introduction & Independent Post-Editing

After providing informed consent to use the gathered data, participants filled in a questionnaire capturing demographics as well as information on CAT usage. Then the experimenter introduced all of the prototype's features in a prepared order to ensure a similar presentation for all participants. The project used during this introduction phase contained a text on the newly announced Game of Thrones Prequel[37] as source, which was pre-translated using IBM Watson.

After that, participants were given 10-15 minutes to explore the prototype by PE a Broadway musical review[38] on their own. We specifically told them that we are more interested in them exploring the presented features than in receiving high-quality translations, and that they do not have to post-edit the whole text as this would not be feasible in the given time. This phase had two main purposes: (1) to let the participants become familiar with the interface (e.g., how to best hold the pen) and to resolve questions early on; (2) to see how participants intuitively work with the prototype without being biased. Furthermore, the hope was to identify missing features in the current implementation. Two experimenters carefully observed the participants' behavior and took notes on interesting behavior and questions asked. Furthermore, the interactions were logged by the system as described above.

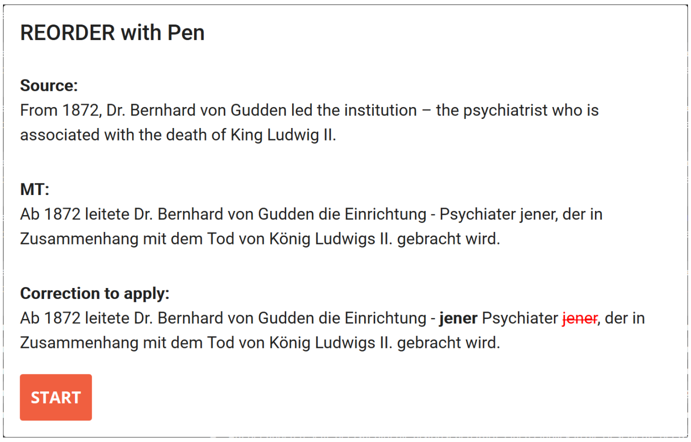
4.2.2 Feature-Wise & General Feedback

The central part of the study was a structured test of each modality for each of our four operations. For this, we used text from the WMT news test set 2018. Instead of actually running a MT system to generate the initial version of the target sentences, we manually introduced errors into the reference set to ensure that there was only a single error per segment. Overall, four sentences had to be corrected per operation (4) using each modality (5), which results in $4 \times 4 \times 5 = 80$ segments per participant. Within the four sentences per operation, we tried to capture slightly different cases, like deleting single words or a group of words, or replacing the ending or a whole word. For this experiment, we adapted the prototype such that a pop-up occurs when changing the segment, which shows (1) the operation to perform and which modality to use, (2) the source and the "MT", which is the reference with the introduced error, as well as (3) the correction to apply, which uses color, bold font, and strike-through to highlight the required change (see Figure 4.5). Participants were instructed to only perform this single requested change, even if they would rephrase the translation for stylistic reasons.

The reason why we provided the correction to apply was to ensure a consistent editing behavior across all participants, thereby making subjective ratings and

[37] https://edition.cnn.com/2019/10/29/entertainment/game-of-thrones-prequel-house-of-the-dragon-hbo/index.html

[38] https://www.nytimes.com/2019/03/21/theater/aint-too-proud-review-the-temptations.html



REORDER with Pen

Source:
From 1872, Dr. Bernhard von Gudden led the institution – the psychiatrist who is associated with the death of King Ludwig II.

MT:
Ab 1872 leitete Dr. Bernhard von Gudden die Einrichtung - Psychiater jener, der in Zusammenhang mit dem Tod von König Ludwigs II. gebracht wird.

Correction to apply:
Ab 1872 leitete Dr. Bernhard von Gudden die Einrichtung - **jener** Psychiater ~~jener~~, der in Zusammenhang mit dem Tod von König Ludwigs II. gebracht wird.

Figure 4.5: Popup giving instructions during the study.

feedback as well as time measurements comparable. The logging functionality was extended, such that times between clicking "Start" and confirming the segment were also logged. To avoid ordering effects, the participants went through the operations in counter-balanced order, and through the modalities in random order. After every operation (i.e., after $4 \times 5 = 20$ segments) and similar to our elicitation study, participants rated each modality for that operation on three 7-point Likert scales ranging from "strongly disagree" to "strongly agree", namely whether the interaction "is a good match for its intended purpose", whether it "is easy to perform", and whether it "is a good alternative to the current mouse and keyboard approach". Furthermore, we asked the translators to give us their thoughts on advantages and disadvantages of the modalities, and how they could be improved. Afterward, participants further had to order the 5 modalities from best to worst for each operation.

4.2.3 Remarks Regarding Methodology

While a direct comparison to state-of-the-art CAT tools would be interesting, the results would be highly questionable as the participants would be expert users of their day-to-day tool and novice users of our tool. Furthermore, the focus of our prototype was on the implemented modalities, while widely used features of fully-fledged CAT tools like a TM or consistency checker are currently missing.

Therefore, our study focused on qualitative feedback, which is more relevant for the main question, namely which of the newly implemented features have potential for which PE operations. To not merely gather free-form statements from the participants, we further captured ratings, orderings and timing information of the modalities for the analyzed operations. To ensure comparable results here, we did not use actual MT output but text with manually prepared errors.

4.3 Results

We first present the demographics of our participants, then report findings on hardware setup, layout, and interface design, before we go over to the subjective ratings, orderings, and the required time per operation with each modality. In the end, we report further qualitative results on the specific modalities, as well as missing features and general feedback.

4.3.1 Participants

Overall, 11 (female=10, male=1, 2 left-handed) professional EN-DE translators participated in the experiment, 3 freelance and 8 in-house translators. Their ages ranged from 30 to 64 (mean=41.6, σ=9.3), with 3 to 30 years of professional experience (mean=13.3, σ=7.4) and a total of 27 language pairs (mean=2.6). All translators translate from EN to DE, and all describe their German Language skills as native and their English skills as C1 to native level (C1=3 times, C2=6, native=2). For most participants, the self-rated CAT knowledge was good (6 times) or very good (4 times, 1 neutral for the oldest participant). However, participants were less confident about their PE skills (4 neutral, 4 good, 3 very good), thereby matching well with the CAT usage surveys. Years of experience with CAT tools ranged from 3 to 20 (mean=11.5, σ=5.1), where participants had used between 1 and 10 distinct CAT tools (mean=4.9, σ=2.7), most frequently using SDL Trados Studio (10), Across (8), MemoQ (8), Transit (6), and XTM (6). Overall, participants are quite satisfied with their current CAT tools (mean=5.3, σ=1.0, on a 7-point scale). The most used features are TM (10 participants), MT (8), and Terminology Management (8). For the following analyses, one should note that the small number of participants and their age distribution (with 10 participants of age 30 to 48, and only one of age 64) did not allow to analyze the effect of age on the results.

4.3.2 Hardware Setup

Four participants commented very positively about the large movable screen ("the screen is great"/"really love the screen"). While one participant decided to work in a standing position throughout the whole study and two participants worked with a tilted screen, we were surprised to see that most participants actually placed it in a normal upright position, probably because this setup was most familiar to them. A reason could be that they explored all features, including mouse and keyboard, and were therefore simply used to this setup. However, one participant who explored handwriting a lot without tilting the screen already asked whether this would be problematic for the shoulder in the long run.

4.3.3 Layout & Interface Design

Regarding interface design, participants highlighted that they really liked the large font size chosen for the current segment, because "you really see what you are working on" which sometimes is not that easy in other CAT tools. The chosen colors were also positively commented two times as they help finding the position one is currently editing, however, one participant did not like the dark mode for the surrounding segments, and would prefer pastel colors for this. While more participants told us that they prefer a horizontal src-tgt layout, one participant said that a switch to change between the two would be nice. Furthermore, some adaption possibilities, like changing the sizes of the windows were requested, enlarging the confirm button and repositioning it to the right or below the target, and that the current segment should be always in the same position. In general, we received mostly positive feedback, claiming it would be "user friedly", "I like the interface", that the layout is nice, etc. One participant suggested that we should further focus on small screens, like tablets, where he sees potential for occasional PE.

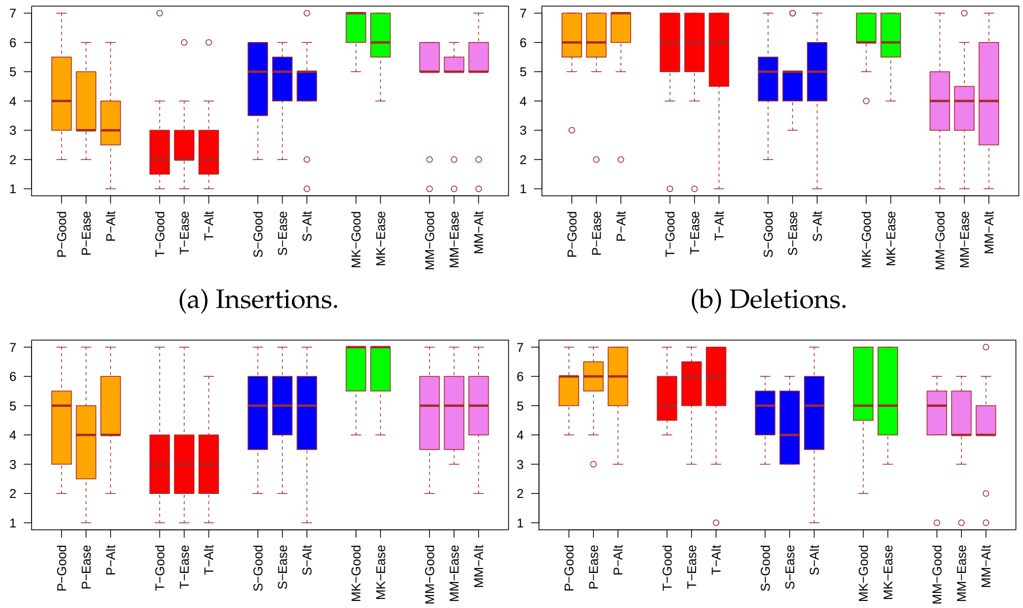
We noticed that some participants were confused which operations could be done on the right-hand and which on the left-hand view, although they apparently became used to it over time and also confirmed that. E.g., two participants tried placing the cursor on the drawing view, which instead painted a dot at that position. A participant also assumed that the cursor on the right-hand side is automatically placed after the last handwritten input on the left, s.t. one could immediately continue typing there, which would be a nice addition to the current prototype. Approaches to solve this confusion were named: Simply labeling them might already help, using different sides of the pen for different actions, or simply having a switch to enable/disable the drawing mode. The consolidation of both views would also make them bigger, which would be useful according to a participant.

4.3.4 Subjective Ratings

Figure 4.6 shows the subjective ratings provided for each modality and operation on the three scales "Goodness", "Ease of use", and "Good alternative to mouse & keyboard" after having tested each feature (see Section 4.2.2). As can be seen, participants tended to give similar ratings on all three scales.

For **insertions** and **replacements**, which required the most text input, the classical mouse & keyboard approach was rated highest; however, the multi-modal combination and speech were also perceived as good, while, pen and especially touch received lower scores.

For **deletions** and **reorderings**, pen, touch, and mouse & keyboard were all perceived as very good, where the P and T were even slightly higher ranked than MK for reorderings. Speech and multi-modal input are considered worse here.



(a) Insertions.

(b) Deletions.

(c) Replacements.

(d) Reorderings.

Figure 4.6: Subjective ratings of the five modalities (P=Pen in orange, T=Touch in red, S=Speech in blue, MK=Mouse & Keyboard in green, MM=Multi-Modal in purple) for the four operations (insertions, deletions, replacements, reorderings) on the 7-point Likert scales for goodness, ease of use, and whether it is a good alternative to MK.

4.3.5 Orderings

After each operation, participants ordered the modalities from best to worst, with ties being allowed. As an example, for "MM & S best, then P, then MK, and last T" we assigned 0.5 times the 1st and 0.5 times the 2nd position to both MM and S, while P got 3rd, MK 4th, and T the 5th position. To get an overall ordering across participants, we then multiplied the total amount of times a modality was rated 1st/2nd/3rd/4th/5th by 1/2/3/4/5 (similar to Zenner and Krüger (2017)). Thus, lower scores indicate that a modality is better suited for an operation.

The scores for each modality and operation are:

- **Insertions:** 1st: **MK**(20.5), 2nd: **MM**(26.5), 3rd: **S**(31.5), 4th: **P**(38.5), 5th: **T**(48)
- **Deletions:** 1st: **P**(21.5), 2nd: **MK**(29), 3rd: **T**(31.5), 4th: **MM**(41), 5th: **S**(42)
- **Replacements:** 1st: **MK**(21), 2nd: **MM**(29), 3rd: **S**(30), 4th: **P**(35), 5th: **T**(50)
- **Reorderings:** 1st: **P**(21.5), 2nd: **T**(31), 3rd: **S**(35.5), 4th: **MK**(36), 5th: **MM**(41)

4.3.6 Timings

We analyzed the logged duration of each modality-operation-pair. Note that this is the time from clicking “Start” until confirming the segment; thus, it includes recognition times (for speech and handwriting) and really measures how long it takes until a participant is satisfied with the edit. Even though participants were instructed to provide feedback or ask questions only while the popup is shown, i.e., while the time is not measured, participants infrequently did so during editing. We filtered out such outliers and averaged the 4 sentences of each modality-operation pair per participant to get a single value, thereby making the samples independent for the remaining analyses.

Figure 4.7 shows boxplots of the dataset for the 20 modality-operation pairs. For statistical analysis, we first conducted Friedman tests per operation, showing us that significant differences exist for each operation (all $p < 0.001$). Afterward, post-hoc analyses using Wilcoxon tests with Bonferroni-Holm correction showed which pairs of modalities are significant and how large the effect r is.

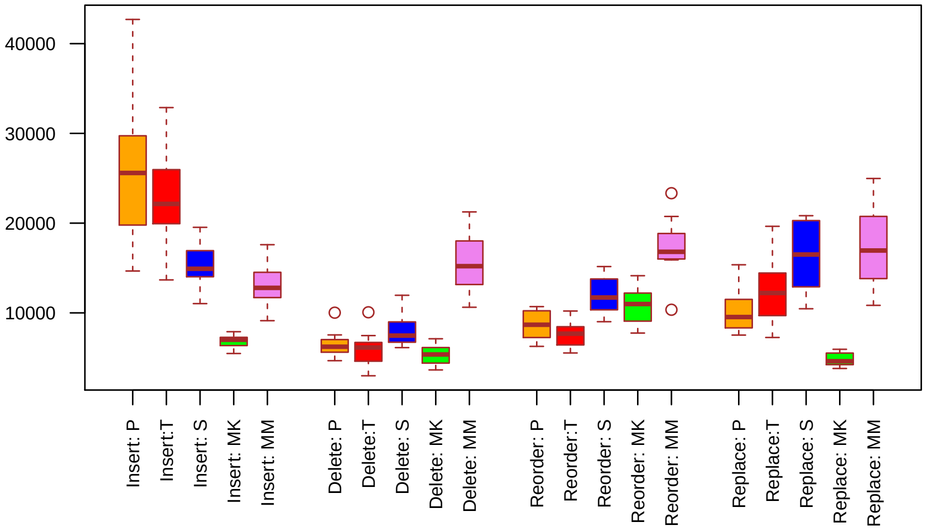


Figure 4.7: Editing durations per operation (insertion, deletion, reordering, replacement) and modality (P=Pen in orange, T=Touch in red, S=Speech in blue, MK=Mouse & Keyboard in green, MM=Multi-Modal in purple).

For **insertions**, MK was by far the fastest modality, followed by MM and S. All differences except for MM vs. S and T vs. P are statistically significant with large effect sizes (all $p < 0.01$, all $r > 0.83$).

As expected, **deletions** were faster than insertions. Here, MK, T, and P were the fastest, followed by S, and by far last MM. Regarding significance, all modalities were significantly faster than MM, and MK was significantly faster than S (all $p < 0.01$, all $r > 0.88$).

For **reordering**, P and T were the fastest, followed by MK and S. The statistical analysis revealed that T is significantly faster than all modalities except P, both P and MK are significantly faster than S, and S is significantly faster than MM (all $p < 0.05$, all $r > 0.83$).

Replacements with MK were fastest, followed by P, T, S, and MM. MK was significantly faster than all other modalities, P significantly faster than S and MM (all p < 0.05, all $r > 0.83$); no significant differences exist between the others.

4.3.7 Qualitative Analysis

Apart from the ratings and timings, we present the main qualitative feedback from the interviews.

Pen & Touch

Insertions & Replacements One participant described handwriting as “fun but not effective”, while others were more positive, finding it “super”, “quick”, and ”very delicate”. A participant stated that the “pen symbolizes the writer, and every translator is a writer, which nicely resembles this”. Especially for short insertions and replacements, handwriting was seen as a suitable input mode; for more traditional translation from scratch or more extended changes, one should instead fall back on typing or dictation. Furthermore, the additional movement compared to normal typing was emphasized.

The pen was considered to have a nice shape and feel pleasant (2 participants). Most participants seemed to prefer the pen to finger handwriting due to its precision, however, two participants preferred finger handwriting as it would be more direct and natural. One participant said that she would even finger-write on paper if ink was coming out of it. In general however, the thickness of the finger seemed to impede the recognition for most.

Space management was the most criticized aspect about handwriting, the main issues being that one has to think about how to ensure having sufficient space, and that the gesture to create space was not always well recognized. To improve this, participants suggested to make the space (font size, line height, etc.) configurable to one’s own handwriting, or to create space automatically. Another issue mentioned by multiple participants was that placing the palm of the hand on the screen resulted in switching the segment, which made handwriting more difficult and should be automatically ignored. Furthermore, strokes that started in the line of the recognized text at the very top of the drawing view (see Figure 4.3a), were ignored, which irritated several participants. To solve this, one could simply add additional space between this line and the first actual line. A participant by accident pressed the button on the pen during writing, resulting in handwriting to be triggered already during hover with the pen, i.e., before an actual touch occurs, and an accidental double-click on the button resulted in the context menu appearing. Thus we should disable this button in the future. Last, ideas to further improve the handwriting were to train the handwriting for the participant, and to automatically convert the handwriting to computer font.

Deletion Both touch/pen deletion mechanisms (strike-through and double-tap) were highlighted as very useful or even "perfect"; as they "nicely resemble a standard correction task". Some favored the double-tap approach, others the strike-through approach. Furthermore, the possibility to delete longer subphrases or even multiple lines through a quick diagonal strike-through was further emphasized. There was no clear opinion whether touch was better or worse then pen, some finding it more natural, others less precise.

Reordering Participants also provided very positive feedback regarding touch/pen reordering, 9 participants stating that it is "super"/"works quite well", that they "loved it", or is "really cool" mostly because it is fast, there is no need to wait and see if the recognition worked, and that it is similar to a standard correction task. However, moving multiple words simultaneously should be supported for touch/pen, but also for multi-modal reordering. While the automatic space management was considered "amazing", a participant argued that the capitalization should also be automatically fixed when moving a word from/to the beginning of a sentence. Last, a participant noticed that the text jumps around a bit while reordering, as the picked up word is removed from the text resulting in a shift of word. Instead one could prevent this by visualizing a large whitespace where the word came from and only adapting the text on drop. Furthermore, the visualization of the end position should be improved. Participants considered touch and pen reordering to work comparably well. 4 participants claimed that touch was better than pen for this task, however, others pointed out that for very small words which frequently occur in Portuguese the finger might not be precise enough, and that the small tip of the pen occludes the text less.

Speech & Multi-modal Combinations

Feelings regarding speech recognition were rather mixed, some thinking it works "super", being "positively surprised by the speech commands", while others thought it is a bit slow or imprecise and that they could not really imagine using it often. One advantage as stated by a participant would be that it helps reduce tenosynovitis, which many translators are dealing with due to frequent typing; furthermore, it would be especially useful for handicapped translators. Two participants argued against speech as it would be exhausting to formulate a command while one is already processing the source and target text. Furthermore, speech was considered impractical for translators working in shared offices. Interestingly, a participant spoke very slowly and well-articulated in the beginning until s/he noticed after a while that talking normally actually works much better.

There was no clear opinion whether the multi-modal approach or speech was preferred. The main advantage of the multi-modal approach was that “one has to speak less”, “could be really fast”, involving “less thinking”. However, it was also argued that you “can't work hands off”, or that “when you talk you can also just say everything”, meaning that the simplified MM command was not seen as an advantage for this participant. An interesting statement was: “if there are no ambiguities, speech is better, but if there are, multi-modal is cool”. Especially for the simple task of deletion “doing two things” was considered “too complex”.

Ideas on how to improve speech ranged from better highlighting the changes in the target view, to adding the possibility to restate the whole segment. While the ASR tool used (IBM Watson) is one of the state-of-the-art APIs, it might still have negatively impacted the results for S and MM, as several times a word was wrongly recognized (e.g., when replacing a word due to an incorrect ending, the ASR did not always correctly recognize the word form). To improve this aspect, participants discussed the idea of passing the text to the speech recognition (similar to the TransTalk project, see Dymetman et al. (1994)) or to train the ASR towards the user. Furthermore, participants proposed to add automatic capitalization depending on punctuation, or applying character-level changes (e.g., remove last character of word), and to offer undo/redo/confirm commands through speech input. Furthermore, a participant said that it would be useful if the ASR could understand both source and target language instead of just one, especially if a word was not translated (undertranslation). Currently, our implementation only supported normal words, so multiple participants requested to be able to manipulate punctuation marks as well, or to use them for specifying position (“after second comma”). We further figured out a few synonyms we should add for our commands, as well as supporting a changed order within our commands (e.g., stating the position before the entity to insert). Sometimes participants started a command, then made a short break, and continued their command afterwards, which made our prototype assume it were two commands, which were both incomplete, and thus resulted in no change at all, which should be fixed in the future.

Insertions & Replacements Both insertions and replacements using speech received lots of positive feedback (from 8 and 7 participants, respectively), interesting findings being that “the longer the insertion, the more interesting speech becomes” and that speech would be more suitable for replace and delete operations as there is usually no need to specify the position. However, a participant also considered it too complex for small replacement changes, saying “while I talk I could just fix it using mouse and keyboard”.

Deletions Speech also received some positive feedback for deletions, stating that it “works fine” and recognizes well, or would be simpler for deletions than for insertions as one has to talk less. The automatic correction of spaces was also highlighted as a nice feature. However, others commented negatively that they cannot really imagine using speech for deletions, that if it misrecognizes you this costs a lot of time, and that it would be bad to have to read 10 words if one wants to delete 10 words. Again, further commands should be supported, but also that deleting the whole sentence should be possible.

Feedback on multi-modal deletions was rather negative, the main disadvantage being that one has to think about and do two things, which is too complex for this simple task. "If all modalities work well by themselves, then why combine them?". Another disliked aspect was that multi-word delete was currently not supported in the multi-modal approach.

Reordering Speech reordering was considered more complicated than using the pen or touch, however, the automatic correction of spaces was again highlighted as a nice feature. Remarks on multi-modal ordering were again mixed, since one has to say less but do two things.

Mouse & Keyboard

Due to their daily usage, participants stated they were strongly biased regarding mouse and keyboard, where "the muscle memory" helps. However, many actually considered MK as very unintuitive if they imagined never having used it before, especially compared to pen and touch, or as one participant stated for reordering: "Oh I did not like that! Why do I have to do all of this, why is it not as simple as the pen". Especially the back and forth between mouse and keyboard when not using hotkeys was considered problematic, as well as the manual space correction after deletion. Surprisingly, one participant even used the mouse to strike-through text in the drawing view. MK was considered to be really precise, and if one uses all the hotkeys also very fast due to the possibility to delete to the right and left, whole words at a time or individual characters etc. Having to manually correct spaces was however highlighted negatively, even though they are used to that. Also placing the mouse first is considered an annoying additional step. One participant put it like "depends on what you do, for small changes precision is good, for whole words it is not so good".

4.3.8 General Feedback & Feature Requests

In general, we received a lot of positive feedback in the final discussion about the prototype, where participants stated that "I am going to buy this once you are ready", it would be "interesting", "fascinating", "respect for the prototype" or that they "could imagine using future versions of it". Multiple participants reported that it would be nice to have multiple options to vary between the modalities, so that one does not do the same thing all the time, which might even improve concentration. It would also be great to "work with the hands" from time to time, which would add an additional "fun factor". Participants did not agree on which modality would in general be most suitable, some mostly preferred speech, others pen, others argued it would be a question of getting used to it. A nice overall conclusion by a participant was: "Anything I add I would use speech, anything I delete I would use pen".

Apart from standard features like translation memory, terminology management, quality assurance, and concordance search, a participant requested the possibility to visualize whitespaces. Furthermore, special language-dependent characters (e.g., apostrophes) should be automatically corrected depending on the target language (3 participants). Three participants found the integration of Linguee as a dictionary and corpus "very cool", and it has been argued twice that there should also be the possibility to lookup multiple source words by marking them on the source view. They also highlighted that further functionality supporting them in researching terms would help. As Balashov (2020) nicely describes, professional translators need to be highly specialized to a domain, but cannot know every detail of it, e.g., they can be experts on medical equipment, but still cannot be expected to know "the distinction between different types of intra-aortic catheters", for which good researching support would be beneficial. While the spell check functionality was already considered a good start, it should be further extended detect double words, incorrect punctuation, or grammar.

4.4 Discussion

This section discusses the main takeaways per modality.

4.4.1 Pen

According to ordering scores, subjective ratings, and comments, we see that pen is among the best modalities for deletions and reordering. However, other modalities are superior for insertions and replacements, where it was only seen as suitable for short modifications, but to be avoided for more extended changes. In terms of timings, P was also among the fastest for deletions and reorderings, and among the slowest for insertions. What is interesting, however, is that P was significantly faster than S and MM for replacements (by 6 and 7 seconds on average) even though it was rated lower. Participants also commented very enthusiastically about pen reordering and deletions, as they would nicely resemble manual copy-editing. The main concern for hand-writing was the need to think about and to create space before actually writing.

4.4.2 Touch

According to subjective ratings, orderings and the participants' comments after the experiment, results for touch were similarly good for deletions and reorderings, but it was considered worse for insertions and replacements. Furthermore, and as we expected due to its precision, pen was preferred to finger touch by most participants. However, in terms of timings, the two did not differ significantly, apart from replace operations (where pen was faster). Even for such replacements, where touch was rated as the worst modality, it actually was (non-significantly) faster than S and MM.

4.4.3 Speech & Multi-Modal Combinations

Speech and Multi-Modal PE were considered the worst and were also the slowest modalities for reordering and deletions. For insertions and replacements, however, these two modalities were rated and ordered 2$^{\text{nd}}$ (after MK) and in particular much better than P and T. Timing analysis agrees for insertions, being 2$^{\text{nd}}$ after MK; for replacements, however, S and MM were the slowest even though the ratings put them ahead of P and T. MM was slower than S for deletion, which can be explained by the fact that our implementation did not support MM deletions of multiple words in a single command which was the case in 2 out of 4 tested sentences. Still, we would have expected a comparable speed of MM and S for reordering tasks, as is the case for replacements. Insertions are the only operation where MM was (non-significantly) faster than S, since the position did not have to be verbally specified. Furthermore, the participants' comments highlighted their concern regarding formulating commands while already mentally processing text. Still, S and MM were considered especially interesting the more text should be added. The main advantage of the multi-modal approach as argued by the participants of our study was that one has to speak less, albeit at the cost of doing two things at once.

4.4.4 Mouse & Keyboard

Mouse & keyboard received the best scores for insertions and replacements, where it was the fastest modality. Furthermore, it got good ratings for deletions and reorderings, where it was also fast but not the fastest for reordering. However, some participants commented negatively, being "unintuitive" especially for the reordering task, and stating that it only works well because of "years of expertise". As can be seen in the timing analysis, MK is still by far the quickest when text needs to be produced (insertion and replacement), and comparably fast to P, T, and S for deletions. For reordering, however, it was slower than P and T. One should note that insertion tasks in this PE study require the insertion of a single or a few words within a sentence; for dictating whole sentences or long phrases, speech might well be faster than typing.

4.4.5 General

Overall, many participants provided very positive feedback on this first prototype combining pen, touch, speech, and multi-modal combinations for PE MT, encouraging us to continue. They especially highlighted that it was nice to have the option to switch between modalities. Furthermore, several promising ideas for improving the prototype were proposed, e.g., to visualize whitespaces.

The focus of our study was to explore the implemented interactions in detail, i.e., each modality for each operation irrespective of frequency. The chosen methodology guaranteed that we receive comparable feedback on all interactions from professional translators by having them correct the same mistakes using different modalities. Nevertheless, a more realistic "natural" workflow follow-up study should be conducted in the future, which will also show if participants swap modalities within sentences depending on the error type, or if they stick to single modalities to avoid frequent modality switches.

Interestingly, our findings are not entirely in line with translators' intuitions reported in our previous elicitation study: While touch worked much better than expected, handwriting of whole subphrases worked worse than they thought. Additionally, it is interesting to note that some newly introduced modalities could compete with mouse & keyboard even though participants are biased by years of training with the latter.

4.5 Conclusion

Based on an initial elicitation study, this chapter proposed and evaluated MMPE, a CAT prototype combining pen, touch, speech, and multi-modal interaction together with common mouse and keyboard input possibilities for PE. Users can directly cross out or hand-write new text, drag and drop words for reordering, or use spoken commands to update the text in place. Our study with professional translators shows a high level of interest and enthusiasm for using these new modalities. For deletions and reorderings, pen and touch both received high subjective ratings, with pen being even better than mouse & keyboard. In terms of timings, they were also among the fastest for these two operations. For insertions and replacements, speech and multi-modal interaction were considered suitable; however, mouse & keyboard were still favored and faster here.

For the goal of this part of the thesis investigating modalities other than mouse and keyboard in the PE process, chapter 4 plays a crucial role by developing a working interface based on the elicitation studies' findings (chapter 3) and evaluating it with professional translators. MMPE combines the most modalities simulatenously explored in the CAT and PE literature, and the well-structured study allows us to gain an understanding of the usefulness of different modalities for different PE tasks. Furthermore, various suggestions for improvement have been outlined, which will be addressed in the next chapter.

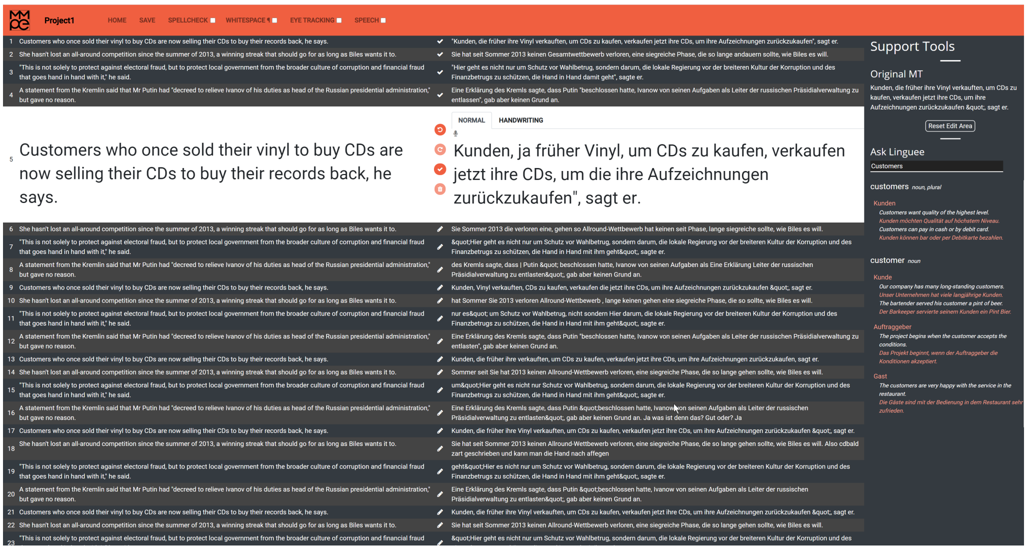
Chapter 5
Improvements and Further Studies with MMPE

In this chapter, we leverage the qualitative feedback from the previous study to extend and improve the MMPE CAT environment through a variety of layout changes, enhanced interaction flexibility, and others, as discussed below. We also present first steps on including eye tracking combined with speech or keyboard input as an additional interaction modality. Finally, we summarize two additional studies conducted with MMPE, namely our investigations of mid-air hand gestures in combination with the keyboard for PE of MT, and the exploration of word-level QE within the PE process.

This chapter is based on publications Herbig et al. (2020d), Jamara et al. (2021), and Shenoy et al. (2021).

5.1 Improvements to Overall Layout

We decided to stick with the enlarged visualization of the current segment, as participants in our study liked the large font size. However, the original prototype had the two target views (handwriting and default editing) next to each other. Thus, overall, our system included three neighboring views. This was perceived as unintuitive, leading to much confusion, especially at the beginning of the experiment, when participants did not remember which target view supported which features. Therefore, we combined the two target views into one with tabs, clearly labeled which mode does what, and allow quickly switching between them. The new combination has the additional advantages that the interface becomes symmetric (as there is only one source and target for the previous, the current, and the remaining segments). Furthermore, the space for hand-writing increases even further, and the layout on smaller displays with insufficient space to nicely visualize three text boxes next to each other is improved. Undo and redo functionality and segment confirmation are also implemented: as before (1) by using hotkeys, or (2) through buttons between source and target, but now also (3) through speech commands. Next to the buttons for undo and redo we added a button for touch deletion, which we will discuss below.



1	Customers who once sold their vinyl to buy CDs are now selling their CDs to buy their records back, he says.	✓	"Kunden, die früher ihre Vinyl verkauften, um CDs zu kaufen, verkaufen jetzt ihre CDs, um ihre Aufzeichnungen zurückzukaufen", sagt er.
2	She hasn't lost an all-around competition since the summer of 2013, a winning streak that should go for as long as Biles wants it to.	✓	Sie hat seit Sommer 2013 keinen Gesamtwettbewerb verloren, eine siegreiche Phase, die so lange andauern sollte, wie Biles es will.
3	"This is not solely to protect against electoral fraud, but to protect local government from the broader culture of corruption and financial fraud that goes hand in hand with it," he said.	✓	"Hier geht es nicht nur um Schutz vor Wahlbetrug, sondern darum, die lokale Regierung vor der breiteren Kultur der Korruption und des Finanzbetrugs zu schützen, die Hand in Hand damit geht", sagte er.
4	A statement from the Kremlin said that Mr Putin had "decreed to relieve Ivanov of his duties as head of the Russian presidential administration," but gave no reason.	✓	Eine Erklärung des Kremls sagte, dass Putin "beschlossen hatte, Ivanow von seinen Aufgaben als Leiter der russischen Präsidialverwaltung zu entlassen", gab aber keinen Grund an.

5 Customers who once sold their vinyl to buy CDs are now selling their CDs to buy their records back, he says.

Kunden, ja früher Vinyl, um CDs zu kaufen, verkaufen jetzt ihre CDs, um die ihre Aufzeichnungen zurückzukaufen", sagt er.

6	She hasn't lost an all-around competition since the summer of 2013, a winning streak that should go for as long as Biles wants it to.	✎	Sie Sommer 2013 die verloren eine, gehen so Allround-Wettbewerb hat keinen seit Phase, lange siegreiche sollte, wie Biles es will.
7	"This is not solely to protect against electoral fraud, but to protect local government from the broader culture of corruption and financial fraud that goes hand in hand with it," he said.	✎	&quot;Hier geht es nicht nur um Schutz vor Wahlbetrug, sondern darum, die lokale Regierung vor der breiteren Kultur der Korruption und des Finanzbetrugs zu schützen, die Hand in Hand mit ihm geht&quot;, sagte er.
8	A statement from the Kremlin said that Mr Putin had "decreed to relieve Ivanov of his duties as head of the Russian presidential administration," but gave no reason.	✎	des Kremls sagte, dass | Putin &quot; beschlossen hatte, Ivanow von seinen Aufgaben als Eine Erklärung Leiter der russischen Präsidialverwaltung zu entlasten&quot;, gab aber keinen Grund an.
9	Customers who once sold their vinyl to buy CDs are now selling their CDs to buy their records back, he says.	✎	Kunden, Vinyl verkauften, CDs zu kaufen, verkaufen die jetzt ihre CDs, um ihre Aufzeichnungen zurückzukaufen &quot;, sagt er.
10	She hasn't lost an all-around competition since the summer of 2013, a winning streak that should go for as long as Biles wants it to.	✎	hat Sommer Sie 2013 verloren Allround-Wettbewerb , lange keinen gehen eine siegreiche Phase, die so sollte, wie Biles es will.
11	"This is not solely to protect against electoral fraud, but to protect local government from the broader culture of corruption and financial fraud that goes hand in hand with it," he said.	✎	nur es&quot; um Schutz vor Wahlbetrug, nicht sondern Hier darum, die lokale Regierung vor der breiteren Kultur der Korruption und des Finanzbetrugs zu schützen, die Hand in Hand mit ihm geht&quot;, sagte er.
12	A statement from the Kremlin said that Mr Putin had "decreed to relieve Ivanov of his duties as head of the Russian presidential administration," but gave no reason.	✎	Eine Erklärung des Kremls sagte, dass Putin "beschlossen hatte, Ivanow von seinen Aufgaben als Leiter der russischen Präsidialverwaltung zu entlasten", gab aber keinen Grund an.
13	Customers who once sold their vinyl to buy CDs are now selling their CDs to buy their records back, he says.	✎	Kunden, die früher ihre verkauften, um CDs zu kaufen, verkaufen jetzt ihre CDs, um ihre Aufzeichnungen zurückzukaufen, sagt er.
14	She hasn't lost an all-around competition since the summer of 2013, a winning streak that should go for as long as Biles wants it to.	✎	Sommer seit Sie hat 2013 keinen Allround-Wettbewerb verloren, eine siegreiche Phase, die so lange gehen sollte, wie Biles es will.
15	"This is not solely to protect against electoral fraud, but to protect local government from the broader culture of corruption and financial fraud that goes hand in hand with it," he said.	✎	um&quot;Hier geht es nicht nur Schutz vor Wahlbetrug, sondern darum, die lokale Regierung vor der breiteren Kultur der Korruption und des Finanzbetrugs zu schützen, die Hand in Hand mit ihm geht&quot;, sagte er.
16	A statement from the Kremlin said that Mr Putin had "decreed to relieve Ivanov of his duties as head of the Russian presidential administration," but gave no reason.	✎	Eine Erklärung des Kremls sagte, dass Putin &quot;beschlossen hatte, Ivanow von seinen Aufgaben als Leiter der russischen Präsidialverwaltung zu entlasten&quot;, gab aber keinen Grund an. Ja was ist denn das? Gut oder? Ja
17	Customers who once sold their vinyl to buy CDs are now selling their CDs to buy their records back, he says.	✎	Kunden, die früher ihre Vinyl verkauften, um CDs zu kaufen, verkaufen jetzt ihre CDs, um ihre Aufzeichnungen zurückzukaufen &quot;, sagt er.
18	She hasn't lost an all-around competition since the summer of 2013, a winning streak that should go for as long as Biles wants it to.	✎	Sie hat seit Sommer 2013 keinen Allround-Wettbewerb verloren, eine siegreiche Phase, die so lange gehen sollte, wie Biles es will. Also cdbald zart geschrieben und kann man die Hand nach affegen
19	"This is not solely to protect against electoral fraud, but to protect local government from the broader culture of corruption and financial fraud that goes hand in hand with it," he said.	✎	geht&quot;Hier es nicht nur um Schutz vor Wahlbetrug, sondern darum, die lokale Regierung vor der breiteren Kultur der Korruption und des Finanzbetrugs zu schützen, die Hand in Hand mit ihm geht&quot;, sagte er.
20	A statement from the Kremlin said that Mr Putin had "decreed to relieve Ivanov of his duties as head of the Russian presidential administration," but gave no reason.	✎	Eine Erklärung des Kremls sagte, dass Putin &quot;beschlossen hatte, Ivanow von seinen Aufgaben als Leiter der russischen Präsidialverwaltung zu entlasten&quot;, gab aber keinen Grund an.
21	Customers who once sold their vinyl to buy CDs are now selling their CDs to buy their records back, he says.	✎	Kunden, die früher ihre Vinyl verkauften, um CDs zu kaufen, verkaufen jetzt ihre CDs, um ihre Aufzeichnungen zurückzukaufen &quot;, sagt er.
22	She hasn't lost an all-around competition since the summer of 2013, a winning streak that should go for as long as Biles wants it to.	✎	Sie hat seit Sommer 2013 keinen Allround-Wettbewerb verloren, eine siegreiche Phase, die so lange gehen sollte, wie Biles es will.
23	"This is not solely to protect against electoral fraud, but to protect local government from the broader culture of corruption and financial fraud	✎	&quot;Hier geht es nicht nur um Schutz vor Wahlbetrug, sondern darum, die lokale Regierung vor der breiteren Kultur der Korruption und des

Support Tools

Original MT

Kunden, die früher ihre Vinyl verkauften, um CDs zu kaufen, verkaufen jetzt ihre CDs, um ihre Aufzeichnungen zurückzukaufen &quot;, sagt er.

Ask Linguee

Customers

customers *noun, plural*

Kunden
Customers want quality of the highest level.
Kunden möchten Qualität auf höchstem Niveau.
Customers can pay in cash or by debit card.
Kunden können bar oder per Debitkarte bezahlen.

customer *noun*

Kunde
Our company has many long-standing customers.
Unser Unternehmen hat viele langjährige Kunden.
The bartender served his customer a pint of beer.
Der Barkeeper servierte seinem Kunden ein Pint Bier.

Auftraggeber
The project begins when the customer accepts the conditions.
Das Projekt beginnt, wenn der Auftraggeber die Konditionen akzeptiert.

Gast
The customers are very happy with the service in the restaurant.
Die Gäste sind mit der Bedienung in dem Restaurant sehr zufrieden.

Figure 5.1: Screenshot of the improved MMPE interface.

A feature that we newly introduced because it was requested in the study was the visualization of whitespaces, which can also be enabled in the navigation bar. Figure 5.2 shows the visualized spaces and line breaks, commonly known from Microsoft Word.

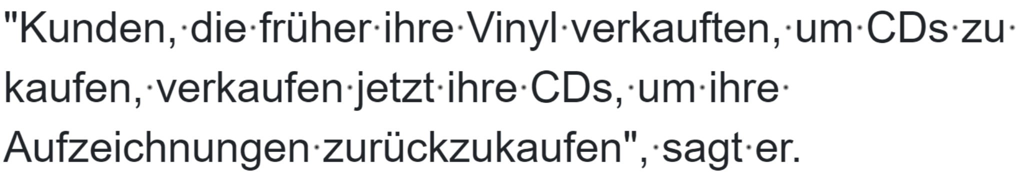

"Kunden,·die·früher·ihre·Vinyl·verkauften,·um·CDs·zu·kaufen,·verkaufen·jetzt·ihre·CDs,·um·ihre·Aufzeichnungen·zurückzukaufen",·sagt·er.

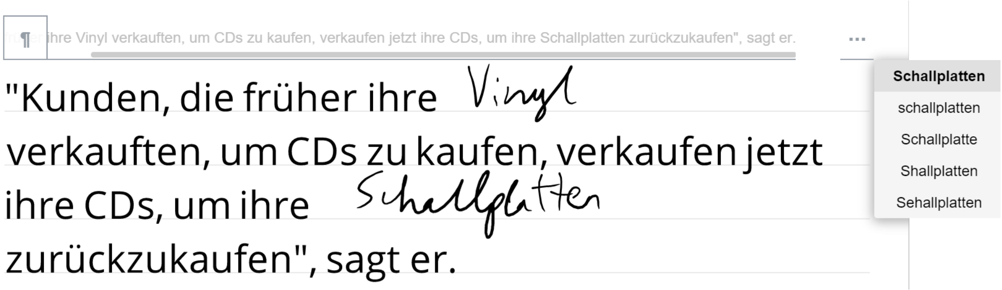
Figure 5.2: Visualization of whitespaces in MMPE.

5.2 Improvements to Handwriting

The handwriting recognition using the MyScript Interactive Ink SDK worked well in the study. Most participants preferred the pen to finger handwriting for insertions and replacements due to its precision, although some considered it less direct than finger input. One participant even used the strike-through deletion with the mouse; therefore, we decided to keep all three handwriting options.

However, handwriting for replacements and insertions was considered to work well only for short modifications. An issue that might have influenced this finding was that in the evaluated version of MMPE, an unintended change of the currently selected segment happened when the palm of the hand touched another piece of text. As it is common to lay down one's hand while writing, we now prevent this unintended segment change by ignoring palm touches.

Furthermore, participants found the gesture to create space (drawing a vertical line) often hard to accomplish, which we improved by increasing the lineheight. For the first line, we added even more space to the top, to prevent a drawing from starting in the text box containing the recognized text, where it would be ignored, thereby improving the user experience. Last, we deactivated the button of the digital pen, as it frequently resulted in unintended right-clicks triggering the context menu. Figure 5.3 shows the handwriting view with the additional space through the interface layout change and the lineheight, as well as some alternatives for recognized text.



früher ihre Vinyl verkaufen, um CDs zu kaufen, verkaufen jetzt ihre CDs, um ihre Schallplatten zurückzukaufen", sagt er.

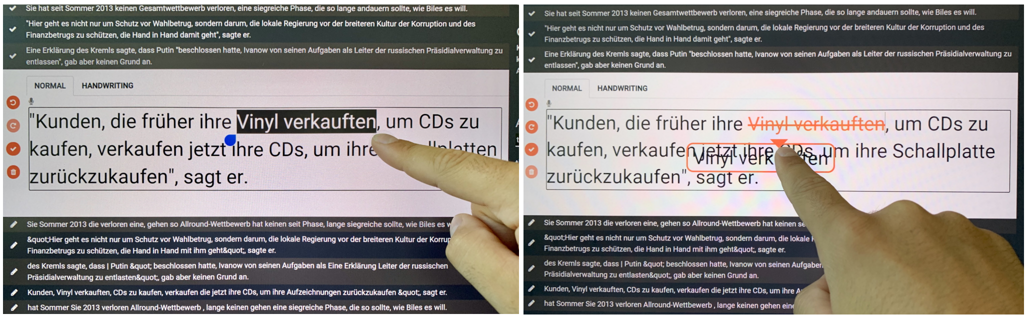
"Kunden, die früher ihre Vinyl
verkauften, um CDs zu kaufen, verkaufen jetzt
ihre CDs, um ihre Schallplatten
zurückzukaufen", sagt er.

Figure 5.3: Handwriting and alternatives after clicking the word "Schallplatten" in the recognized text at the top.

5.3 Improvements to Touch Reordering and Deletion

Touch reordering was highlighted as particularly useful and received the highest subjective scores and lowest time required for reordering. Nevertheless, the old reorder only supported moving one word at a time, not whole sub-phrases or parts of words, which is naturally needed in actual PE settings. Now, users have two options: (1) they can drag and drop single words by starting a drag directly on top of a word, or (2) they can double-tap to start a selection process, define which part of the sentence should be selected (e.g., multiple words or a part of a word, see Figure 5.4a), and then move it (see Figure 5.4b).

While this allows a much more flexible reorder functionality, it has the disadvantage that double-tap can no longer be used to delete words, as was the case in the previous prototype version. However, as strike-through in the handwriting view



✓ Sie hat seit Sommer 2013 keinen Gesamtwettbewerb verloren, eine siegreiche Phase, die so lange andauern sollte, wie Biles es will.

✓ "Hier geht es nicht nur um Schutz vor Wahlbetrug, sondern darum, die lokale Regierung vor der breiteren Kultur der Korruption und des Finanzbetrugs zu schützen, die Hand in Hand damit geht", sagte er.

✓ Eine Erklärung des Kremls sagte, dass Putin "beschlossen hatte, Ivanow von seinen Aufgaben als Leiter der russischen Präsidialverwaltung zu entlassen", gab aber keinen Grund an.

"Kunden, die früher ihre Vinyl verkauften, um CDs zu kaufen, verkaufen jetzt ihre CDs, um ihre llplatten zurückzukaufen", sagt er.

Sie Sommer 2013 die verloren eine, gehen so Allround-Wettbewerb hat keinen seit Phase, lange siegreiche sollte, wie Biles es w

&quot;Hier geht es nicht nur um Schutz vor Wahlbetrug, sondern darum, die lokale Regierung vor der breiteren Kultur der Korrup Finanzbetrugs zu schützen, die Hand in Hand mit ihm geht&quot;, sagte er.

des Kremls sagte, dass | Putin &quot; beschlossen hatte, Ivanow von seinen Aufgaben als Eine Erklärung Leiter der russischen Präsidialverwaltung zu entlasten&quot;, gab aber keinen Grund an.

Kunden, Vinyl verkauften, CDs zu kaufen, verkaufen die jetzt ihre CDs, um ihre Aufzeichnungen zurückzukaufen &quot;, sagt er.

hat Sommer Sie 2013 verloren Allround-Wettbewerb , lange keinen gehen eine siegreiche Phase, die so sollte, wie Biles es will.

(a) Multi-word selection using touch.

✓ Sie hat seit Sommer 2013 keinen Gesamtwettbewerb verloren, eine siegreiche Phase, die so lange andauern sollte, wie Biles es will.

✓ "Hier geht es nicht nur um Schutz vor Wahlbetrug, sondern darum, die lokale Regierung vor der breiteren Kultur der Korruption und des Finanzbetrugs zu schützen, die Hand in Hand damit geht", sagte er.

✓ Eine Erklärung des Kremls sagte, dass Putin "beschlossen hatte, Ivanow von seinen Aufgaben als Leiter der russischen Präsidialverwaltung zu entlassen", gab aber keinen Grund an.

"Kunden, die früher ihre ~~Vinyl verkauften~~, um CDs zu kaufen, verkaufen jetzt ihre CDs, um ihre Schallplatte zurückzukaufen", sagt er.

Vinyl verk ften

Sie Sommer 2013 die verloren eine, gehen so Allround-Wettbewerb hat keinen se

&quot;Hier geht es nicht nur um Schutz vor Wahlbetrug, sondern darum, die lokale Finanzbetrugs zu schützen, die Hand in Hand mit ihm geht&quot;, sagte er.

des Kremls sagte, dass | Putin &quot; beschlossen hatte, Ivanow von seinen Aufgaben Präsidialverwaltung zu entlasten&quot;, gab aber keinen Grund an.

Kunden, Vinyl verkauften, CDs zu kaufen, verkaufen die jetzt ihre CDs, um ihre A

hat Sommer Sie 2013 verloren Allround-Wettbewerb , lange keinen gehen eine

(b) Touch reorder.

Figure 5.4: Touch multi-word selection and reordering in MMPE.

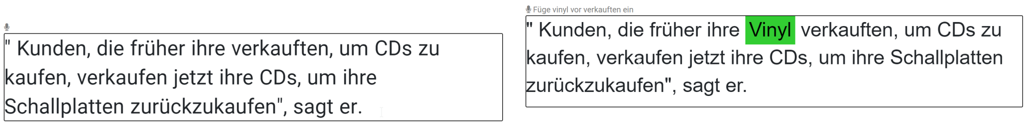
was also highly liked for deletion, we think removing this functionality does not harm overall usability. Furthermore, we added a delete button alongside the undo/redo/confirm buttons so that users can still delete using touch by selecting text through double-tap and pressing the button then. Overall we believe that the increased flexibility should enhance usability, even though touch deletion of single words became slightly more complicated.

Several participants in our study noted that the text was jumping around when reordering a word from the end of a line: By immediately removing the picked-up word from the text, all remaining words moved to the front, and the placeholder element was taking up space that also pushed words from line to line while dragging. We have now solved this issue by keeping the word(s) in the old position in a struck-through appearance (see Figure 5.4b), showing a copy of the word(s) below the finger/pen, and only removing the actual word(s) on drop. Furthermore, the visualization was redesigned to make the drop position clearer without taking up any space and highlighting the picked-up text better. As before, spaces between words and punctuation marks are automatically fixed, and, for German, nouns are automatically capitalized.

5.4 Improvements to the Speech Modality

According to the participants, speech would become especially compelling for longer insertions and would be preferable when commands remain simple. However, it was considered problematic in shared offices and would be complex to formulate commands while mentally processing text. To limit the complexity of speech commands, we added further synonyms (e.g., "*write*" or "*put*" as alternatives to "*insert*") and allow users to specify anchors by occurrence (e.g., "*delete last A*"). Thus, we increase flexibility and offer more natural commands that participants had used in our study in chapter 4, but which were not supported back then. Furthermore, we now allow modifying punctuation marks (e.g., "*delete comma after nevertheless*"), automatically capitalize words inserted at the beginning, uncapitalize them when reordered to other positions, capitalize the second word when deleting the first in the sentence, and so on. Users can now also choose to restate the whole sentence when MT quality is low, and in general, dictations are supported.

We also improved user feedback regarding speech commands: On the one hand, invalid commands display why they are invalid below the transcription (e.g., "Cannot delete comma after nevertheless, as nevertheless does not exist", or "There are multiple occurrences of nevertheless, please specify further"). On the other hand, it previously was hard to see if the speech module correctly interpreted the requested change because the text was simply replaced. Thus, the interface now temporarily highlights insertions in green, deletions in red (the space at the position), and combinations of green and red for reordering and replacements, where the color fades away 0.5s after the command. That way, the user can quickly see if everything worked as expected, or if further corrective commands are required, in which case a simple undo operation can be triggered (e.g., by simply saying "*undo*"). The updated speech module including change highlighting can be seen in Figure 5.5.



" Kunden, die früher ihre verkauften, um CDs zu kaufen, verkaufen jetzt ihre CDs, um ihre Schallplatten zurückzukaufen", sagt er.

(a) Target before speech command.

Füge vinyl vor verkauften ein

" Kunden, die früher ihre Vinyl verkauften, um CDs zu kaufen, verkaufen jetzt ihre CDs, um ihre Schallplatten zurückzukaufen", sagt er.

(b) Target after speech command with change highlighting.

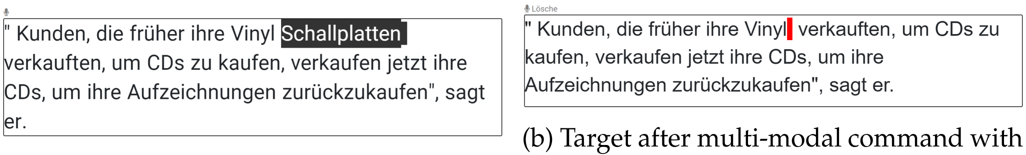
Figure 5.5: Speech commands with change highlighting in MMPE.

Other ideas we are currently working on include passing the text to the speech recognition to improve transcription results by considering the context (similar to the TransTalk project, Dymetman et al. (1994)) or training the automatic speech recognition towards the user to improve the received transcription.

5.5 Improvements to Multi-Modal Combinations of Pen/-Touch/Mouse with Speech

In the study in chapter 4, multi-modal interaction received good ratings for insertions and replacements, but worse ratings for reorderings and deletions. One big issue for deletions and reorderings was that multi-word (or partial word) reorder/delete was not supported in the previous implementation; thus, the translator had to place the cursor followed by a speech command multiple times. Due to the possibility of touch selection of multiple (or partial) words, this is now possible using multi-modal combinations of pen/touch/mouse combined with simplified speech commands, thereby hopefully enhancing the user experience. We further offer the possibility to keep the selection more straightforward, i.e., allowing the user to place the cursor at one position, but then state, e.g., "delete two words". This should improve situations where speech-only commands are particularly complex due to ambiguities, in which the combined approach was highlighted as advantageous to the speech-only approach. Furthermore, commands like "delete two words" based on a single cursor position are useful for multi-modal combinations including eye tracking, as discussed below.

Naturally, all other improvements for the speech case discussed above like the enhanced flexibility of commands, and especially the change highlighting also work for the multi-modal case, thus, hopefully making multi-modal interaction even more natural. An example image is shown in Figure 5.6.



" Kunden, die früher ihre Vinyl Schallplatten verkauften, um CDs zu kaufen, verkaufen jetzt ihre CDs, um ihre Aufzeichnungen zurückzukaufen", sagt er.

(a) Target before multi-modal command.

Lösche

" Kunden, die früher ihre Vinyl verkauften, um CDs zu kaufen, verkaufen jetzt ihre CDs, um ihre Aufzeichnungen zurückzukaufen", sagt er.

(b) Target after multi-modal command with change highlighting.

Figure 5.6: Multi-modal command of selection and speech in MMPE.

5.6 Introducing Eye Tracking for Multi-Modal Combinations

In the study reported in chapter 4, insertions are the only operation where the multi-modal approach was (non-significantly) faster than speech-only commands, since the position did not have to be verbally specified. We therefore investigate other approaches to enhance the multi-modal case: Apart from improving it by supporting multi-word reorder/delete and simplifying the speech commands as discussed above, we integrated an eye tracker. The idea is to simply fixate the word to be replaced/deleted/reordered or the gap used for insertion, and state the simplified speech command (e.g., "*replace with A*"/"*delete*"), instead of having to manually place the cursor through touch/pen/mouse/keyboard. Apart from possibly speeding up multi-modal interaction, this approach would also solve the issue reported by several of our participants that one would have to "do two things at once", while keeping the advantage of having simple commands in comparison to the speech-only approach.

In terms of hardware we now additionally integrated a remote eye tracker (the Tobii 4C, ca. 200 Euro[39]), which is attached to the screen and calibrated to the user with the eye tracking software provided. Other eye trackers could also be integrated in a similar way. Figure 5.7 shows the setup including the eye tracker.

Upon activation in the Angular client's navigation bar, a Python script is launched for the communication with the eye tracker. This script retrieves the raw gaze positions and pupil diameters of both eyes, further calculates fixations using a

[39] At the time of writing.

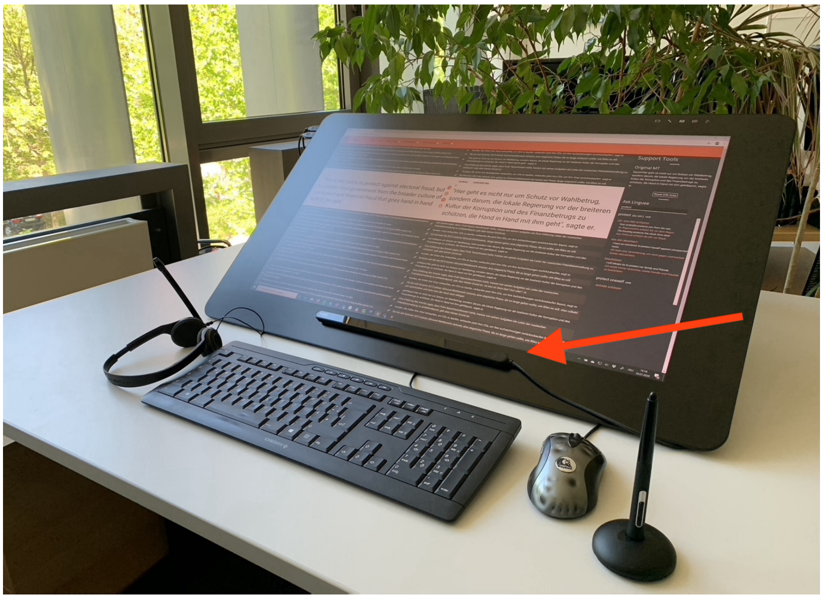


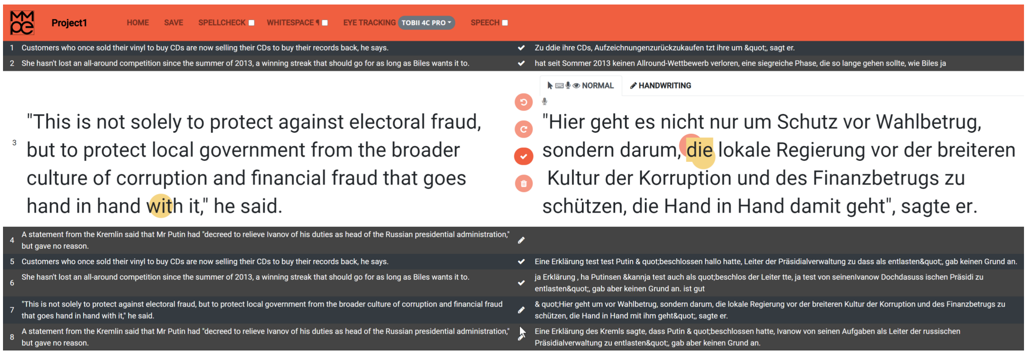
Figure 5.7: Hardware setup with a large tiltable screen, a digital pen, mouse, keyboard, a headset, and the newly integrated eye tracker (highlighted through an arrow).

dispersion-based algorithm, and sends the data back to the browser. Figure 5.8 shows the recognized current gaze position with a red circle. Based on the fixation, the client calculates the word belonging to the fixation and highlights it (yellow rectangle), which can then be used for multi-modal commands. If more than one word need to be changed, eye tracking naturally is sub-optimal at marking a text range. We therefore extended the speech commands to support the manipulation of multiple words, e.g., the user can look at the first word to be deleted and say "remove three words", or reorder multiple words by fixating the first one and stating "move four words after X".

Furthermore, we memorize and visualize (with a yellow circle) the last fixation on the source and target views, thereby helping translators navigating through the text not to get lost when switching their attention back and forth between source and target. This approach is similar to Gazemarks (Kern et al., 2010), which has shown its efficiency in visual search tasks with attention shifts. Apart from combining eye tracking with speech commands, we also plan to combine it with the keyboard, similar to the ReType approach (Sindhwani et al., 2019) but adapted towards the PE domain.

5.7 Improvements to the Log Files

We also worked on improvements and extensions to the logging functionality: Apart from bug fixes, we improved logs for copy and paste by adding the



1 Customers who once sold their vinyl to buy CDs are now selling their CDs to buy their records back, he says. ✓ Zu ddie ihre CDs, Aufzeichnungenzurückzukaufen tzt ihre um &quot;, sagt er.

2 She hasn't lost an all-around competition since the summer of 2013, a winning streak that should go for as long as Biles wants it to. ✓ hat seit Sommer 2013 keinen Allround-Wettbewerb verloren, eine siegreiche Phase, die so lange gehen sollte, wie Biles ja

3 "This is not solely to protect against electoral fraud, but to protect local government from the broader culture of corruption and financial fraud that goes hand in hand with it," he said.

"Hier geht es nicht nur um Schutz vor Wahlbetrug, sondern darum, die lokale Regierung vor der breiteren Kultur der Korruption und des Finanzbetrugs zu schützen, die Hand in Hand damit geht", sagte er.

4 A statement from the Kremlin said that Mr Putin had "decreed to relieve Ivanov of his duties as head of the Russian presidential administration," but gave no reason.

5 Customers who once sold their vinyl to buy CDs are now selling their CDs to buy their records back, he says. ✓ Eine Erklärung test test Putin & quot;beschlossen hallo hatte, Leiter der Präsidialverwaltung zu dass als entlasten&quot;, gab keinen Grund an.

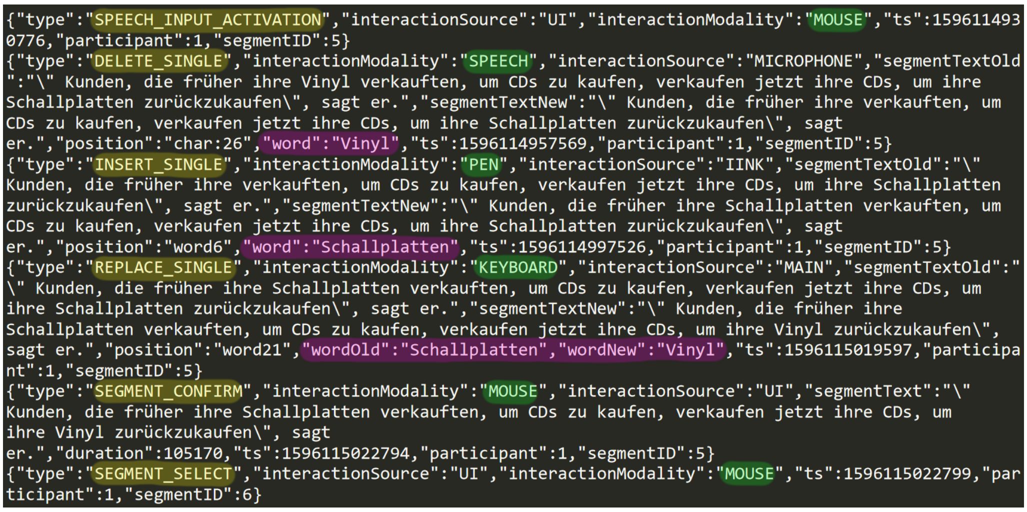
6 She hasn't lost an all-around competition since the summer of 2013, a winning streak that should go for as long as Biles wants it to. ✓ ja Erklärung , ha Putinsen &kannja test auch als quot;beschlos der Leiter tte, ja test von seinenIvanow Dochdasuss ischen Präsidi zu entlasten&quot;, gab aber keinen Grund an. ist gut

7 "This is not solely to protect against electoral fraud, but to protect local government from the broader culture of corruption and financial fraud that goes hand in hand with it," he said. & quot;Hier geht um vor Wahlbetrug, sondern darum, die lokale Regierung vor der breiteren Kultur der Korruption und des Finanzbetrugs zu schützen, die Hand in Hand mit ihm geht&quot;, sagte er.

8 A statement from the Kremlin said that Mr Putin had "decreed to relieve Ivanov of his duties as head of the Russian presidential administration," but gave no reason. Eine Erklärung des Kremls sagte, dass Putin & quot;beschlossen hatte, Ivanow von seinen Aufgaben als Leiter der russischen Präsidialverwaltung zu entlasten&quot;, gab aber keinen Grund an.

Figure 5.8: Eye tracking: The red circle (that is partly hidden) shows the current gaze position. The yellow circle on the left is the memorized last fixation on the source side. The highlighted word "die" shows the word that this fixation was mapped to. Applying multi-modal speech commands will take this marked word as input.

clipboard content, better distinguished between delete followed by an insert in comparison to replace operations, improved logs for reordering (better distinction into reorder-single, reorder-group, and reorder-partial), and provided more understandable logs for undo/redo. Furthermore, we improved logging for multi-modal commands: We do not merely save whether the interaction was multi-modal, but store whether it was a combination of speech and pen, or speech and mouse, or speech and finger touch. Last, we extended the logging functionality by adding gaze positions, fixations, and especially pupil diameter, which can be used for cognitive load analyses (see chapter 7). Figure 5.9 shows an extraction of a log file, to showcase the level of granularity of the high-level log, which simplifies subsequent analysis.



```
{"type":"SPEECH_INPUT_ACTIVATION","interactionSource":"UI","interactionModality":"MOUSE","ts":159611493
0776,"participant":1,"segmentID":5}
{"type":"DELETE_SINGLE","interactionModality":"SPEECH","interactionSource":"MICROPHONE","segmentTextOld
":"\" Kunden, die früher ihre Vinyl verkauften, um CDs zu kaufen, verkaufen jetzt ihre CDs, um ihre
Schallplatten zurückzukaufen\", sagt er.","segmentTextNew":"\" Kunden, die früher ihre verkauften, um
CDs zu kaufen, verkaufen jetzt ihre CDs, um ihre Schallplatten zurückzukaufen\", sagt
er.","position":"char:26","word":"Vinyl","ts":1596114957569,"participant":1,"segmentID":5}
{"type":"INSERT_SINGLE","interactionModality":"PEN","interactionSource":"IINK","segmentTextOld":"\"
Kunden, die früher ihre verkauften, um CDs zu kaufen, verkaufen jetzt ihre CDs, um ihre Schallplatten
zurückzukaufen\", sagt er.","segmentTextNew":"\" Kunden, die früher ihre Schallplatten verkauften, um
CDs zu kaufen, verkaufen jetzt ihre CDs, um ihre Schallplatten zurückzukaufen\", sagt
er.","position":"word6","word":"Schallplatten","ts":1596114997526,"participant":1,"segmentID":5}
{"type":"REPLACE_SINGLE","interactionModality":"KEYBOARD","interactionSource":"MAIN","segmentTextOld":"
\" Kunden, die früher ihre Schallplatten verkauften, um CDs zu kaufen, verkaufen jetzt ihre CDs, um
ihre Schallplatten zurückzukaufen\", sagt er.","segmentTextNew":"\" Kunden, die früher ihre
Schallplatten verkauften, um CDs zu kaufen, verkaufen jetzt ihre CDs, um ihre Vinyl zurückzukaufen\",
sagt er.","position":"word21","wordOld":"Schallplatten","wordNew":"Vinyl","ts":1596115019597,"participa
nt":1,"segmentID":5}
{"type":"SEGMENT_CONFIRM","interactionModality":"MOUSE","interactionSource":"UI","segmentText":"\"
Kunden, die früher ihre Schallplatten verkauften, um CDs zu kaufen, verkaufen jetzt ihre CDs, um
ihre Vinyl zurückzukaufen\", sagt
er.","duration":105170,"ts":1596115022794,"participant":1,"segmentID":5}
{"type":"SEGMENT_SELECT","interactionSource":"UI","interactionModality":"MOUSE","ts":1596115022799,"par
ticipant":1,"segmentID":6}
```

Figure 5.9: Logging of text manipulations in an easily interpretable granularity.

5.8 Mid-Air Hand Gestures for Post-Editing

Apart from integrating eye tracking, we further conducted detailed analyses on the use of mid-air hand gestures for PE. For this, we (i) investigate which mid-air gestures combined with the keyboard (GK) are suitable for which text-editing operations in PE, (ii) build a prototype supporting PE using GK, and (iii) analyze editing times and subjective feedback on mid-air hand gestures compared to mouse and keyboard (MK) for specific PE operations[40]. This section is a summary of publication Jamara et al. (2021), to which we refer the interested reader for further details.

5.8.1 Gesture Elicitation Study

While gestures on their own are naturally not a good modality for text input (as also stated by the participants of our elicitation study, see chapter 3), the use of gestures in combination with the keyboard might be a suitable alternative to the mouse. To investigate this, we started our research with a gesture elicitation study with 14 professional translators, conducted online due to the Corona pandemic. We analyzed the proposals of our participants for the referents delete (a single word or a group of items), reorder (single and group), replace (single and group), as well as insert, to define a suitable set of gestures for PE. Performing these referents implicitly includes other operations, namely selecting a position, a word, or a group of words/characters.

While analyzing the data, consistent patterns emerged: Similar to the way the mouse is used, participants performed all referents by first selecting the text, then performing the editing operations, e.g., deleting. Consequently, we decided in our analysis to separate the selection gestures from the editing operation gestures, analyzing and discussing each separately. In addition, the proposed selection gestures are divided into two types: the selection of a single item and the selection of a group of items.

Among the 8 unique gestures proposed for **group selection**, two were the most common: *both indices* (pointing with index fingers and moving them apart to select: see Figure 5.10a) and *index + thumb* (pointing with pinched index finger and thumb and separating them to select a range). *Index + thumb* and *both indices* appear to also be preferred in **selecting a single item**. In addition, the gesture *pointing* (where a participant points with the index finger to place the cursor on the item) was often proposed for single item selection.

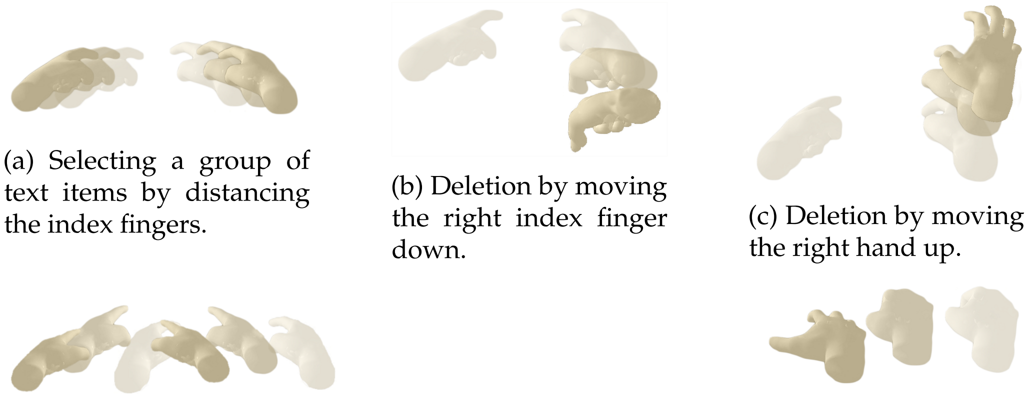
For the **deletion** referents, three gestures were suggested most often among the participants. Those were: move *right index down* (Figure 5.10b), move *right index up*, and move the *right hand up* (Figure 5.10c). Moving the *right hand up* to delete was also common for the **replace** referents. Analyzing participants' thoughts, which were captured via think-aloud protocol, it appears that they wanted to

[40] Both studies have been approved by the university's ethical review board.

delete first and then type the replacement item. Another common proposal for replacement was to simply type after selecting a text.

The **reordering** referents received three distinct gestures: The first one was to select and move the text with *both hands* by moving them simultaneously (Figure 5.10d). The second gesture was to point with the *right index* finger and start moving it to move the text immediately after selecting. The third gesture was to *grab* with the right hand and move the hand to reorder the text, then open it to release (Figure 5.10e).

Finally, the **insertion** referent received five unique gestures. One of the proposals was to *point* with the right index finger and then move it to place the cursor in the required place. Once the cursor was placed in the target position, the user would switch to the keyboard for typing.



(a) Selecting a group of text items by distancing the index fingers.

(b) Deletion by moving the right index finger down.

(c) Deletion by moving the right hand up.

(d) Reordering by moving both hands simultaneously.

(e) Reordering by grabbing, moving, and releasing.

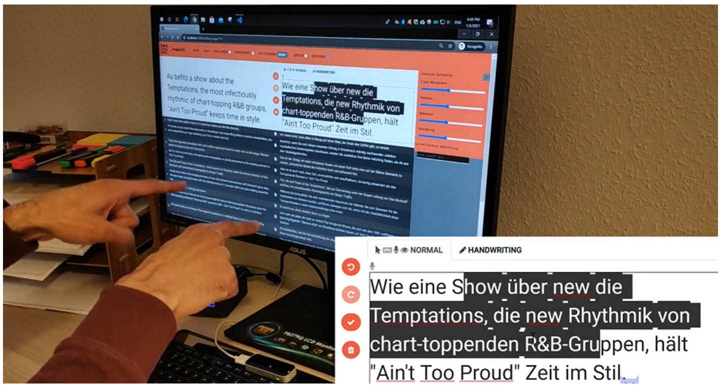
Figure 5.10: Common hand gestures for PE tasks proposed in our gesture elicitation study.

Together, these findings constitute a gesture set for text editing. Our separation into selection (for single items and groups) and editing operations makes the PE tasks more consistent and better represents our participants' mindsets. What is interesting is that selection of single items achieved high agreement by simply placing the cursor on the item, without selecting it from start to end as with the mouse. The deletion and replacement referents shared some gesture proposals because participants often wanted to replace by deletion followed by typing.

5.8.2 Prototype

We used the elicitation study results to define our final gesture set and implement a prototype. For this, the frequently proposed gestures were explored in terms of implementation feasibility. If two gestures were conflicting, we dropped the less popular one; otherwise we slightly modified it to resolve the conflict.

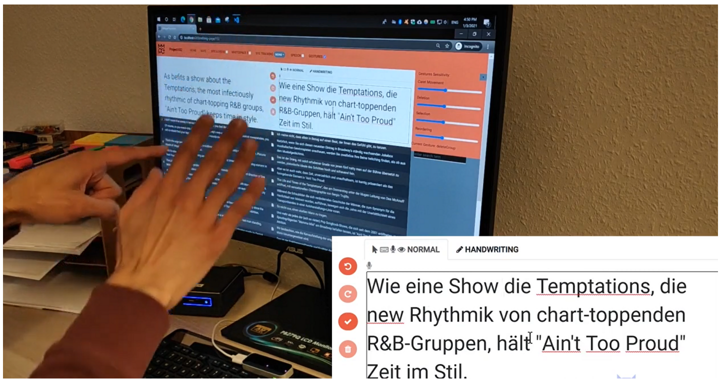
For **group selection**, we found that the proposed *index + thumb* gesture practically fails upon selection across multiple lines; thus, we dropped it. In contrast, using *both indices* can perform this kind of selection, so we implemented it as depicted in Figure 5.11. Note that in contrast to the mouse, the group selection using both index fingers allows the user to manipulate both ends of the selection continuously instead of having one side fixed. For **single item/position selection**, we implemented *pointing* with the right index finger. For multi-line text, both single and group selection allow moving the index finger vertically and horizontally.



Wie eine Show über new die
Temptations, die new Rhythmik von
chart-toppenden R&B-Gruppen, hält
"Ain't Too Proud" Zeit im Stil.

Figure 5.11: Mid-air gesture-based group selection by pointing with both indices.

For **deletion**, we implemented *hand or finger movement down and up* to to support all often proposed gestures (Figure 5.12). Note that for single item deletion it is sufficient if the cursor is placed somewhere on the word; there is no need to define the start and end of the word through a group selection.

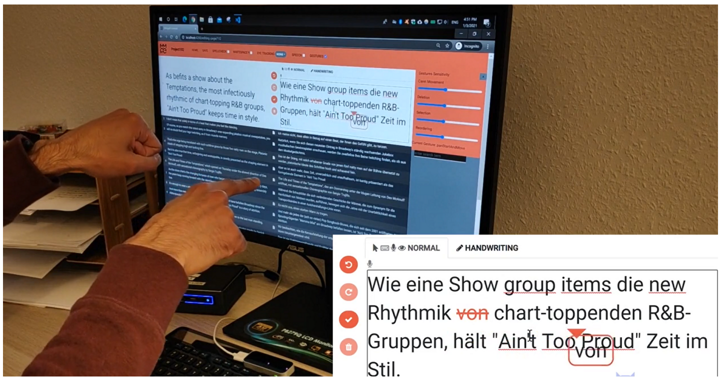


Wie eine Show die Temptations, die
new Rhythmik von chart-toppenden
R&B-Gruppen, hält "Ain't Too Proud"
Zeit im Stil.

Figure 5.12: Mid-air gesture-based deletion by moving the right hand or finger up or down.

Replacement can be achieved by either performing a group selection and typing directly, or by selecting a single item or group of items, deleting, and then typing. Note that single item replacement can thus also be achieved without performing a group selection.

The most complicated gestures were proposed for **reordering**; the gestures are a compound of several sub-gestures. Since reordering using the *right index* conflicts with cursor movement, we dropped it. Moving *both hands* while in the selection position turned out to be difficult to perform, as maintaining the same distance between the hands at all times is challenging. Therefore, we decided to merge it with the *grab* proposal; thus, after selection, a grab with the left hand indicates the start of the reordering process. Then moving both hands or just the right index finger reorders the text (Figure 5.13). Once the required position is reached, closing the right hand ends the reordering process and drops the text in the target position. For single item reordering, it is again sufficient to place the cursor on the item without selecting the whole text.



Wie eine Show group items die new Rhythmik ~~von~~ chart-toppenden R&B-Gruppen, hält "Ain't Too Proud" Zeit im Stil.

Figure 5.13: Mid-air gesture-based reordering by selecting, left grab, pointing with the right index finger to the target position, and right grab.

Our gesture detection relies on the *Leap Motion Controller*[41], which is small in size (8cm * 3cm) and can be placed on the top of the keyboard (Figure 5.13). The device provides frames of detected hands with 3D positions of finger joints, as well as some basic detection such as whether the fingers are extended or not. Based on this information our gesture detection algorithm determines if one of the above gestures is being performed. If only the right hand is detected with the index fingers extended, then the cursor gets updated based on hand movement. Moving both index fingers selects the corresponding text in the interface (Figure 5.11). When a deletion gesture is detected, the selected text (for group selection), or the word that the cursor is currently positioned on, is removed (Figure 5.12). A grab with the left hand puts the currently selected text/word containing the cursor in a reordering visualization. Then, movements of the right index are tracked and move the highlighted text as well as an arrow indicator visualizing the currently calculated drop position. Releasing the grab then places the text back into the input field at the indicated position (Figure 5.13). To avoid unintended gestures while moving the hands back to the keyboard, the user can form a grab in both hands after executing a gesture. Since people move their hands at different speeds, we further added sensitivity settings for gestures,

[41] https://www.ultraleap.com/product/leap-motion-controller/

similar to the standard mouse settings. A video showing the interactions in practice can be found under: `https://youtu.be/qIRYeojkFVc`. Note that the prototype facilitates all editing types required for PE.

5.8.3 Prototype Evaluation

To evaluate our implementation in practice, we conducted another study similar to the evaluation of the main prototype (see chapter 4), that tested every editing operation in isolation against the traditional mouse and keyboard approach. Unfortunately, due to the pandemic, we could only run a small scale study with 8 participants from our DFKI research department, instead of consulting professional translators. To mitigate the difference between non-translation professional subjects (computer scientists) and translation professionals, we ensured that similar to professional translators, (i) all our participants have academic training (computing degrees instead of translation degrees), (ii) that they are also highly familiar with traditional mouse and keyboard interfaces and use them in their day-to-day work, (iii) all subjects have relevant language proficiency (source EN, target DE), and (iv) all work in a multilingual EN-DE environment. Furthermore, as the evaluation required participants only to perform pre-specified text editing operations in the two conditions gesture and keyboard (GK) as well as mouse and keyboard (MK), without involving any linguistic translation decisions, we hope to minimize the effect of not having translators as participants.

Qualitative data was collected by the semi-structured interviews and 7-point Likert rating scales (7 = "strongly agree") as to whether the gesture is (a) a good match for its intended purpose, (b) easy to perform, and (c) a good alternative to MK. Figure 5.14 shows that operations manipulating single items were generally rated higher than operations on groups of items. Deletion of a single item was rated best, especially in terms of goodness and ease of use. The majority of our participants commented that group selection was hard to perform, whereas the editing operations themselves were considered easy. While comments differed depending on the referent, most of them were positive, and we frequently got statements such as "it is great, [GK] felt like the same level of MK".

Quantitative data, shown in Figure 5.15, captured the editing duration of both GK and MK for each referent, showing that the GK interquartile range was higher than the standard mouse and keyboard, except for group reordering. However, the most interesting finding was that, although the participants had years of experience using mouse and keyboard and were new to gestures, the average editing time in the gesture condition was very close to the average in the mouse condition in 4 out of 7 referents. Similar to what we found in the qualitative analysis, gestures operating on single items were more efficient than group operations in the GK condition. Single item deletion was the fastest, followed by single item replacement and insertion. On the other hand, group operations turned out to be the most time-consuming in both conditions, with the biggest differences between conditions for group deletion and group replacement.

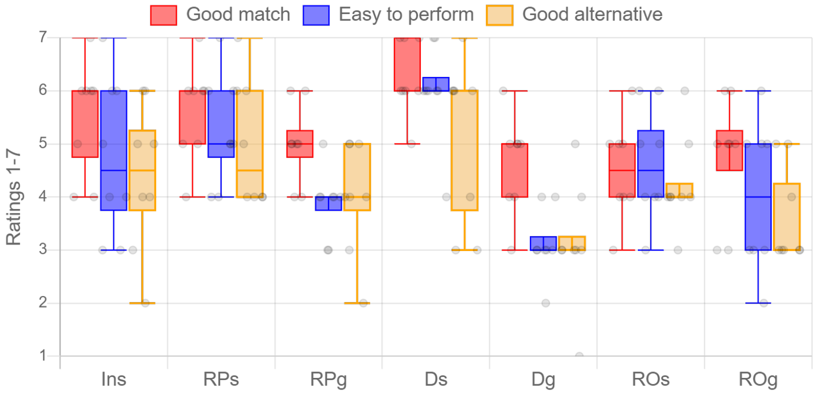


Figure 5.14: Subjective ratings in the gesture prototype evaluation for the referents insertion (Ins) single and group replacement (RP_s, RP_g), single and group deletion (D_s, D_g), and single and group reordering (RO_s, RO_g).

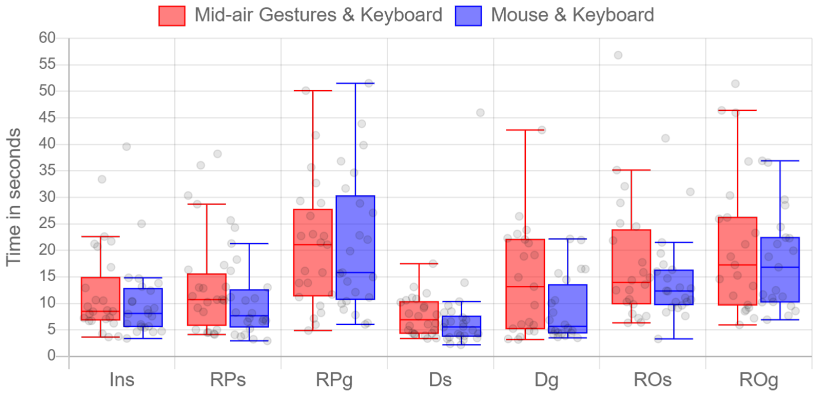


Figure 5.15: Editing duration of "gestures and keyboard"/"mouse and keyboard" in the gesture prototype evaluation for the referents insertion (Ins) single and group replacement (RP_s, RP_g), single and group deletion (D_s, D_g), and single and group reordering (RO_s, RO_g).

5.8.4 Conclusion

The findings overall suggest that GK could be a suitable interaction modality for PE and thus merits further research: Even though participants had years of experience with MK, our quantitative analysis of editing time showed that GK was only slightly slower for most operations, especially when manipulating single items. Similarly, qualitative data shows that manipulating single items was rated higher than operations working on groups of items, as participants found the group selection gesture "cumbersome" to perform, especially when selecting across multiple lines. This finding indicates that further effort should be invested in improving group operations, which are also common in PE (e.g. by exploring if a different placement of the detection device could increase detection accuracy). However, the appealing results on single item operations and the satisfactory results on group operations bode well and could provide benefit to the PE process as a complement, not replacement, to traditional mouse- and keyboard-based editing, which warrants further exploration with professional translators in a realistic PE scenario.

5.9 Guiding Post-Editors through Quality Estimation

Word-level QE assigns a quality label to each word or gap in a sentence (without access to a reference) and can thereby guide the post-editor to potentially erroneous parts of the sentence as well as good subphrases that can be recombined or adapted. However, the quality of QE systems is crucial, as incorrect QE might lead to translators missing errors or wasting time on already correct MT output. While word-level QE research has considerably advanced, achieving accurate automatic word-level QE remains very hard: even the best models currently only achieve F1 scores in the range 60% to 63% (Lee, 2020; Specia et al., 2020) depending on the text domain and the underlying MT system used to generate the data. With the few exceptions discussed in section 2.3.1, existing studies usually focus on detailed technical analyses on the test data, without user studies to investigate (i) if the quality levels achieved by the QE systems are already useful for human PE, and if not, what quality levels are required, and (ii), how to best present word-level QE information to translators. We address these research questions in terms of a well-controlled experiment with professional translators who are presented with (real and simulated) word-level QE output of varying quality in two different visualization schemes. This section is a summary of publication Shenoy et al. (2021).

5.9.1 Concept and Implementation

To ascertain which quality level would be sufficient to start helping the PE process, we artificially generate data at quality levels higher than what can be achieved by current state-of-the-art word-level QE models. The process of simulating QE output, as well as integrating their visualization into a translation environment, is explained in the following sections.

Artificial Generation of QE Output

We are not able to predict what exactly the output of a word-level QE model achieving 95% F1 would look like, i.e., which kinds of MT errors the model could detect well and which might be classified wrongly. The best we can do is to assume that a higher quality QE model would be similar to current QE models, but gradually improving on all parts of the MT output where current models fail. We therefore first conduct a pre-analysis to understand the kinds of errors of a current QE model, which we then leverage to generate artificial QE output of higher quality. For this, we flip labels of the ground truth annotations, while taking into consideration the parts of speech (PoS) that current QE models are more likely to classify incorrectly, instead of flipping labels fully randomly.

Our pre-analysis using the TextBlob library[42] for PoS tagging and the "QEBrain" model (Wang et al., 2018)[43] for generate quality predictions reveals the probability of each PoS and the corresponding conditional probability for being incorrect (given as ($P(PoS)$, $P(error|PoS)$)): nouns are most often wrong (32%, 48%), followed by prepositions (10.9%, 15.6%), pronouns (8.69%, 14.8%), determiners (13.04%, 14.3%), conjunctions (7%, 13.9%), interjections (4.5%, 12.9%), verbs (28%, 9.6%), adjectives (4.34%, 7.9%) and adverbs (5%, 2.9%).

Using those values, we simulate the error behavior of QE models achieving a certain quality level by flipping the ground truth QE label (OK or BAD) of the words depending on the conditional probability of the corresponding PoS. We use the following equation to determine the flip probability per PoS:

$P(flip|PoS) = P(error|PoS) * \frac{F1_{base}}{F1_{target}}$

where $P(error|PoS)$ is the conditional probability computed in the pre-analysis, $F1_{base}$ is the F1 score of the real QE model used in the pre-analysis, and $F1_{target}$ is the F1 quality score that we artificially generate. Since the error likelihood of nouns is highest, this ensures that it is more likely to flip a noun's label than that of an adverb. To get confidence scores for simulated QE models, we simply randomly sample values above 0.5 for 'OK' predictions, and values below 0.5 for 'BAD' predictions.

The limitations of our approach are that it assumes a constant error distribution, in the sense that higher quality QE models would just make proportionally fewer errors in each category, and that confidence scores are simply randomized. Of course, this is debatable, but, given that we cannot know exactly what a higher quality QE model would look like, we believe that this simple approach is a reasonable starting point for our investigations.

Visualization of QE output

Apart from QE quality, the visualization of QE output might also impact whether QE helps or hinders PE. We designed two alternatives, called *Binary-* and *Gradient-*based visualization schemes, as shown in Figure 5.16. In the binary visualization, quality is represented by simply coloring words in green and red depending on

[42] https://pypi.org/project/textblob/. Since PoS tagging accuracy is fairly high for high-resource languages and general text domains, we expect limited impact of PoS errors.

[43] This QE model was chosen because it was the best performing system according to the assessment carried out by Shterionov et al. (2019).

the QE output based on a level threshold of 0.5[44]. While this seems intuitive and easy to understand, uncertainties of the QE model cannot be depicted in the binary visualization. To tackle this, the gradient-based visualization directly shows the floating point number output of the QE model in the interval [0....1] by mapping the output to a color gradient ranging from red to green. Thus, the darker the shade of green, the more correct the model estimates the word to be. At the same time, this additional information about model uncertainties may well be confusing or overwhelming for the human post-editors.

Hawkeye, ein reformierter Unterwelt Snoop, der Gangster und andere Kriminelle verfolgt.

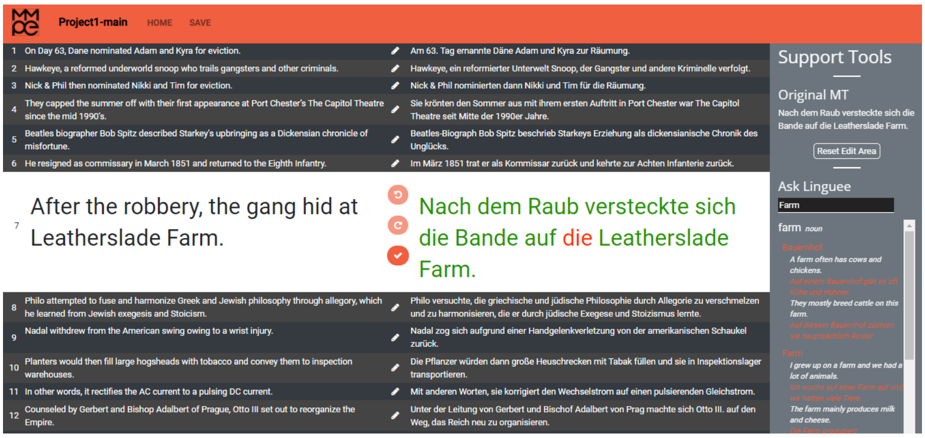
(a) Binary QE visualization.

Hawkeye, ein reformierter Unterwelt Snoop, der Gangster und andere Kriminelle verfolgt.

(b) Gradient-based QE visualization.

Figure 5.16: Binary and gradient-based QE visualization schemes.

We extend MMPE for QE by loading and visualizing the (real and simulated) QE models' quality predictions per word of the MT output from the project files (Figure 5.17) and extending the logging functionality to capture the QE condition and visualization scheme. When a user manually changes a word flagged by the QE system, its color is changed to black because we assume post-edits done by a user to be correct. Ideally, we could re-run the QE to obtain new scores for all the unchanged parts of the segment after each edit. However, as our simulated QE models rely on ground truth data, which we only have for the original MT output without modification, this is not a possibility.



1 On Day 63, Dane nominated Adam and Kyra for eviction. — Am 63. Tag ernannte Däne Adam und Kyra zur Räumung.

2 Hawkeye, a reformed underworld snoop who trails gangsters and other criminals. — Hawkeye, ein reformierter Unterwelt Snoop, der Gangster und andere Kriminelle verfolgt.

3 Nick & Phil then nominated Nikki and Tim for eviction. — Nick & Phil nominierten dann Nikki und Tim für die Räumung.

4 They capped the summer off with their first appearance at Port Chester's The Capitol Theatre since the mid 1990's. — Sie krönten den Sommer aus mit ihrem ersten Auftritt in Port Chester war The Capitol Theatre seit Mitte der 1990er Jahre.

5 Beatles biographer Bob Spitz described Starkey's upbringing as a Dickensian chronicle of misfortune. — Beatles-Biograph Bob Spitz beschrieb Starkeys Erziehung als dickensianische Chronik des Unglücks.

6 He resigned as commissary in March 1851 and returned to the Eighth Infantry. — Im März 1851 trat er als Kommissar zurück und kehrte zur Achten Infanterie zurück.

7 After the robbery, the gang hid at Leatherslade Farm. — Nach dem Raub versteckte sich die Bande auf die Leatherslade Farm.

8 Philo attempted to fuse and harmonize Greek and Jewish philosophy through allegory, which he learned from Jewish exegesis and Stoicism. — Philo versuchte, die griechische und jüdische Philosophie durch Allegorie zu verschmelzen und zu harmonisieren, die er durch jüdische Exegese und Stoizismus lernte.

9 Nadal withdrew from the American swing owing to a wrist injury. — Nadal zog sich aufgrund einer Handgelenkverletzung von der amerikanischen Schaukel zurück.

10 Planters would then fill large hogsheads with tobacco and convey them to inspection warehouses. — Die Pflanzer würden dann große Heuschrecken mit Tabak füllen und sie in Inspektionslager transportieren.

11 In other words, it rectifies the AC current to a pulsing DC current. — Mit anderen Worten, sie korrigiert den Wechselstrom auf einen pulsierenden Gleichstrom.

12 Counseled by Gerbert and Bishop Adalbert of Prague, Otto III set out to reorganize the Empire. — Unter der Leitung von Gerbert und Bischof Adalbert von Prag machte sich Otto III. auf den Weg, das Reich neu zu organisieren.

Original MT: Nach dem Raub versteckte sich die Bande auf die Leatherslade Farm.

Bauernhof
A farm often has cows and chickens.
Auf einem Bauernhof gibt es oft Kühe und Hühner.
They mostly breed cattle on this farm.
Auf diesem Bauernhof züchten sie hauptsächlich Rinder.

Farm
I grew up on a farm and we had a lot of animals.
Ich wuchs auf einer Farm auf und wir hatten viele Tiere.
The farm mainly produces milk and cheese.
Die Farm produziert

Figure 5.17: Screenshot of the interface after integrating QE information.

[44] The chosen threshold can trade off how sensitive the shown QE annotations are, so it can potentially trade off editing time for correctness. Our approach of using 0.5 is simple and straightforward, also often used in logistic regression and similar classification by just showing the tendency of the model; nevertheless future research should investigate different thresholds.

5.9.2 Evaluation

Using our implementation, we conducted a user study to assess the quality threshold when word-level QE starts facilitating and stops hindering the PE process and to investigate which word-level QE visualization is preferred.

Method

Due to the ongoing pandemic, the evaluation was conducted online[45] by 17 non color-blind translators working as freelancers on a platform called Upwork Global Inc[46]. All participants fulfilled the requirement of having either C1 or C2 language proficiency in both English and German. After filling in a demographics questionnaire, participants received an explanation of all of the prototype's features and pointers regarding the execution of the main experiment in a four minute introductory video[47]. After the video, translators had time to get accustomed to MMPE with word-level QE before delving into the main experiment.

The main experiment was the central part of the study and entailed PE of 32 text segments (8 text blocks of 4 segments each) with QE support. As text, we used EN-DE text from the training set of the WMT 2020 QE shared task (Specia et al., 2020) originating from Wikipedia, which relies on an up-to-date NMT model from the fairseq toolkit (Ott et al., 2019).

The segments were labeled by either the real QE model or simulated QE output with 75% F1, 85% F1 and 95% F1 quality levels created as described above. The real QE model proposed by Wang et al. (2018) was pre-trained on the training set of WMT 2018's QE shared task and fine-tuned on the Wikipedia domain, achieving a quality level of 63.5% F1 on the training set[48] of WMT 2020's QE shared task. QE information for each sentence was visualized either with the binary or gradient-based visualization. Since there are four quality levels and two visualization schemes, the experiment follows an 8*8 Balanced Latin Square: the text order is kept identical for each participant, but the QE quality and visualization are counter-balanced accordingly on the 8 text blocks of 4 sentences each. Thus, participants might be more exhausted for the same sentences towards the end of the experiment and better concentrated on the initial sentences; however, the effects of text and tiredness should cancel out for quality level and visualization due to the counter-balancing. This methodology allows us to analyze if visualization-x with QE-y is better than visualization-x' with QE-y' across text blocks. Moreover, the impact of translation skill or technical skill of first-time users of MMPE factors into all the conditions equally due to the chosen within-subject design.

[45] The study has been approved by the university's ethical review board.

[46] `https://www.upwork.com/`

[47] `https://youtu.be/6LgUzia_3pM`

[48] Note that the model was not trained on this data.

After post-editing each segment, participants had to press confirm, which we used to record the required time. When confirming, we also showed a pop-up asking "Was the word-level quality estimation helpful?" which the participant had to rate on a 9-point Likert scale ranging from 1 ("very strongly disagree") to 9 ("very strongly agree"). Apart from *duration* and *subjective ratings*, we measure per condition *edits* done and the *final quality* of the translations in terms of TER (Snover et al., 2006) by comparing the post-edited version of the translation to the reference. Overall, the study took approximately one hour per participant.

Results

We present our results in 4 categories: (1) subjectively assessed helpfulness per QE quality, (2) preferred visualization, (3) editing duration per QE quality, and (4) translation quality per QE quality.

Subjective Helpfulness of QE Quality We analyze subjective ratings across all segments with the same word-level QE quality level. To ensure independence of samples, we average the ratings for the same QE quality level per participant. As shown in Figure 5.18, the lower QE quality levels of 63.5% F1 and 75% F1 are consistently rated as less helpful (mean values of 1.5/3.25), while the higher quality levels of 85% F1 and 95% F1 are rated more helpful (mean values of 7/8). This indicates that in comparison to high quality QE, bad QE is not considered helpful. Two-tailed t-tests, testing each group against 5, which is the middle value along the subjective rating scale depicting neutrality, support this finding.

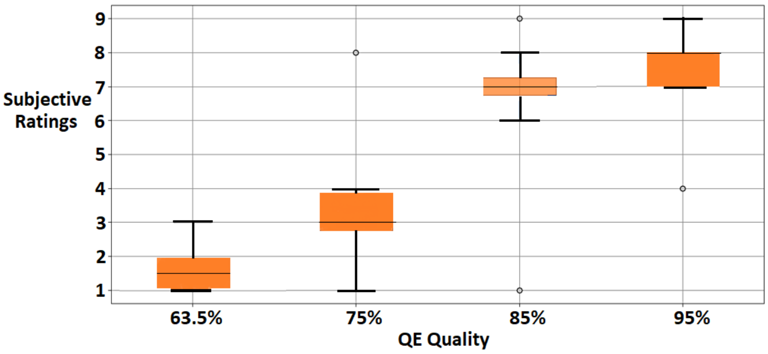


Figure 5.18: Subjective ratings per QE quality level.

Preferred Visualization per QE Quality To analyze preference of visualization, we use the same ratings[49], however, we multiply the ratings corresponding to sentences shown in binary visualization by -1 and normalize the obtained scores to the range [0...1]. Therefore, values close to 0 indicate a preference for the

[49] Independence of samples is achieved by averaging ratings per participant across segments with the same QE quality and visualization.

binary visualization scheme, while values close to 1 indicate a preference for the gradient-based visualization. The corresponding box-plot in Figure 5.19 shows that for word-level QE quality levels 63.5% F1 and 75% F1 binary is the preferred visualization scheme (preference value below 0.5). In contrast, for higher quality levels of 85% F1 and 95% F1, the gradient-based visualization is preferred (preference value above 0.5). Again, two-tailed t-tests against the middle value (0.5) representing equal preference towards both visualizations confirms this finding statistically.

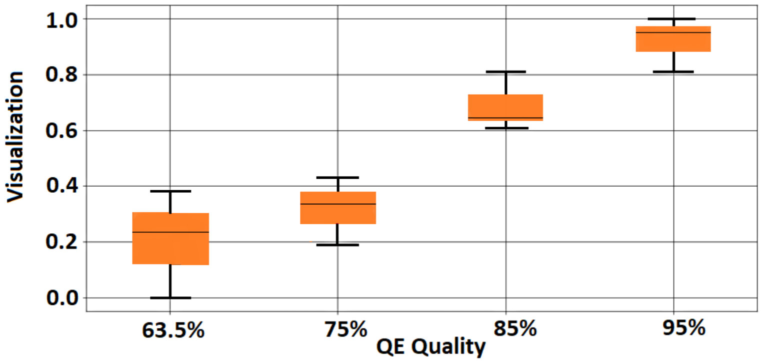


Figure 5.19: Visualization preference per QE quality level.

Editing Duration per QE Quality Apart from subjective ratings (used for QE helpfulness and visualization preference), we also capture the time taken to post-edit the segments per QE quality. We average the duration across the segments having the same QE quality per participant to make the observations independent. The box plot in Figure 5.20 depicts that when the QE quality is low the duration taken to post-edit the segments is high, whereas translators are fast when the QE quality is high. To find out whether the differences in duration are significant, we run a one-way ANOVA followed by Tukey HSD post-hoc test, showing that all pairs except for 85% F1 vs. 95% F1 are significantly different.

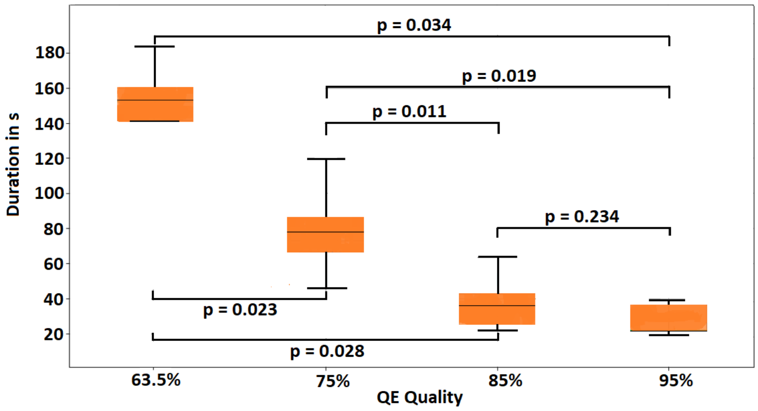


Figure 5.20: Duration per QE quality level with Tukey HSD p-values.

Translation Quality per QE Quality We have seen that translators subjectively find high quality QE helpful and post-edit fast with it. In order to analyze whether they are fast just because they blindly trust and follow the QE system (even when they should not) or because the system actually helps, we evaluate the quality of the resulting translations. As before, the scores were averaged across segments having the same QE model per participant to make the observations independent. The box plot in Figure 5.21 shows that the TER of the post-edited version against the reference is low when the QE quality is high, and by contrast, it is high when the QE quality is low. Since a low TER score implies a high quality translation, translations get better with increasing QE quality.

We speculate that the reason for the high TER score for the 63.5% F1 QE model is that the translators produced a different translation than the reference. This different translation may or may not be accurate; we cannot know for sure without manual evaluation. Nonetheless, from our automatic quality evaluation we are certain that with better QE the final translations get closer to the reference. In order to find out whether the differences are significant, we again run a one-way ANOVA followed by Tukey HSD test showing that indeed all pairs are significantly different.

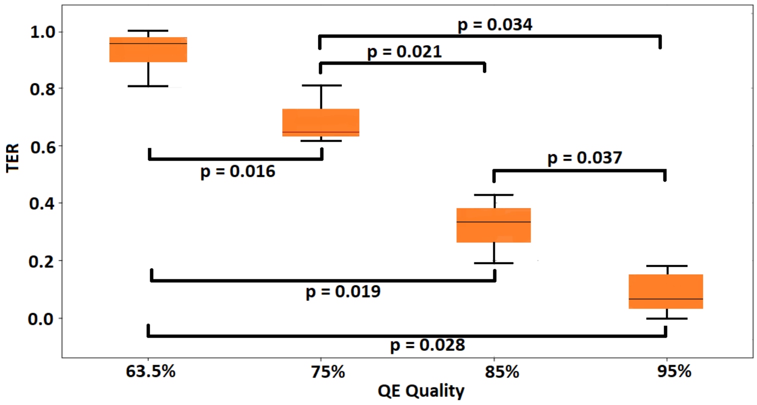


Figure 5.21: TER scores per QE quality level with Tukey HSD p-values.

5.9.3 Conclusion

Our results show that existing state-of-the-art word-level QE systems are not yet good enough to be helpful during PE. Instead, all our analyses agree that QE systems need an F1 score of at least 80% to support PE in terms of subjective helpfulness, editing duration, and quality of the final translations. This establishes a target for future QE research. In terms of visualization, the word-level quality scores should be visualized using gradient-based visualization which also shows uncertainties of the model, since the binary approach was considered superior only in cases where QE was not helpful. This preference is interesting as the exact color chosen for the gradient was randomly sampled in the red/green range for BAD/OK ratings for the artificial QE output. A reason for preferring the gradient-based visualization could be a stronger involvement in the decision process. With increasing QE quality, PE becomes more efficient, where in particular the higher quality levels of 85% and 95% require less editing time than the lower quality levels. Lastly, we found that translation speed gains are not merely a result of blindly trusting the QE system, but indeed help producing translations that are closer to the reference. To sum up, a QE quality level of at least 80% F1 sets the approximate boundary where word-level QE starts helping translators, and for these QE quality levels, a gradient-based visualization is preferred.

5.10 Conclusion

In this chapter, we first used the previously presented feedback from professional translators to improve and extend the existing prototype: we redesigned the layout, added visualization of whitespaces, fixed issues in hand-writing, allowed multi-word reordering using touch drag and drop and improved its visualization, extended the speech commands, provided better feedback for the user on what the speech commands changed, and improved the logging functionality. Furthermore, we showcased how eye tracking can be integrated, not only for logging but as an actual interaction modality that can be used in combination with speech recognition or the keyboard to quickly correct errors. Guided by another elicitation study, we also integrated mid-air hand gestures that can be used in combination with the keyboard for PE, which turned out much more promising than initially expected in our study. Finally, we extended MMPE to visualize word-level QE predictions and conducted a study that showed that current state-of-the-art QE models are not yet good enough to support the PE process; instead, quality levels of at least 80% F1 must be reached, and for these quality levels a gradient-based QE visualization is preferred. The MMPE CAT tool is now available open source on Github[50].

The next obvious steps would be running follow-up studies, as we will discuss in chapter 14. Implementation-wise, to transform MMPE into a fully fledged translation workbench, further project management functionality, direct loading of common file types like .docx, and support for more language pairs are required.

This chapter ends our investigations regarding the goal of investigating modalities other than mouse and keyboard in the PE process. We have seen that, as hypothesized, different modalities perform well for different PE tasks and could indeed benefit the PE process according to our studies. MMPE combines the most modalities simulatenously explored in the CAT and PE literature and is now available to the public. In fact, we have already used MMPE in follow-up research, e.g., on interactive PE (Akmal, 2021) and for exploring whether seeing more than one MT proposal per segment is beneficial for PE (Wang, 2021), but exclude these investigations from the scope of this dissertation.

[50] https://github.com/NicoHerbig/MMPE

Part III

Multi-Modal Cognitive Load Estimation

Now that we have discussed how *explicit* multi-modal user input can support PE, we move on and explore if and how *implicit* multi-modal input in the form of sensory devices can be used to model the user state, in particular the perceived CL. We first discuss the vision of cognition-aware CAT tools and present findings from interviews conducted with professional translators to understand user expectations, in particular which kind of CL-based dynamic adaptations they imagine to be useful (chapter 6).

Afterwards, chapter 7 presents a multi-modal framework we developed, that combines a large variety of sensors from the literature (Demberg and Sayeed, 2016; Rowe et al., 1998; Stuyven et al., 2000; Villarejo et al., 2012) to estimate and analyze CL. With it, we ran three studies (chapter 8): First, a study with translation students who PE segments in a realistic environment while being monitored by the first version of our framework. Then two studies with the complete framework: The first one almost identical to the study with translation students but this time with professional translators. The second within the e-learning domain, to explore if our measures also work well in other areas than PE, where learning was a reasonable choice because CL theory originates from educational psychology.

Finally, chapter 9 presents results of a survey to understand if users are more concerned about the use of some types of sensors then about other types for estimating CL, and in particular, to figure out how big the improvements achieved through cognition-awareness need to be for users to consider sharing their data for this purpose.

Part III is based on publications Herbig et al. (2019c), Herbig et al. (2021), Herbig et al. (2020a), and Herbig et al. (2019d).

Chapter 6
Cognitive Load Adaptations

This chapter starts with a discussion why and how CL should be considered in the PE process. Then, we present translators' thoughts regarding adapting the translation user interface to measured CL based on interviews conducted as a side-investigation to the study presented in chapter 3.

This chapter is based on publications Herbig et al. (2019a) and Herbig et al. (2019c).

6.1 Rethinking the Post-Editing Process from a Cognitive Load Perspective

While it is widely acknowledged that MT can, in many situations, not be immediately used as is but requires further PE, the cognitive dimension of PE is often overlooked. This leads to a neglect of PE costs related to the way in which post-editors work with MT output, and instead to a focus on creating MT output that is as close as possible to an independently provided reference translation. Especially the PE task, however, 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 possibly 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. 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. Another 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 repeated 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).

Apart from CL, other proposals to measure the usefulness of MT for PE were made, e.g., recording PE time and effort (Guerberof, 2009; Zampieri and Vela, 2014b), quantifying in seconds and keystroke logs the difference between MT output and a human-acceptable translation. We argue that it is not only the amount of PE necessary or the PE time that should be considered, but the actual 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. 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.

6.2 Method

To gain insights from professional translators regarding this vision we conducted semi-structured interviews. The goal was to understand (1) if users have interesting ideas on how CL measurements could be used within the context of PE, and (2) which proposed adaptations to CL they would find useful. For this, participants are asked to suggest ideas themselves, and we discuss possibly interesting adaptations, which are rated on a simple 7-point scale ranging from "very bad" to "very good". Overall 13 (female=9, male=4) translators aged 28 to 62 participated. Since these interviews were conducted with the same translators as our elicitation study, further details on the demographics can be found in section 3.2.1.

6.3 Results

We present the results of our interviews by clustering the opinions regarding the different adaptation ideas.

Providing Additional Resources The idea to automatically receive additional resources when high loads are detected was often proposed by participants, e.g., to display a corpus or terminology proposals, to automatically provide MT alternatives, or to trigger concordance search. We also discussed this idea with participants who did not propose something related on their own. On average a rating of 5.15 was achieved on a 7-point Likert scale (min=1, max=7, σ=1.68). Overall, 10/13 participants were positive regarding such automatic adaptations (ratings in range [5–7]) stating that especially TM and MT alternatives would help, that research activities should be triggered automatically, and that this could avoid wasting time. However, it was also noted that adaptations need to be configurable, and that they risks showing irrelevant information.

Simplifying the Interface We also discussed the opposite idea: to hide elements when a high CL is detected, the hypothesis being that CAT interfaces are too complex (Zaretskaya and Seghiri, 2018) and considering all information might be overwhelming. Here, we see rather contrary opinions: only three participants were very positive in this regard (range [6–7]), while all others were against such simplifications or had a neutral opinion (rating in [1–4]). The average rating is therefore rather low: mean=3.39, min=1, max=7, σ=1.94.

Estimating CL of New Text We further discussed the idea of learning a CL estimation, mapping high/low CL experienced in the past to new tasks to estimate how demanding the translation will be. 9/13 participants provided very positive feedback, stating that this "is better assessment than just words", should be used for color-coding, or could help estimate the time and effort required for new jobs. 3 participants however showed mixed and 1 participant negative opinions towards this idea. On average the ratings were rather good: mean=4.92, min=3, max=7, σ=1.19.

Suggesting Breaks The idea to propose breaks automatically was also discussed: One translator proposed to display a coffee cup icon, while another argued that choosing break times more carefully might help achieve better quality. In general the feelings were mixed, with an average rating of 4.23 (min=1, max=7, σ=2.05), and many participants stating that automatic break proposals might help if they are configurable and only suggested but not enforced.

Reordering Segments Similarly, the idea arose to work on the texts/projects in a changed order, to avoid long periods of too high or too low CL and achieve a better balance. Of course, this would only be possible if the context allows it, but the participants claimed that this is feasible and many already adapt the order, e.g., by translating a table or a caption in between or by switching to a simpler translation project. This would help with "making better use of cognitive resources" because you are sometimes "blocked by focusing too much". However, apart from these many positive voices, four participants also stated that they "can do it on [their] own" or "would not want that". On average the translators gave their idea a rating of 4.15 (min=1, max=6, σ=1.46).

Other Ideas Other interesting ideas proposed by the participants were to pay translators based not only on time but also on CL, to adapt the font size, to make a profile about what is easy for which translator to find a good match between text and translator, or to estimate the remaining time based on complexity. Further suggestions were notifications about detected boredom, or increasing error tolerances for speech and handwriting recognition when under load.

In general, participants found potential adaptations to one's own cognitive state exciting and offering room for lots of improvement, but at the same time some found it to be "frightening" and "feeling manipulative".

6.4 Conclusion

PE is often considered merely from a productivity perspective and neglects the cognitive dimension involved in the process. Furthermore, MT research often ignores the fact that MT is often subsequently post-edited, and thus only focuses on quality as a comparison to a ground truth translation and not in terms of suitability for PE. We thus argued for the need to consider the cognitive dimension more closely in PE environments. With this motivation in mind, we interviewed professional translators to understand which CL adaptations could offer benefit during PE, and found that they were especially positive regarding the idea of a user interface that adapts to measured CL, particularly if it automatically provides additional resources like TM matches or MT proposals. We also received a vast amount of further ideas how CAT tools could adapt to estimated CL, which motivates us to further explore this vision.

We next focus on creating a framework combining a variety of sensors to estimate CL, and after that, present findings from 3 studies with the framework that indicate which modalities work better or worse for CL estimation.

This chapter thus formed the basis for our second research question: it motivated why CL should be considered more closely in the PE process, provided initial ideas how to adapt CAT tools to perceived CL, and finally discussed these ideas with professional translators to capture their expert knowledge and gain a better understanding of a possible CL-aware CAT environment.

Chapter 7
Multi-Modal Cognitive Load Estimation Framework

As stated in the beginning of Part III of this thesis, 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 (see section 2.4) provides a lot of studies in other domains; however, the question remains which of the related approaches are applicable here. Within the translation domain 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 from a variety of sensors during PE, which can be combined in a multi-modal fashion. 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.

This chapter is based on publications Herbig et al. (2020a), Herbig et al. (2021), and Herbig et al. (2019c), and focuses on the data capturing framework for robust CL estimation. The next chapter then uses this framework in three studies and discusses the subsequent data analysis, including the regression analysis from raw sensor values to subjective CL.

7.1 Architecture

To assess data from multiple modalities during PE, we implemented a framework combining all data in a simple and effective manner. 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 (like heart rate variability) based on the raw data (like the RR intervals), it is also possible to subscribe to specific events, process them, and send the resulting high-level feature back to the server. Thus, scripts post-processing raw sensor data effectively act as further high-level sensors themselves. A schematic overview of the architecture is depicted in Figure 7.1.

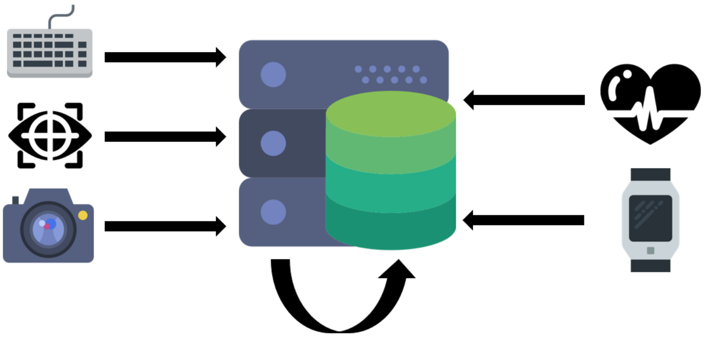


Figure 7.1: The architecture of our CL estimation framework with all sensors communicating with a central server in an event-based fashion[51].

All sensors used are depicted in Figure 7.2. We ensured that all devices included in our analysis do not significantly hinder the translation process or feel uncomfortable. This is also the reason why we did not include two-finger GSR sensors, head-mounted eye trackers, or EEG sensors. Furthermore, we aimed for a combination of rather cheap consumer-oriented devices (like the Polar heart belt) and research-oriented devices (like the Empatica E4) to compare both classes. We will now look at the individual subjective, performance, behavioral, and physiological measures.

[51] Icons by Smashicons & Freepik from `flaticon.com`.

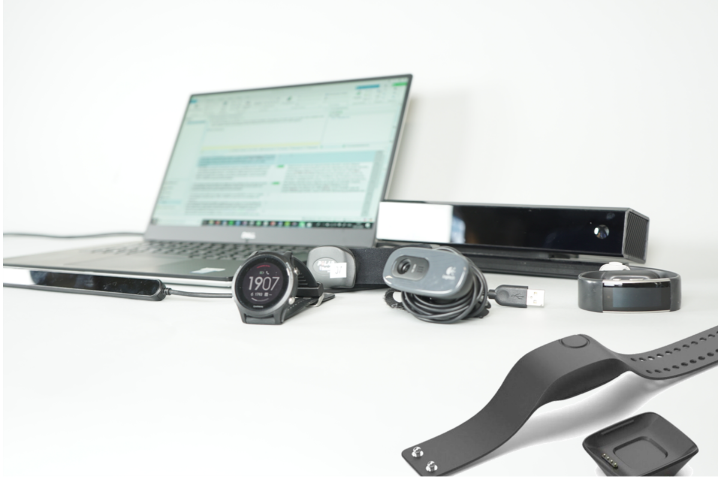


Figure 7.2: Overview of sensors used for CL estimation[52].

7.2 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 show a pop-up asking for a **subjective CL rating** (SubjCL[53]) using the scale proposed by Paas and Van Merriënboer (1994) after every single segment. This scale was chosen because it focuses on CL and not on quality, has been widely used and verified in many application areas, and further since it was used in the two most related studies by Vieira (2014, 2016). Furthermore, it 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 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".

7.3 Performance Measures: Time & Text

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 off quality for time and because translation quality is a partly subjective measure. Nevertheless, we integrate the following simple time and text measures:

[52] Empatica image taken from https://i.pinimg.com/originals/04/82/54/0482548efb5dfca4394c56610952800a.jpg.

[53] Here and in the following feature names used in the subsequent studies are defined.

7.3.1 Time Measures

For the time features we integrate **PE time** (PeTime) and **length-normalized PE time** which also considers the segment length (LNPeTime).

7.3.2 Text Measures

The text features consist of smoothed **BLEU**, **HBLEU** (Lin and 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. Thus, the H-based measures can be applied to any post-edited text, while the non-H measures only work in controlled experiments when a reference translation is available.

7.4 Behavioral Measures: Typing, Facial Expressions, & Body Posture

Behavioral measures can be extracted from user activity while performing a task.

7.4.1 Typing-Based Measures

Especially interesting in the context of PE, where traditionally 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 Average Pause Ratio (**APR**) (“the average time per pause in the segment divided by the average time per word in the segment”) and Pause to Word Ratio (**PWR**) (the number of pauses in a segment divided by the number of words in a segment) by Lacruz et al. (2012; 2014), which were shown to correlate with PE effort, are automatically calculated from the keyboard events.

7.4.2 Facial Expressions

A web-cam records images at a fixed interval which are then sent to the emotion recognition API from Microsoft Cognitive Service[54], returning a simple JSON

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

format with the likelihood of each of the **basic emotions** based on a trained neural network (EmotionName). While emotion recognition and CL estimation cannot be considered equal, it seems reasonable to include it here as well.

7.4.3 Body Posture-Based Measures

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).

7.5 Physiological Measures: Eyes, Heart, & Skin

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

7.5.1 Eye-Based Measures

For eye-based features, we use a web-cam and an eye tracker (see Figure 7.3).

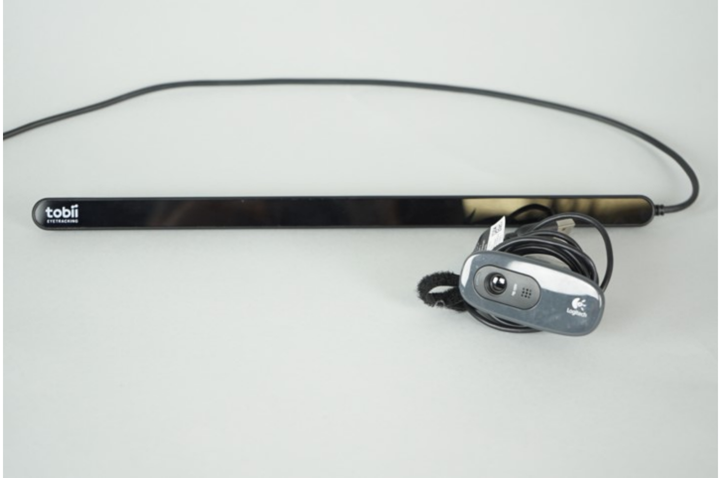


Figure 7.3: A Logitech webcam and the Tobii 4C eye tracker for eye-based CL measures.

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 and Cech, 2016). Even though the work did not explore if EAR is a suitable indicator of CL, it was included because intuitively a link might exist and the simplicity of using web-cams would make the CL measurement easily applicable in practice.

We further 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. Based on the recorded raw data, we calculate the amount of **blinking** (of less than 2 s length) ($\mathrm{BlinkAmount}$) and also normalize this by the PE time ($\mathrm{NormBlinkAmount}$) (Van Orden et al., 2001). Similarly, we calculate the amount of **fixations** ($\mathrm{FixAmount}$) and normalize it by PE time ($\mathrm{NormFixAmount}$). We further compute the fixation durations (FixDur) and **saccade** durations ($\mathrm{SaccDur}$) (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 and Kotval (1999) to calculate the **probability of visual search** based on the eye movements ($\mathrm{SearchProb}$), 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.

Note that we listed these features as *physiological* measures, even though eye movement and blinks are to a large extent controllable and are therefore often counted as *behavioral* features. However, we wanted to present and analyze all eye features, including the non-controllable pupil diameter (see below), together, and therefore list them as physiological measures.

To capture the **pupil diameter** (O'Brien, 2006a) ($\mathrm{PupilDiameter}$), we use the same eye tracker, however, with the Tobii Pro SDK. Even though this SDK is expensive (ca. 2300 Euro[55]), making it harder to establish pupil measures in real-world scenarios, we include the pupil diameter for comparison as it has been frequently used in CL studies. 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 and 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 amount of rapid dilations ($\mathrm{ICA}^{\mathrm{wave}}$), 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 ($\mathrm{ICA}^{\mathrm{count}}$). Last, we also implemented the work of Hossain and Yeasin (2014), which checks for sharp changes and continuations of the ramp in the Hilbert unwrapped phase of the pupil diameter signal ($\mathrm{Hilbert}$).

7.5.2 Heart-Based Measures

For heart measures, we integrate three devices: a Polar H7 heart belt, a Garmin Forerunner 935 sports watch, and the Empatica E4 wristband (see Figure 7.4). 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. Furthermore, we can compare the wrist-worn devices (Empatica and Garmin) to the chest-worn measurements by the Polar belt.

[55] At the time of writing.

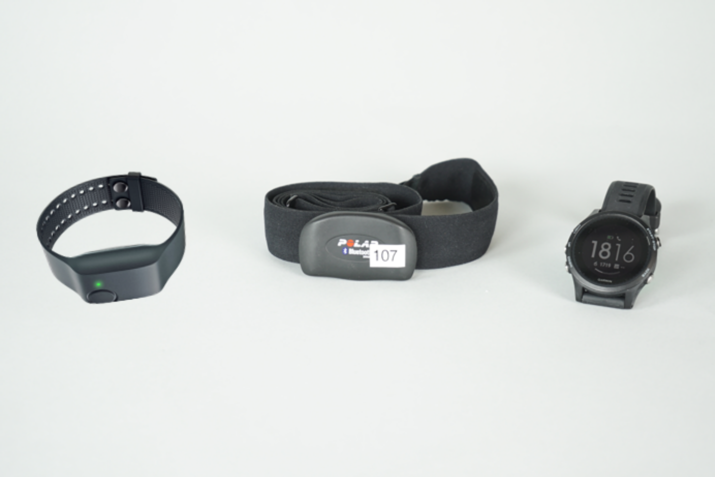


Figure 7.4: The Empatica E4, Polar H7 and Garmin Forerunner 935 for heart-based CL measures[56].

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 "R"s (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 is the number and percentage of successive NN intervals that differ by more than 50 ms (Shaffer and Ginsberg, 2017), for both the Empatica and the Polar to the analysis.

The Empatica further 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 (Iani et al., 2004) (BVPAmp), 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.

[56] Empatica image taken from `http://empatica.com`.

7.5.3 Skin-Based Measures

For skin-based features, we integrate the Microsoft Band v2 and again use the Empatica and the Garmin devices, as depicted in Figure 7.5.

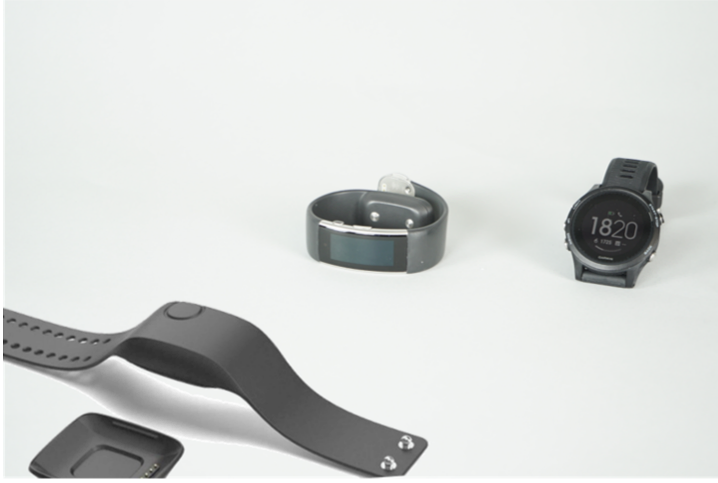


Figure 7.5: Empatica E4, Microsoft Band v2, and Garmin Forerunner 935 used for skin-based CL measures[57].

The Microsoft Band 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). Following 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[58] to calculate higher level skin conductance features on the Empatica raw data. It provides us with "global" features, namely the mean value ($\mathrm{Leda}_{\mathrm{avg}}$) and the maximum positive deflection ($\mathrm{Leda}_{\mathrm{MaxDefl}}$), and "through-to-peak (TTP)/min-max" analysis, namely the number of significant (i.e., above-threshold) skin conductance responses (SCRs) ($\mathrm{Leda}_{\mathrm{TTP.nSCR}}$), the sum of SCR amplitudes ($\mathrm{Leda}_{\mathrm{TTP.AmpSum}}$) of significant SCRs, and the response latency ($\mathrm{Leda}_{\mathrm{TTP.Lat}}$) of the first significant SCR. Furthermore, and most interestingly, we use Ledalab to perform a Continuous Decomposition Analysis (CDA) (Benedek and 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 ($\mathrm{Leda}_{\mathrm{CDA.nSCR}}$, $\mathrm{Leda}_{\mathrm{CDA.AmpSum}}$, $\mathrm{Leda}_{\mathrm{CDA.Lat}}$). Furthermore, the average phasic driver ($\mathrm{Leda}_{\mathrm{CDA.SCR}}$), the area of phasic driver ($\mathrm{Leda}_{\mathrm{CDA.ISCR}}$), as

[57] Empatica image taken from https://i.pinimg.com/originals/04/82/54/0482548efb5dfca4394c56610952800a.jpg.

[58] http://www.ledalab.de/

well as the maximum value of phasic activity ($\mathrm{Leda}_{\mathrm{CDA.PhasMax}}$) and the mean tonic activity ($\mathrm{Leda}_{\mathrm{CDA.Ton}}$) features are created by the Ledalab software.

The Empatica and Garmin devices also measure the **skin temperature**, which we use as a feature ($\mathrm{SkinTemp}$).

7.6 Data Normalization & Segment-Wise Feature Calculation

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

By *single features* we mean features that yield only one value per segment: This class comprises subjective measures (SubjCL), time measures (PeTime, $\mathrm{LNPeTime}$), text measures (BLEU, HBLEU, TER, HTER, SL), keyboard measures (APR, PWR), the amount-based eye features ($\mathrm{BlinkAmount}$, $\mathrm{FixAmount}$, $\mathrm{NormBlinkAmount}$, $\mathrm{NormFixAmount}$), 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.

Apart from these single 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 ($\mathrm{max} - \mathrm{min}$), which is comparable to many related works, e.g., Borys et al. (2017) and Ishimaru et al. (2017). As an example, this means that GSR, actually consists of the 6 features $\mathrm{GSR}_{\mathrm{acc}}$, $\mathrm{GSR}_{\mathrm{avg}}$, $\mathrm{GSR}_{\mathrm{std}}$, $\mathrm{GSR}_{\mathrm{min}}$, $\mathrm{GSR}_{\mathrm{max}}$, and $\mathrm{GSR}_{\mathrm{range}}$.

7.7 Conclusion

While the last chapter focused on the vision of a cognition-aware translation environment, this chapter presented our multi-modal data capturing framework which combines a variety of subjective, performance, behavioral, and especially physiological features. It thus builds the basis for investigating which sensor modalities work better or worse for estimating CL during PE of MT, and therefore forms a necessary step towards achieving this vision. The next chapter will use this framework in three studies to gain practical insights.

Chapter 8
Cognitive Load Estimation Experiments

Using the previously defined framework, we overall conducted 3 studies. Section 8.1 presents a PE study with translation Master's students using only parts of the framework. Section 8.2 then shows a follow-up study with the full framework and professional translators instead of translation students. Finally, section 8.3 explores the same framework with university students consuming videos and answering questions in e-learning, to explore whether the same approach also yields robust CL estimates in a different domain.

Section 8.1 is based upon publications Herbig et al. (2019a) and Herbig et al. (2019c), section 8.2 is based on publication Herbig et al. (2021), and section 8.3 is based on publication Herbig et al. (2020a).

8.1 Multi-Modal Cognitive Load Estimation with Translation Students

This first study explores large parts of the previously presented framework to capture physiological, behavioral, and performance measures from translation students while PE MT. Even though it does not involve the whole framework, several of the tested features have not previously been explored in the translation domain and especially PE domain. With the captured data, we investigate how well predictive models based on feature combinations from these modalities can predict perceived CL, as measured by our subjective rating scale (see section 7.2). The different modalities and their combinations are then compared in terms of regression performance. Finally, we analyze correlations between the best performing features and the corresponding subjective ratings to better understand what benefits a multi-modal approach has. The results of our analyses indicate that combining multiple modalities helps in detecting CL.

Compared to the literature, the works by Vieira (2014, 2016) presented in subsubsection 2.4.3 are probably the most closely related studies. However, 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.

We will first present which of the features from the framework were used in this study, and then outline the MT system used to generate text for the experiment. After that, we will discuss the experiment to capture data in more detail, before we present the corresponding data analysis. Finally, we will present limitations to the approach and discuss the main take-aways.

8.1.1 Analyzed Measures of Cognitive Load

Among the CL measures presented in chapter 7, we use SubjCL, so the subjective CL rating provided after editing each segment as the core measure to which we link all other measures. The goal is thus to see how well SubjCL can be assessed by other means, because introspection and reporting of CL was shown to be sensitive to small differences and reliable (Paas et al., 2003), but unable to capture real-time changes in CL (Moissa et al., 2019).

All measures that we compare against SubjCL and against each other can be categorized as follows: (1) *time*-based measures, (2) *text*-based measures, (3) *sensor*-based measures consisting of *typing-*, *eye-*, *heart-*, *skin-*, *emotion-*, and *body posture*-based measures. Finally, we explore (4) *combinations* of the previous three.

The *time* measures are the post-editing time (PeTime) or the length-normalized post-editing time ($\mathrm{LNPeTime}$); the *text* features consist of BLEU, HBLEU (Lin and Och, 2004), TER, HTER (Snover et al., 2009), and sentence length (SL), as well as all combinations thereof.

The *typing*-based measures comprise APR and PWR.

Using a Tobii 4C eye tracker, we further capture the *eye*-based measures $\mathrm{BlinkAmount}$, $\mathrm{FixAmount}$, $\mathrm{SaccDur}_{\mathrm{avg}}$, and $\mathrm{SearchProb}_{\mathrm{avg}}$. All of these features also work without the expensive Pro SDK, thus, making real-world usage more realistic. Furthermore, we use a web-cam to capture the eye openness ($\mathrm{EAR}_{\mathrm{avg}}$).

For *heart*-based measures, we integrate a Polar H7 heart belt communicating with the computer via Bluetooth Low Energy. From its raw RR data, we capture $\mathrm{RMSSD}^{\mathrm{Polar}}_{\mathrm{avg}}$ and $\mathrm{SDNN}^{\mathrm{Polar}}_{\mathrm{avg}}$, thus, for each participant the normalized RMSSD and SDNN signal averaged across segment.

For *skin*-based measures, we include the Microsoft Band v2, a small bracelet offering a galvanic skin response sensor. In terms of CL measures, we again normalize by participant and capture $\mathrm{GSR}_{\mathrm{avg}}^{\mathrm{MSBand}}$, $\mathrm{GSR}_{\mathrm{acc}}^{\mathrm{MSBand}}$, and the frequency domain feature $\mathrm{FreqGSR}_{\mathrm{avg}}^{\mathrm{MSBand}}$.

Using a web-cam, we further capture *emotion* measures from facial expressions ($\mathrm{EmotionName}$).

Last, a Kinect v2 captures the *body posture*, from which we extract the head distance ($\mathrm{HeadDist}_{\mathrm{avg}}$).

For the *sensor* features, 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.

8.1.2 Machine Translation System

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 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) using BPE. We followed the best hyperparameter 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. Even though MT quality has improved further since this experiment, the used system can still be considered high-quality and thus results remain relevant.

8.1.3 Data Capturing 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. The study was approved by the university's ethical review board and the data protection officer. All data used throughout this experiment is publicly available at `http://mmpe.dfki.de/data/MTJournal2019`.

Text Selection

Similar to Vieira (2016), we used a subset of the WMT 2017 news translation task test set as text for this study. After using our NMT system, we extracted 300 sentences and their translations, 100 each within the TER intervals [35-50], [60-70], and [80-95]. 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.

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. 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 segments 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 CL 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 with respect to the data 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 was 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.

Apparatus

For the study, the post-editor is equipped with a Microsoft Band v2 on her right wrist, the Polar H7 heart belt on her chest, and the Tobii 4C 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 SDL Trados Studio for this study as it is by far the most used CAT tool in professional applications (see section 2.3.2).

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. 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.

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 8.1 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).

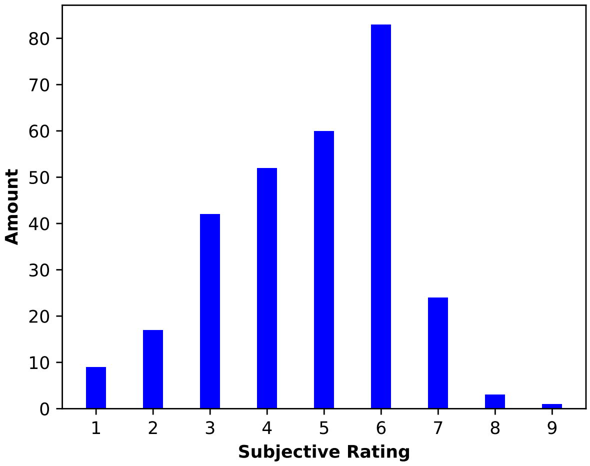


Figure 8.1: Segment distribution across subjective CL scale ranging from 1 = “very very low mental effort” to 9 = “very very high mental effort”.

8.1.4 Data Analysis Overview

Based on the 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 (see subsection 8.1.5). 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 (see subsection 8.1.6).

For both analyses, we use the above (see section 8.1.1) 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.

8.1.5 Multi-Modal Cognitive Load Detection – Regression Analysis

Method

Problem Framing 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 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. The reason why we focus on subjectively assessed CL is that it is good at capturing inter-translator differences in contrast to any general measure of quality. 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.; hence, the translations are not objectively hard or easy.

Baselines 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 ($\mathrm{SubjCL}_{\mathrm{avg}}$), and (2), always predicting the median subjective rating ($\mathrm{SubjCL}_{\mathrm{median}}$).

Explored Models 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[59] 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 hyperparameter 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).

Analysis Details 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*

[59] `http://scikit-learn.org/`

sets, we ensure that at least one feature of *time* or *text* is combined with one or multiple *sensor*-based features.

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 precisely 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.

Results

General The regression results are presented in Tables 8.1 and 8.2. The results are 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).

Time and Text Features For *time* features, we notice that PeTime performs better than its length-normalized alternative ($\mathrm{LNPeTime}$), 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.

Sensor 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.

Combined Features 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.

Pairwise Comparison 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 8.3 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 statistically significant differences.

Discussion

Although the concrete ratings differ between post-editors, the methods to measure CL, especially the multi-modal ones, are apparently transferable.

Time and Text Features 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.

	MSE	
Features	**1x10-CV↓(Reg.)**	**5x2-CV↓ (SD)**
Baselines		
$\mathrm{SubjCL}_{\mathrm{avg}}$	2.466 (-)	2.465 (0.093)††
$\mathrm{SubjCL}_{\mathrm{median}}$	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		
$\mathrm{RMSSD}^{\mathrm{Polar}}_{\mathrm{avg}}$	2.285 (Ridge)	2.282 (0.054)**††
$\mathrm{SDNN}^{\mathrm{Polar}}_{\mathrm{avg}}$	2.352 (Ridge)	2.379 (0.078)**††
$\mathrm{RMSSD}^{\mathrm{Polar}}_{\mathrm{avg}}$, $\mathrm{SDNN}^{\mathrm{Polar}}_{\mathrm{avg}}$	2.304 (SVR)	2.309 (0.057)**††
Eyes		
BlinkAmount	2.034 (Ridge)	2.040 (0.062)**††
FixAmount	2.276 (SVR)	2.292 (0.131)**††
$\mathrm{SaccDur}_{\mathrm{avg}}$	2.415 (Lasso)	2.421 (0.122)††
$\mathrm{SearchProb}_{\mathrm{avg}}$	2.462 (Lasso)	2.247 (0.094)††
$\mathrm{EAR}_{\mathrm{avg}}$	2.424 (Ridge)	2.438 (0.093)**††
BlinkAmount, FixAmount, $\mathrm{SearchProb}_{\mathrm{avg}}$, $\mathrm{EAR}_{\mathrm{avg}}$	1.704 (RF)	1.803 (0.175)**††
Skin		
$\mathrm{GSR}^{\mathrm{MSBand}}_{\mathrm{avg}}$	2.462 (Lasso)	2.461 (0.093)††
$\mathrm{GSR}^{\mathrm{MSBand}}_{\mathrm{acc}}$	2.181 (Lasso)	2.185 (0.041)**††
$\mathrm{FreqGSR}^{\mathrm{MSBand}}_{\mathrm{avg}}$	2.402 (Ridge)	2.383 (0.082)*††
$\mathrm{GSR}^{\mathrm{MSBand}}_{\mathrm{avg}}$, $\mathrm{GSR}^{\mathrm{MSBand}}_{\mathrm{acc}}$, $\mathrm{FreqGSR}^{\mathrm{MSBand}}_{\mathrm{avg}}$	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		
$\mathrm{HeadDist}_{\mathrm{avg}}$	2.445 (SGD)	2.460 (0.095)**††
Emotions		
$\mathrm{Anger}_{\mathrm{avg}}$	2.430 (SGD)	2.445 (0.089)**††
$\mathrm{Anger}_{\mathrm{avg}}$, $\mathrm{Neutral}_{\mathrm{avg}}$, $\mathrm{Sadness}_{\mathrm{avg}}$, $\mathrm{Surprise}_{\mathrm{avg}}$	2.383 (RF)	2.420 (0.101)**††

Table 8.1: Feature evaluation results (1) 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 $\mathrm{SubjCL}_{\mathrm{avg}}$, while a dagger (†) indicates a significant difference to $\mathrm{SubjCL}_{\mathrm{median}}$. */† represent $p < 0.05$, **/†† represent $p < 0.01$ after Bonferroni correction.

	MSE	
Features	**1x10-CV↓(Reg.)**	**5x2-CV↓ (SD)**
Combined Sensors		
Keyboard (TER) Eye (BlinkAmount, FixAmount, SaccDur$_{\text{avg}}$, EAR$_{\text{avg}}$) Skin (GSR$^{\text{MSBand}}_{\text{acc}}$, GSR$^{\text{MSBand}}_{\text{avg}}$, FreqGSR$^{\text{MSBand}}_{\text{avg}}$) Heart (SDNN$^{\text{Polar}}_{\text{avg}}$, RMSSD$^{\text{Polar}}_{\text{avg}}$)	1.512 (RF)	1.639 (0.153)**††
Eye (BlinkAmount, FixAmount, EAR$_{\text{avg}}$) Skin (GSR$^{\text{MSBand}}_{\text{acc}}$, GSR$^{\text{MSBand}}_{\text{avg}}$)	1.595 (RF)	1.646 (0.115)**††
Combined Features		
Time (PeTime) Keyboard (APR, PWR) Eye (BlinkAmount, FixAmount, EAR$_{\text{avg}}$, SaccDur$_{\text{avg}}$, SearchProb$_{\text{avg}}$) Skin (FreqGSR$^{\text{MSBand}}_{\text{avg}}$) Heart (RMSSD$^{\text{Polar}}_{\text{avg}}$)	1.434 (Ridge)	1.487 (0.069)**††
Time (PeTime) Skin (FreqGSR$^{\text{MSBand}}_{\text{avg}}$) Eye (BlinkAmount, FixAmount) Heart (RMSSD$^{\text{Polar}}_{\text{avg}}$)	1.490 (Ridge)	1.508 (0.084)**††

Table 8.2: Feature evaluation results (2) 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 $\text{SubjCL}_{\text{avg}}$, while a dagger (†) indicates a significant difference to $\text{SubjCL}_{\text{median}}$. $*/^\dagger$ represent $p < 0.05$, $**/^{\dagger\dagger}$ represent $p < 0.01$ after Bonferroni correction.

Sensor Features 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$_{\text{avg}}$).

Combinations of GSR-based *skin* 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 as a *body posture* feature 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 diminishingly 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

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

Table 8.3: 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).

other features and the gains 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.

Combined Features 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 (see Table 8.3), 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.

Summary 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, contrary 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 hyperparameter 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 MSEs in the future.

8.1.6 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.

Results

Figure 8.2 shows violin plots of the individual feature values plotted against the subjective CL ratings provided by each participant. 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 8.4, 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 $\mathrm{RMSSD}_{\mathrm{avg}}^{\mathrm{Polar}}$ (-0.220) that is weak to moderate. For the eye features $\mathrm{BlinkAmount}$, $\mathrm{FixAmount}$, and $\mathrm{SaccDur}_{\mathrm{avg}}$, 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 complex 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.

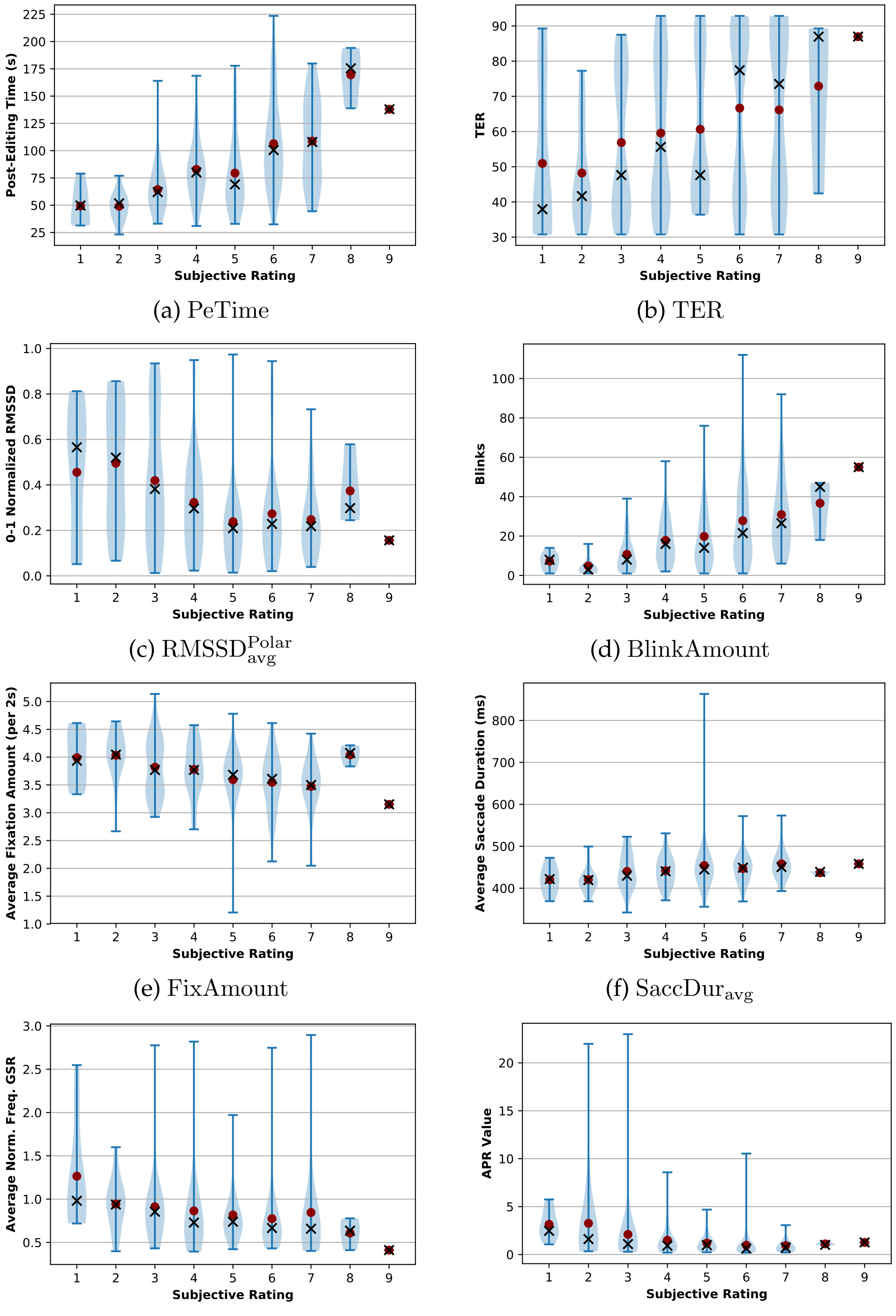
Feature	**Spearman's ρ**	**Interpretation**	**p-value**
PeTime	$+0.505$	Strong	< 0.001
TER	$+0.276$	Moderate	< 0.001
$\mathrm{RMSSD}_{\mathrm{avg}}^{\mathrm{Polar}}$	-0.220	Weak to moderate	< 0.001
$\mathrm{BlinkAmount}$	$+0.453$	Strong	< 0.001
$\mathrm{FixAmount}$	-0.262	Moderate	< 0.001
$\mathrm{SaccDur}_{\mathrm{avg}}$	$+0.193$	Weak to moderate	< 0.001
$\mathrm{FreqGSR}_{\mathrm{avg}}^{\mathrm{MSBand}}$	-0.264	Moderate	< 0.001
APR	-0.308	Moderate	< 0.001

Table 8.4: Spearman's correlation results between different features and subjective CL ratings.

Discussion

These results show why multi-modality helps: Apart from PeTime and $\mathrm{BlinkAmount}$, 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 (see Tables 8.1 and 8.2), 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 smartwatches 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



(a) PeTime

(b) TER

(c) $\text{RMSSD}^{\text{Polar}}_{\text{avg}}$

(d) BlinkAmount

(e) FixAmount

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

(g) $\text{FreqGSR}^{\text{MSBand}}_{\text{avg}}$

(h) APR

Figure 8.2: Violin plots for different feature values per subjective rating (x-axis). The circles indicate the means, the crosses the medians.

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.

8.1.7 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.

8.1.8 Conclusion

This first study using parts of the multi-modal CL estimation framework focused on predicting and correlating perceived cognitive PE effort. 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.

Regarding our second overall research question, we showed that multi-modal sensor input indeed helps estimate CL during PE of MT. The study thus contributes to our overall goal: decreasing the perceived CL, and thereby stress and exhaustion, during PE, e.g., by optimizing MT systems on the user's CL measurements to produce less demanding outputs or by intervening in the PE process within CAT tools. Naturally, the next steps involve using larger parts of the framework, and running a similar study with professional translators, which is the focus of the next section. This second study will also apply data filtering approaches and feature selection approaches to make better use of the available data than the regression approach chosen for this initial investigation.

8.2 Multi-Modal Cognitive Load Estimation with Professional Translators

The previous study has shown that automatically estimating perceived CL based on a variety of sensors is feasible, and that a multi-modal approach facilitates this. However, the study only utilized a small subset of our measurement framework to which the presented results are limited. Furthermore, the target users were translation Master's students and not professionals. In this study, we thus use an even wider 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. We again 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 and Van Merriënboer, 1994). The different modalities and their combinations are again compared in terms of regression performance. Furthermore, and 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, skin, as well as combined measures perform very well on their own, 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, which should be sufficient for most application scenarios discussed in chapter 6.

As we did in our previous study with translation students, we begin by clearly specifying the features used, then present the methodology, and finally go over to the results and discussion, which consists of the two main parts: (i) multi-modal regression analysis and (ii) pairwise correlations and PCA.

8.2.1 Analyzed Measures of Cognitive Load

Compared to Vieira (2016), our last study with translation students already increased the amount of analyzed features significantly by adding heart-, skin-, and camera-based features. For this study, we add even more and higher quality sensors and add further high-level features.

As in the last study, we show a pop-up asking for SubjCL after each segment using the same scale as before (Paas and Van Merriënboer, 1994). These subjective ratings are simultaneously the target values when training the regression models.

For the *time features* we again integrate PE time (PeTime) and length-normalized PE time which also considers the segment length (LNPeTime).

Identical to our last study, *text features* consist of smoothed BLEU, HBLEU (Lin and Och, 2004), TER, HTER (Snover et al., 2009), and sentence length (SL).

As *typing features* we again capture the higher-level pause features APR and PWR by Lacruz et al. (2012), which were shown to correlate with PE effort.

For *eye-based features*, the same hardware as before (a web-cam and the Tobii 4C eye tracker) is used to capture the eye aspect ratio (EAR, Soukupova and Cech (2016)), the amount of blinking (of less than 2 s length) (BlinkAmount), its normalization by the PE time (NormBlinkAmount, Van Orden et al. (2001)), and similarly the amount of fixations (FixAmount) and its time-normalized version (NormFixAmount). Furthermore, fixation durations (FixDur) and saccade durations (SaccDur) (Doherty et al., 2010; Moorkens et al., 2015) as well as the probability of visual search based on the eye movements (SearchProb, Goldberg and Kotval (1999)) are added. As a main distinction from our last study, we also capture the pupil diameter (PupilDiameter, O'Brien (2006a)) which we get from the Tobii Pro SDK. As discussed in chapter 7, we calculate the higher level features Index of Cognitive Activity (ICA) with two approaches: one using a wavelet transformation to calculate the amount of rapid dilations ($\mathrm{ICA}^{\mathrm{wave}}$), 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 ($\mathrm{ICA}^{\mathrm{count}}$). We also check for sharp changes and continuations of the ramp in the Hilbert unwrapped phase of the pupil diameter signal (Hilbert, Hossain and Yeasin (2014)).

For *heart measures*, we integrate the Polar H7 belt and the Garmin Forerunner 935 watch to capture the heart rate (HR), and use the Polar and the Empatica E4 wristband to capture the RR interval (RR) and all our Heart Rate Variability (HRV) (Rowe et al., 1998) measures (RMSSD, SDNN, NN50, and pNN50, Shaffer and Ginsberg (2017)). Furthermore, the Empatica captures the blood volume pulse from which we capture a variety of features (BVP, BVPAmp,BVPMedAbsDev,BVPMeanAbsDiff as discussed in chapter 7). The main difference compared to our last study 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 use the Microsoft Band v2 and Empatica to capture the GSR and frequency domain features ($\mathrm{FreqGSR}$,$\mathrm{FreqFrameGSR}$) as described in chapter 7. Furthermore, we now use the Ledalab software to calculate the "global" features ($\mathrm{Leda}_{\mathrm{avg}}$, $\mathrm{Leda}_{\mathrm{MaxDefl}}$), and "through-to-peak (TTP)/min-max" features ($\mathrm{Leda}_{\mathrm{TTP.nSCR}}$, $\mathrm{Leda}_{\mathrm{TTP.AmpSum}}$,$\mathrm{Leda}_{\mathrm{TTP.Lat}}$), as well as features based on Continuous Decomposition Analysis (CDA) (Benedek and Kaernbach, 2010), separating skin conductance data into continuous signals of tonic (background) and phasic (rapid) activity ($\mathrm{Leda}_{\mathrm{CDA.nSCR}}$, $\mathrm{Leda}_{\mathrm{CDA.AmpSum}}$, $\mathrm{Leda}_{\mathrm{CDA.Lat}}$,$\mathrm{Leda}_{\mathrm{CDA.SCR}}$, $\mathrm{Leda}_{\mathrm{CDA.ISCR}}$, $\mathrm{Leda}_{\mathrm{CDA.PhasMax}}$,$\mathrm{Leda}_{\mathrm{CDA.Ton}}$). Finally, the Empatica and Garmin devices also measure the skin temperature, which we use as a feature ($\mathrm{SkinTemp}$). The differences from our last study for the skin features are as follows: We further use the skin resistance data delivered by the Empatica E4 wristband, on which we calculate the same features as before, but additionally add the Ledalab features. Furthermore, we integrate the skin temperature features.

Emotions are not used in this study, as their results did not yield much performance in the last study, and the approach of sending the data to an emotion recognition API led to privacy concerns among some participants.

As *body posture* features, we again use the distance to the head and normalize it per participant ($\mathrm{HeadDist}$).

In terms of data normalization, this time we use the whole approach described in chapter 7, which means keeping the *global features* and calculating the 6 very simple features from the normalized signal of the *continuous features*: the accumulated, average, standard deviation, minimum, maximum, and range ($\max - \min$). Note that the previous study focused solely on the per-participant normalized average value per segment, and did not include the other 5 features.

Furthermore, we applied a more strict data filtering approach: For this, 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 100000 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 $\mathrm{HR}^{\mathrm{Polar}}$ and $\mathrm{RR}^{\mathrm{Polar}}$ samples which fall outside the acceptable 50–120 beats per minute or 500–1200 ms ranges were ignored (Shaffer and Ginsberg, 2017).

8.2.2 Method

To explore this enhanced feature set for capturing CL in PE, we perform a user study that was again approved by the university's ethical review board. We first describe the text selection and apparatus, and then present our data analysis approach, which is an extension to the approach used for the previous study.

Text Selection and Apparatus

As text, we used the same 30 sentences as in the last study (see subsubsection 8.1.3) which participants see within SDL Trados Studio in randomized order 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 widely-used SDL Trados Studio 2017 translation environment.

Data Analysis Approach

We analyze the gathered data in a multi-step approach: Similar to the previous study, we first analyze the subjective ratings provided by our participants, and then estimate the subjective CL ratings 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 and how they cluster together. One should note here that we analyze many more features than Vieira (2016), so we aim to both reproduce and extend his findings. The goal of the regression analysis is to be able 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 insights into why some measuring approaches perform well while others contribute little information.

For all analyses, we discuss the features in terms of the feature sets described in Section 8.2.1: *subjective*, *time*, *text*, *keyboard*, *body posture*, *heart*, *eye*, and *skin* features. Finally, we also investigate *combinations* of these sets.

Subjective Ratings & Sensor Correlations 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. We further report correlations between measures produced by different sensors, e.g., how similar is data captured by a heart belt compared to wrist-based heart measurements, that would be easily applicable in practice?

Multi-Modal CL Regression Analysis This part is closely related to our previous study with CL students (see section 8.1), having the goal of automatically gathering CL values for segments through different sensors. This is again achieved by learning a function that fits our features to the subjective CL ratings in range 1 to 9, which also captures inter-translator differences. 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 result in different physiological responses. Thus, instead of potentially biasing our data by transforming the target variable, we keep it as is. However, compared to our last analysis we now also perform a comparison between models with a random effect for participant and segment 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 in chapter 6, quality and CL cannot be considered equal, and time could be traded off for quality, thereby limiting findings based solely on these measures.

As before, we compare the different regression models based on different feature sets against each other and to the a simple baseline always predicting the mean subjective rating ($\mathrm{SubjCL_{avg}}$).

What is new, is that we compare two approaches for training regression models.

The first approach is almost identical to our last experiment, as it 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. We again explore a variety of models that can represent different function spaces that might be needed for the different feature sets, namely linear models with different regularizers (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 set is very similar to our last study, except that we exclude the Support Vector Regression (SVR) model[60].

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[61]. 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 the same value, 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% when loaded. By incorporating the participant and segment as a feature into the models, we ensure

[60] Since SVRs do 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.

[61] 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.

that they can learn such individual difference. This is also a major distinction from our last experiment, where we did not incorporate these measures. However, this approach of training the models is only relevant for strictly controlled experiments, because in practice no two translators will PE the same segment.

To avoid over-fitting, all regression functions use regularization or averaging, and we perform 10 by 1 and 5 by 2-fold stratified cross-validation (CV) as in our last experiment. The stratification is important here as it is better suited for an imbalanced distribution of the target variable (that we happen to have, see Section 8.2.3). Naturally, every regression model is trained on the same folds, to make results comparable. We kept MSE as our main metric, as it strongly penalizes large errors, which are particularly undesirable for our goal. 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 the `scikit-learn`) to decide on the amount of features to select.

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. Furthermore, the above regression naturally is only interested in correlations to subjective CL, which we selected as our target variable, and is not concerned with correlations between different features themselves.

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 large number of features.

To keep the reporting concise, we only report PCs that together explain 95% of the variance, i.e., components capturing less then 5% variance are not plotted. 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 regression model 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.

Participants and User Evaluation Procedure

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

Identical to the experiment with translation students, after the data protection form and demographics questionnaire, the professional translators were given the same instruction how to PE to ensure similar editing behavior across participants. 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 and then they 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 amount of data gathered for this participant.

8.2.3 Results & Discussion

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

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 8.3). We also observe rating differences between post-editors, with an average standard deviation across segments of 1.2 (minimum 0.8, maximum 1.8) on our 9-point scale. When asked after the experiment how demanding it was overall, i.e., a retrospective overall judgement, the scores are very much in line with the ratings given immediately after each segment, with a mean of 5.4 and a standard deviation of 0.8. In general, the rating distribution and the inter-rater differences are strongly comparable to those of our previous study, were the extreme cases were rarely chosen. Note that we use these individual CL ratings (without any aggregation on segment level) for the remaining analyses to also capture inter-participant differences. Furthermore, 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 again shows that strong differences in editing behavior do not necessarily impact the CL.

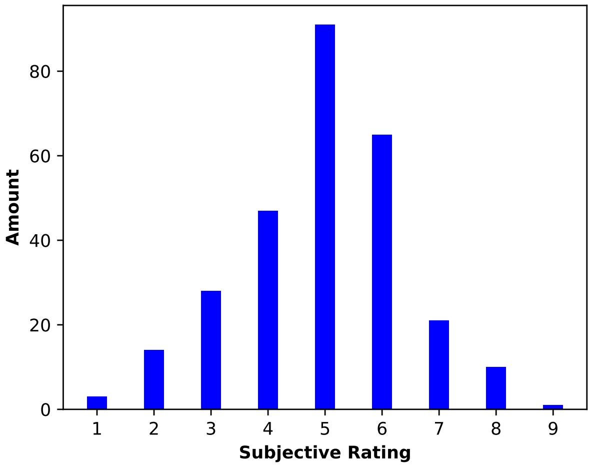


Figure 8.3: Segment distribution across subjective CL scale ranging from 1 = “very very low mental effort” to 9 = “very very high mental effort”.

Sensor Correlations

We further investigated correlations between measures produced by different sensors. The results can be seen in Table 8.5: While the heart rate and temperature measures indeed correlate strongly, they still do not match exactly. We expected this, since measuring at different positions and using different sensors will never give exactly the same results. What we did not expect, however, is that the RR intervals by the Empatica and Polar belt are only correlated to a medium extent. We do assume the Empatica data to be of high quality as it satisfies several regulatory compliances and was designed as a research device. However, since the Polar device is worn at the chest, we had expected it to produce rather accurate measures as well even though it is only a sports device.

Feature	Spearm.’s ρ	Pearson’s r	Interpret.
$\mathrm{RR}^{\mathrm{Empatica}}$ vs. $\mathrm{RR}^{\mathrm{Polar}}$	0.38	0.32	Medium
$\mathrm{HR}^{\mathrm{Garmin}}$ vs. $\mathrm{HR}^{\mathrm{Polar}}$	0.81	0.71	Strong
$\mathrm{SkinTemp}^{\mathrm{Garmin}}$ vs. $\mathrm{SkinTemp}^{\mathrm{Empatica}}$	0.83	0.74	Strong

Table 8.5: Spearman’s correlation results between different features and subjective CL ratings, all $p < 0.01$.

Multi-Modal CL Regression Analysis

Without Including the Participant and Segment ID The results of the first regression analysis approach, that is without passing the participant and segment ID alongside the features to the model, are reported in Table 8.6. 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, as this is one (among many) measures of what this feature contributes. 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 8.6 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 8.6 shows that all models are significantly better than the mean baseline (after Bonferroni correction).

When looking at the individual results in Table 8.6, 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.

Including the Participant and Segment IDs Table 8.7 shows how the results change when including LMEMs to the list of potential regressors and adding the participant and segment ID as additional features to the other regression models, both during training and testing. This allows the trained models to react differently to different participants and segments, thereby losing generality but allowing better model fits. This time only LMEMs and Random Forest models were chosen, and again the 1 by 10 and 5 by 2-fold CV scores are roughly comparable. We again use a univariate ANOVA (including Greenhouse-Geisser correction due to a violation of sphericity) and find that all models are significantly better than the baseline (after Bonferroni correction).

Comparison When comparing the results of Table 8.7 to Table 8.6, we see that the results with participant and segment improved substantially for the *time*, *text*,

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

Table 8.6: Feature evaluation results **without considering LMEMs/without adding participant and segment**. For 10 by 1 and 5 by 2-fold CV with standard deviation (SD). Asterisk (*) in the right column indicates a significant difference ($p < 0.01$) from $\text{SubjCL}_{\text{avg}}$ after Bonferroni correction.

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

Table 8.7: Feature evaluation results when **considering LMEMs/adding participant and segment**. For 10 by 1 and 5 by 2-fold CV with standard deviation (SD). Asterisk (*) in the right column indicates a significant difference ($p < 0.01$) from $\mathrm{SubjCL_{avg}}$ after Bonferroni correction.

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 8.8. Note that these results are using the models without incorporating participant and segment (Table 8.6), 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.

Features	**Test Statistics**
Time vs. Combined	$\tilde{t} = -4.06$*
Text vs. Combined	$\tilde{t} = -6.03$*
Keyboard vs. Combined	$\tilde{t} = -5.35$*
Body Posture vs. Combined	$\tilde{t} = -6.32$*
Eyes vs. Combined	$\tilde{t} = -0.98$
Heart vs. Combined	$\tilde{t} = -1.42$
Skin vs. Combined	$\tilde{t} = -1.34$

Table 8.8: Pairwise comparisons between the feature selected models without LMEM/without participant and segment (Table 8.6). * shows significance with $p < 0.05$ after Bonferroni-Holm correction. $\tilde{t}$ is the test statistics for the modified paired t-test (Dietterich, 1998).

Summary Summarizing, Tables 8.6 and 8.8 suggest that CL measurement without special adaptations per participant and segment (Table 8.6) 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 8.7), 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.

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 8.4, 8.5 and 8.6 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 and color shows the strength of the loading; blue continuous lines represent positive loadings, while red dashed lines indicate negative loadings. We only summarize the most interesting results, which are all statistically significant:

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

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

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

As expected, the most relevant *eye features* $\mathrm{FixAmount}$, $\mathrm{SaccDur}_{\mathrm{acc}}$, and $\mathrm{FixDur}_{\mathrm{acc}}$ correlate by almost 1, load on the same PC, and strongly relate to SubjCL.

For the *heart features*, the correlations between $\mathrm{NN50}^{\mathrm{polar}}_{\mathrm{acc}}$, $\mathrm{SDNN}^{\mathrm{polar}}_{\mathrm{acc}}$, and $\mathrm{RMSSD}^{\mathrm{polar}}_{\mathrm{acc}}$ are again very close to 1, and the PCA plot nicely visualizes that they cluster together. $\mathrm{BVPMedAbsDev}$ shows the strongest correlation to SubjCL.

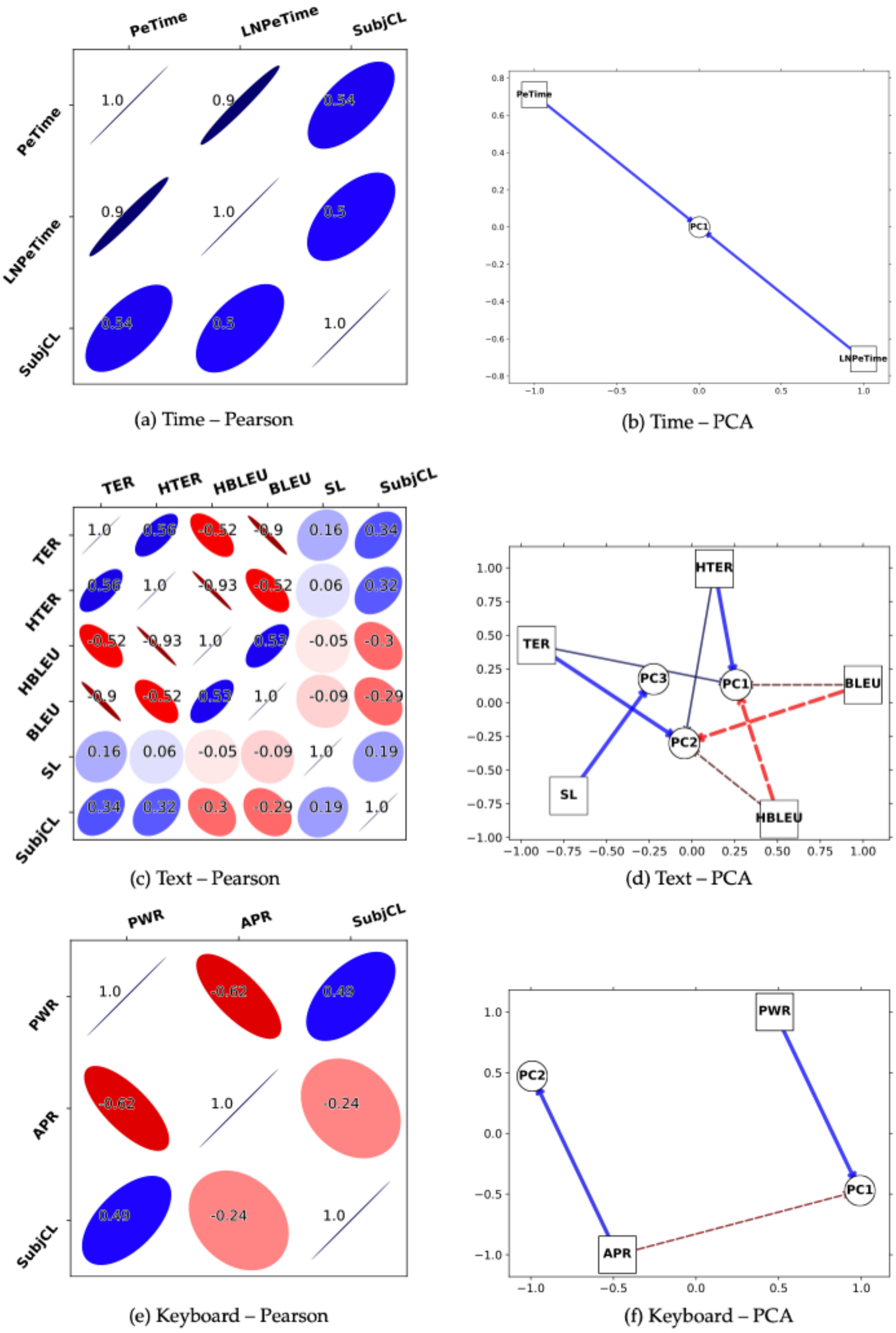
Inspecting the most relevant *skin features*, we see very strong correlations between $\mathrm{FreqFrameGSR}^{64,\mathrm{Empatica}}_{\mathrm{avg}}$ and $\mathrm{Leda}_{\mathrm{avg}}$, as well as medium to strong correlations between the frequency frame and $\mathrm{SkinTemp}^{\mathrm{Garmin}}_{\mathrm{acc}}$ features.

Most interestingly, for the *combined features* we can again see that $\mathrm{SDNN}^{\mathrm{polar}}_{\mathrm{acc}}$ and $\mathrm{NN50}^{\mathrm{polar}}_{\mathrm{acc}}$, as well as $\mathrm{FixAmount}$ and $\mathrm{SaccDur}_{\mathrm{acc}}$, correlate with almost a value of 1. There also seems to be a strong link between the HRV measures and the eye measures $\mathrm{SaccDur}_{\mathrm{acc}}$ and $\mathrm{FixAmount}$. The PCA further shows that there is one PC for the HRV measures, one for the ICA, and another one for the eye features $\mathrm{FixAmount}$ and $\mathrm{SaccDur}_{\mathrm{acc}}$.

Discussion

Overall, very good regression results of up to 0.7 MSE on a 9-point scale were achieved by our regression models. This amount of error should be acceptable for most applications scenarios discussed in chapter 6. While the 5 by 2-fold CV results are often slightly worse, which might be because less training data was seen, the results of 1 by 10 and 5 by 2-fold are comparable, and the very small standard deviations indicate that the models are rather robust.

Without Passing the Participant and Segment When comparing the regression results without adding participant and segment to our previous study, whose ap-



(a) Time – Pearson

(b) Time – PCA

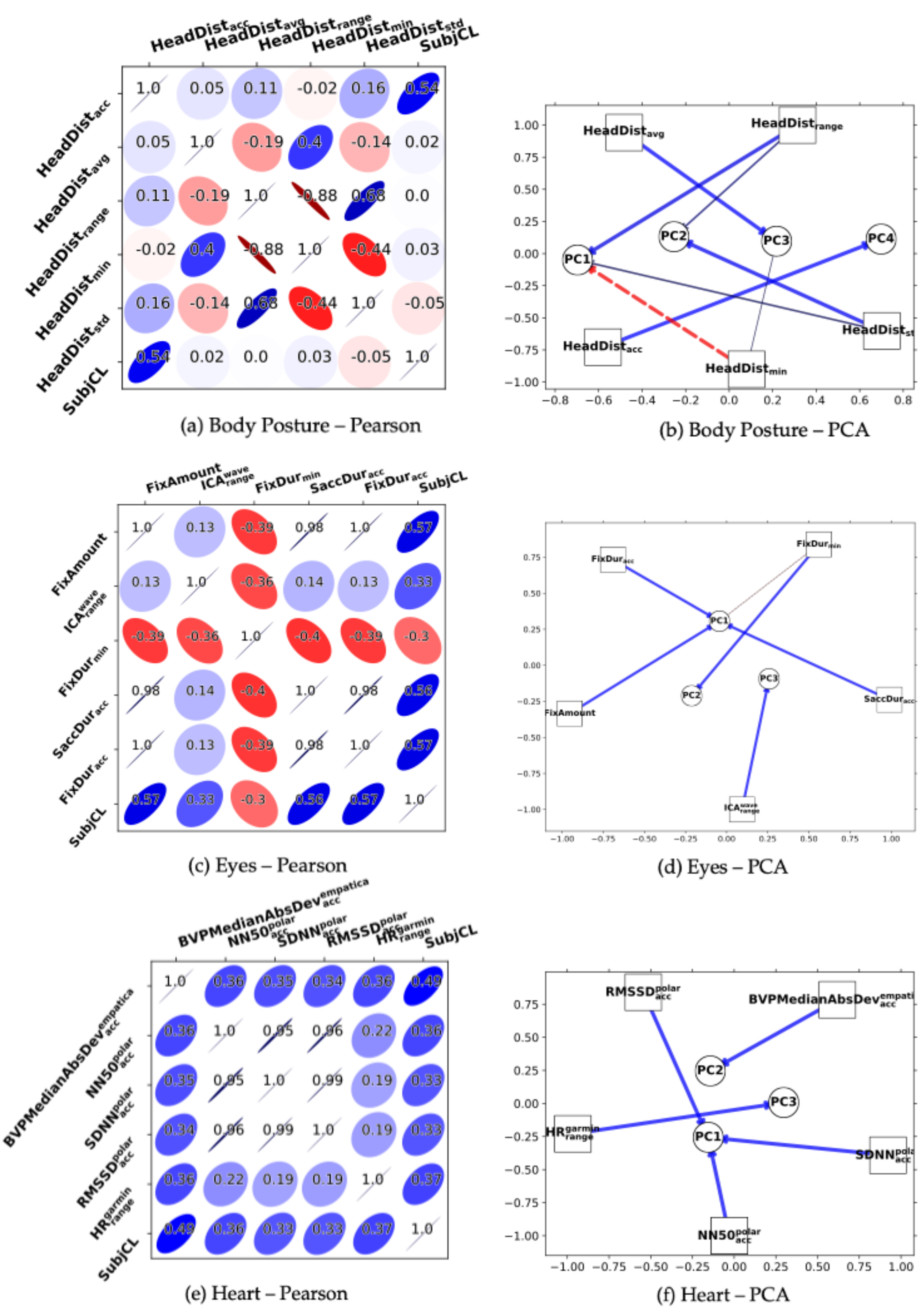
(c) Text – Pearson

(d) Text – PCA

(e) Keyboard – Pearson

(f) Keyboard – PCA

Figure 8.4: Correlations and PCA for *time*, *text*, and *keyboard* modalities.



(a) Body Posture – Pearson

(b) Body Posture – PCA

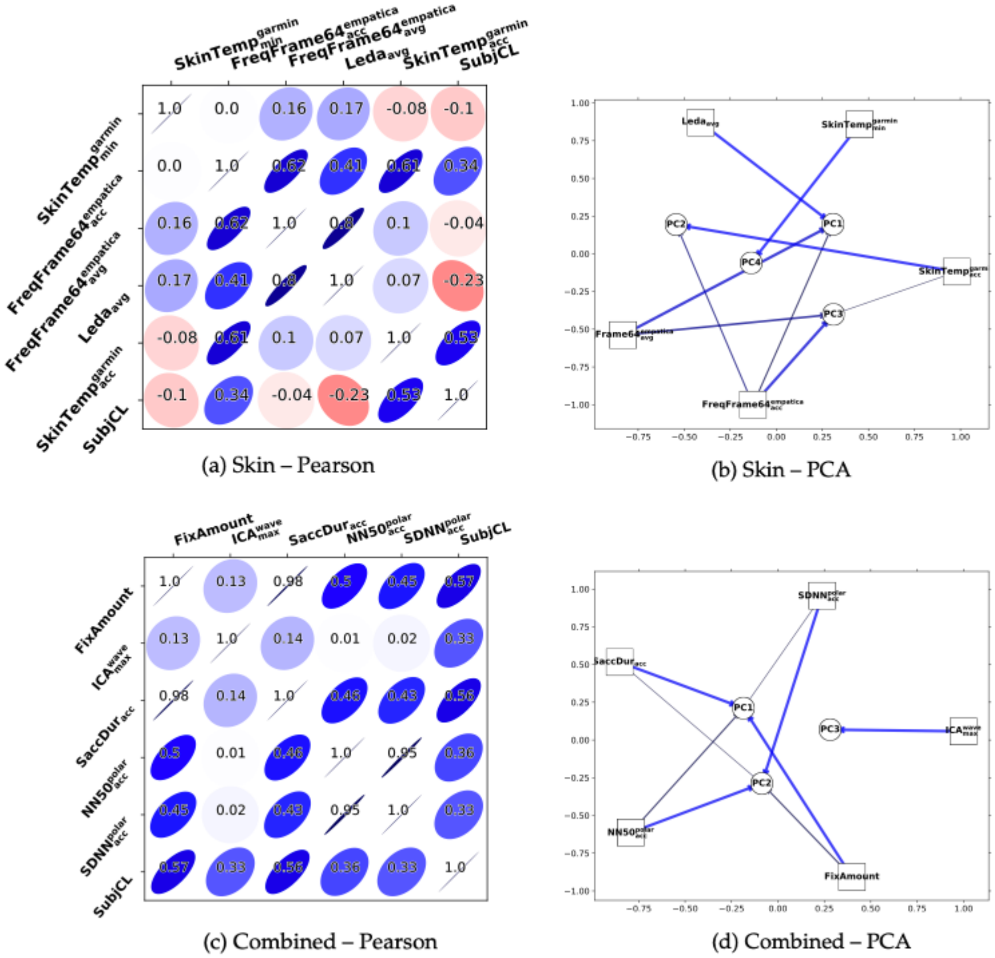
(c) Eyes – Pearson

(d) Eyes – PCA

(e) Heart – Pearson

(f) Heart – PCA

Figure 8.5: Correlations and PCA for *body posture*, *eye*, and *heart* modalities.



(a) Skin – Pearson

(b) Skin – PCA

(c) Combined – Pearson

(d) Combined – PCA

Figure 8.6: Correlations and PCA for the *skin* and *combined* modalities.

proach is almost the same apart from having fewer sensors and features, we note a few similarities and differences: First of all, in the new experiment, we found consistently better results across all modalities; however, already the baseline yields better results on our dataset, which indicates that the ratings of our professional participants are a bit more stable around the mean. While the *time* features in our last experiment were rather good, they are among the worst modalities here. A reason might be that we considered many more features, that helped the other modalities improve over the *time* feature. While in the last study the *eye* features were by far the best among the three main categories *eye*, *skin*, and *heart*, all three show similar results here. This could be due to the many newly added *skin* and *heart* features. Whereas in both studies the *combined* approach leads to the best results, the performance gains when combining multiple modalities were much stronger in our first experiment, probably again because the three main categories are already very good by themselves. That way, the *combined* approach was not significantly better than *eye*, *skin*, or *heart* based approaches.

So when we do not consider the individual participant and the segment s/he is post-editing (Table 8.6 or our last experiment), we can achieve the best results already only with our main categories, *eyes*, *heart*, *skin*, or by *combining* features from several modalities. This is relevant for less controlled and more practical applications, e.g., adapting the user interface to perceived CL, where it is impossible to use participant and segment information, as ideally no two translators should post-edit the same sentence (which would otherwise be contained in TM).

Passing the Participant and Segment In contrast, when we do consider participant and segment (Table 8.7), modalities of lesser quality, like *time*, *text*, *keyboard*, or *body posture* can also achieve good results. So considering *who is editing what* seems to yield enough information to learn from when combined with these features, while without considering participant and segment, the generalization is impeded. This is probably due to a strong overfit to the participant and segment, which simplifies the problem considerably. However, if the goal is to conduct a controlled experiment, e.g., to investigate the impact of different sentence features on subjectively felt CL, integrating participant and segment into the models allows to also achieve valuable estimates with these other modalities. The above experiment therefore also suggests that quality, keyboard, and time measures, which are frequently used in the literature to estimate effort, only work well in controlled settings.

Correlations and PCA While we cannot compare all our correlation and PCA results to Vieira (2016), since we considered many more features, there is still some interesting overlap: The *time* features in both studies correlated strongly to SubjCL. Furthermore, the link between the PWR and SubjCL also seems comparable, while that between APR and SubjCL appears weaker in our dataset. However, the correlation between these two *keyboard* features is similarly strong in both studies. The *eye* features FixAmount and FixDur also correlate to a similar extent with SubjCL in both studies. To summarize, we could both reproduce (except APR vs. SubjCL) and extend the findings by Vieira (2016).

The correlation and PCA especially revealed that many highly redundant features were selected by the feature selection approach (e.g., the HRV measures). The reason for this probably is their strong correlation to SubjCL; however, due to the redundancy, it is unclear whether incorporating multiple such features really helps. Therefore, we want to explore if handcrafting a set of features with fewer redundancies, or using a more sophisticated feature selection approach than RFECV, could boost performance further. As a simple example, the time features which strongly correlate to SubjCL and improved the results in our first study, were not selected for the *combined* model. While we only analyzed a few features in terms of correlations and PCA, we also plan to investigate the link to the non-selected features, as well as a PCA including more features from all different modalities than the few reported here.

Limitations

The results presented in this study are subject to the following limitations: Only 10 professional translators participated, resulting in a rather small dataset. Furthermore, we did not perform a leave-one-subject-out cross validation, thus, fine-tuning on new users might be required to achieve these results in practice. Last, as our multivariate analysis showed, a more guided feature selection approach, potentially combined with a thorough hyperparameter selection, might yield further improvements over the automatic "top-down" regression approach relying on feature selection.

8.2.4 Conclusion

This study was again motivated by the goal to robustly measure CL during PE and adapt to it. Compared to our last study, which already involved a large variety of measuring approaches, we explored even more features, comprising the largest set to date in the translation domain. The main novelties compared to the last experiment were the additional analysis of the *eye's* pupil diameter, as well as using the Garmin Forerunner 935 and the Empatica E4 wristband and adding further heart- and skin-related features.

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

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

This study thus contributed to our second research question by enhancing the findings from the previous study through results based on professional translators, a wider range of features, and more in-depth analysis. Overall, we still see that multi-modal sensor input helps estimate CL of PE, but that combining enough measures from either heart, skin, or eye modalities also allows good CL estimation. The long-term goal to which these results contribute remains the same: being able to decrease the perceived CL, and thereby stress and exhaustion, during PE through the various adaptations discussed in chapter 6, which require robust CL estimation during PE of MT.

8.3 Exploring the Cognitive Load Framework in E-Learning

We have now presented two studies using the CL framework to estimate perceived CL during PE. Naturally, the same framework can also easily be applied in other domains. To showcase this and to explore which measures are rather specific to PE and which are rather general purpose, we ran another study in the e-learning domain.

8.3.1 Why E-Learning?

The first and most obvious reason for choosing the e-learning domain as an additional test-bed is that CL theory stems from educational psychology, so the educational setting is probably the best researched area for CL measurements. However, we wanted to stick to a setting similar to PE, with a single user in front of a computing device, thus, choosing e-learning instead of learning in groups or classrooms. Furthermore, the e-learning industry is continuously growing, with a predicted compound annual growth rate of 7% until 2025 (Accuray Research LLP, 2017), making it a highly interesting market.

Modern e-learning systems already offer a variety of customization possibilities, including the possibility to work through the content in a self-chosen order or speed. Recommendation engines support the customization process, however, they often only consider the previous behavior of the current or other learners. This neglects the user's cognitive state, i.e., the cognitive load experienced, or factors like the perceived stress, tiredness, boredom, or attention, which we argue can strongly influence the content or speed that is appropriate for his/her current state (see e.g., Vogel and Schwabe (2016)). Those factors are also given appropriate consideration by human teachers in traditional learning, as they react to their students' needs and moods by asking follow-up questions or adding additional explanations. Similarly, a cognition-aware e-learning system could provide further clarifying contents when a high CL is detected, or decide to move on to more complex topics when the CL drops. Taking this to the e-learning domain, an application-oriented video showing a learned technique in practice might be well suited when a user feels overloaded, while a complex technical article could be overwhelming and therefore not effective in this situation. In contrast, the same video might feel boring in another cognitive state, where the technical article could be more appropriate. Furthermore, informing the instructor about the learners' cognitive states could help improve the learning content and tailor it to individual needs.

Thus, similar to CAT tools reacting to CL, a variety of adaptations of e-learning systems, aiming to keep the learner in the optimal range of CL (Spüler et al., 2016), would be possible if these systems could reliably estimate the cognitive state of a user. As discussed before, plenty of approaches to measure cognitive load, stress, etc. have been proposed in the literature and allow some form of cognition awareness (Bahreini et al., 2016; Rodrigues et al., 2013; Shen et al., 2009). Often the sensors used in these works are nowadays even integrated into consumer devices like smartwatches, making the concepts feasible in practice. However, the interplay between those individual sensors and the power of using multiple modalities simultaneously in a multi-modal setup, remain underexplored.

In this study, we thus deploy our CL framework in the e-learning domain, thereby exploring combinations of heart, skin, eye, body posture, performance, and subjective measures. Our study uses a realistic e-learning setting, where participants learn through videos and quizzes instead of using an unrealistic memorization or mental calculation task which related works often use as a proxy for learning. Based on the captured data, we investigate how well predictive models using feature combinations from the explored modalities can predict intrinsic difficulty as well as the perceived CL and difficulty. In particular, we analyze which sensor modalities are more or less suitable for estimating CL, thereby guiding researchers and developers of future cognition-aware e-learning systems and allowing an in-depth comparison to our previous studies in PE of MT.

8.3.2 Analyzed Measures of Cognitive Load

To keep comparability to the PE study with professional translators (section 8.2) high, we use almost identical features from our CL measurement framework (chapter 7).

As *subjective measures*, we now ask participants for two ratings after each video or quiz: For estimating subjective CL (SubjCL), again the commonly used scale proposed by Paas and Van Merriënboer (1994) is utilized. Additionally, the difficulty measure proposed by Kalyuga et al. (1998), which is a 7-point scale, ranging from 1 (extremely easy) to 7 (extremely difficult), asking about the difficulty of the task, is analyzed (SubjDiff). The reason for a second scale was to also capture the concept of difficulty as proposed in the literature on instructional design. While we could have adapted the answer possibilities of one scale to the other, i.e., to have two 9-point scales, we decided to use both scales in their original unaltered form, in which they have been validated and used for a large variety of experiments.

For *performance measures*, we deviate from the PE studies as the task is different. While the time required to watch a video is not relevant, due to the constant duration of the video, we analyze the quiz time, where we expect more difficult quizzes to require more time. Furthermore, the percentage of quiz questions answered correctly is used as a measure of performance.

As for the *behavioral measures*, the distance to the head is captured by a Microsoft Kinect v2 as in our last experiment, hypothesizing that learners come closer for harder content. Since video tasks are rather passive and the quizzes require only few clicks and no text production, we excluded the previously used keyboard features from this analysis. As before, we also do not perform emotion recognition based on images of participants' faces since they worked poorly in the first study and lead to privacy concerns.

For *eye-based features*, we again use data from a webcam and the Tobii 4C eye tracker to capture the openness of the lids, blinks, fixations, and saccades, as well as the probability of visual search (EAR, BlinkAmount, NormBlinkAmount, FixAmount, NormFixAmount, FixDur, SaccDur, SearchProb). As in the last study, we use the pupil diameter measures (PupilDiameter, $\mathrm{ICA}^{\mathrm{wave}}$, $\mathrm{ICA}^{\mathrm{count}}$, Hilbert). Thus, an identical set of eye features to our previous study with professional translators was used.

For *heart measures*, we also use an identical setup as in the study with professional translators, capturing the heart rate from both the Polar belt and the Garmin watch (HR), and using the Polar belt, as well as the Empatica wristband to capture the RR interval (RR) and based on it various HRV (Rowe et al., 1998) measures RMSSD, SDNN, NN50, and pNN50 (Shaffer and Ginsberg, 2017). Similarly, all blood volume pulse (BVP) measures are also integrated as before (BVPAmp, BVPMedAbsDev, BVPMeanAbsDiff).

Skin-based features, captured by the Microsoft Band v2 and Empatica E4 and Garmin Forerunner 935 are also identical to the last study, measuring the galvanic skin response (GSR) and corresponding frequency domain features (FreqGSR, FreqFrameGSR), Ledalab features ($\mathrm{Leda}_{\mathrm{avg}}$, $\mathrm{Leda}_{\mathrm{MaxDefl}}$, $\mathrm{Leda}_{\mathrm{TTP.nSCR}}$, $\mathrm{Leda}_{\mathrm{TTP.AmpSum}}$, $\mathrm{Leda}_{\mathrm{TTP.Lat}}$, $\mathrm{Leda}_{\mathrm{CDA.nSCR}}$, $\mathrm{Leda}_{\mathrm{CDA.AmpSum}}$, $\mathrm{Leda}_{\mathrm{CDA.Lat}}$, $\mathrm{Leda}_{\mathrm{CDA.SCR}}$, $\mathrm{Leda}_{\mathrm{CDA.ISCR}}$, $\mathrm{Leda}_{\mathrm{CDA.PhasMax}}$, $\mathrm{Leda}_{\mathrm{CDA.Ton}}$), and skin temperature (SkinTemp).

In terms of analyzed measures, the difference to the previous study in section 8.2 is thus merely that no typing measures are used, and that slightly adapted subjective and performance measures are employed. The remainder is kept identical. However, the captured continuous signals are processed slightly differently. Instead of transforming the normalized data per content into 6 features, we now only transform it into the 5 features average, standard deviation, minimum, maximum, and range ($\mathrm{max} - \mathrm{min}$). Thus, we do not use the accumalated values as a feature, as these could leak information about the time, and time itself could reveal the content, as the videos chosen for the different levels of difficulty slightly differ in duration. Given all our single features and calculating the 5 subfeatures for the continuous ones, we analyze 202 features overall.

We also again manually inspected the data distribution per content item and participant for outliers and overall data quality. Values were filtered according to visual inspection and related literature: Data above 100000 kΩ for the raw skin resistance, as well as Polar RMSSD above 300, SDNN values above 250 (van den

Berg et al., 2018), and finally Polar HR and RR samples which fall outside the acceptable 50–120 beats per minute or 500–1200 ms ranges were removed (Shaffer and Ginsberg, 2017). Here, the HRV filtering is even more strict than in the previous study, while the other filters are identical.

8.3.3 Procedure, Apparatus, & Content Used

Using this feature set, we explore approaches for estimating CL in e-learning. As before, the study was approved by Saarland university's ethical review board.

Overview

For the study, we use a Moodle[62] environment, as it is one of the most widely used open-source e-learning platforms. After signing a data protection form and granting permission to use the collected data, participants are asked to fill out a quick pre-questionnaire. Then, they go through six pairs of mathematics videos and corresponding quizzes in counter-balanced order. After each quiz, there is a small break task and at the very end, a final questionnaire is filled out.

Pre-Questionnaire

The initial questionnaire captures demographics, previous e-learning experience, and information about last night's sleep quality and length, already performed actions since waking up on the day of the study, as well as perceived exhaustion and tiredness. Last, the math background of the participants is captured to (a) confirm that they match our targeted group (see Section 8.3.4), and (b) to see if effects found might depend on differences in prior knowledge.

Apparatus

Then the learner is equipped with a Microsoft Band v2 on her right wrist, a Garmin Forerunner 935 sports watch and an Empatica E4 wearable on the left wrist (the Garmin is further up), a Polar H7 heart belt on her chest, and a Tobii eye tracker 4C with Pro SDK, as well as a web-cam and a Microsoft Kinect v2 camera facing her. As input possibilities, keyboard and mouse are used, and a 22-inch monitor displays the Moodle environment. Thus, an identical setup to the previous PE study except for a slightly smaller screen and of course a Moodle environment instead of SDL Trados Studio for the task is used.

[62] https://moodle.de/

Videos

We chose 3 mathematical topics for the experiment: vectors, integration, and eigenvectors. For each topic two videos are presented, one considered easy as it is part of the curriculum for the high school certificate, and one considered hard as it is part of the university's "mathematics for computer scientists" curriculum. Note here that this distinction into easy/hard is based solely on the concept of intrinsic CL, while extraneous and germane load also depend on how the material is taught (Sweller et al., 1998). We aimed to make the teaching style as comparable as possible by using videos from the popular German math Youtube channel "Mathe by Daniel Jung"[63], thereby ensuring the videos to have the same speaker, presenting in a very similar fashion, being filmed from the same perspective, etc. The length of the videos is roughly 5 minutes each (mean=312s, min=247s, max=368s). After each video, participants quickly assess their CL as well as the content difficulty on the two rating scales for SubjCL and SubjDiff (see Section 8.3.2).

Quizzes

Afterwards, and similar to Ishimaru et al. (2017), participants take a multiple choice quiz of 2 to 4 questions on the previously watched content, testing whether they understood what they saw. In contrast to the videos, which are consumed rather passively, the quizzes ensured that participants had to actively work. The quizzes were created by us, then refined after discussions with two students matching our participant profile, and afterwards tested in a pre-study with two participants. While we cannot guarantee that the quizzes are didactically 100% comparable, the question design and the iterative testing aimed to make the quizzes as consistent as possible. Participants also had to rate each quiz on the same two subjective scales as the videos (see Section 8.3.2).

Break Task

Following each quiz, participants had to engage in a break task to limit the interference between content items. The task encompassed connecting numbers drawn on paper in increasing order (see Vernon (1993)) and verbally stating each number while drawing, to clear both the visual as well as the verbal working memory (see Baddeley and Logie's (1999) model of working memory).

Post-Questionnaire

At the very end, participants had to fill in a final questionnaire which again captured tiredness, stress, and exhaustion, as well as motivation, to be compared

[63] https://www.youtube.com/channel/UCPtUzxTfdaxAmr4ie9bXZVA

to these factors before the experiment, thereby analyzing tiredness effects. Furthermore, participants had to judge the relative differences in difficulty between the three content topics.

8.3.4 Data Analysis Results & Discussion

For the data analysis, we first analyze the questionnaire results and look at the subjective ratings and time required for individual content items, thereby validating that the chosen method for data acquisition works as planned.

Then we analyze how well our three main metrics, (1) subjective CL (SubjCL, regression in range 1 to 9), (2) subjective difficulty (SubjDiff, regression in range 1 to 7), and (3) intrinsic difficulty (IntrinDiff, binary classification whether the content is part of the university's or high school curriculum), can be estimated based on the captured sensor data. Analysis (1) here is thus identical to our last experiment on PE of MT.

The last part of the analysis aims to better understand which feature modalities consistently perform better or worse than others, thereby providing suggestions on how to implement cognition-aware e-learning systems in practice. For this, we present results from multiple analyses, including intraclass correlation coefficients and an analysis of how the performance of the predictive models changes when we leave out different modalities.

Participants and Questionnaire Results

Overall 21 students, aged 20 to 33 years (mean=25.2), participated (m=17), 9 participants at the end of their Bachelor's studies, while 12 were within their Master's studies. Roughly half (9 participants) described their e-learning experience as rather good or good and used platforms like Moodle, Udacity, or Coursera. To ensure a comparable background, we required all participants to be enrolled in a computer-science-related course of study, and to have already successfully passed the mathematics lectures covering our selected topics. We thereby ensure that the participants' background in the selected topics is comparable. Further requirements, e.g., not having taken any additional math-related lecture, would have limited the amount of matching participants too strongly, therefore, we only captured such lectures instead of imposing further restrictions. Furthermore, participants had to self-assess their background in the three chosen topics on 5-point scales, where they claimed to have the most prior knowledge for the topic of vectors (mean=3.38, σ=0.81), closely followed by integration (mean=3.14, σ=0.85), and last, eigenvectors (mean=2.67, σ=0.97). In the post-questionnaire at the very end, participants were asked to rate the three topics in terms of difficulty on a 7-point scale: Vectors were rated the easiest (mean=2.19, σ=1.12), followed by integration (mean=3, σ=1.18), and eigenvectors (mean=3.05, σ=1.28), where 3 corresponds to "rather easy". According to a univariate ANOVA for the three topics, vectors are significantly easier than the other two topics, which are on the same level (F(2,40) = 4.84, $p < .05$). This means that we should also investigate each topic separately and not only analyze the differences between all easy and all hard content items. Using an ANCOVA to test if these differences only come from a higher prior knowledge in the topic of vectors shows that this is not the case (F(2,18) = 0.37, p = .693 for interaction between topic and prior knowledge).

The current tiredness (mean=2.57, σ=0.98), exhaustion (mean=2.05, σ=0.87), and stress (mean=2.0, σ=0.95, all ratings on a 5-point scale) were in an acceptable state at the beginning of the experiment, probably since the previous activity was not considered strenuous (mean=1.67, σ=0.91, all ratings on a 5-point scale). The corresponding values after the experiment (exhaustion (mean=2.29/5, σ=0.78), stress (mean=1.95/5, σ=0.97), and tiredness (mean=2.58/5, σ=0.96)) showed no significant differences from the ratings before. This, combined with the rated demand of the experiment (mean=3.57/5, σ=0.81), shows that the data should not be substantially distorted by tiredness effects. The post-questionnaire further showed that participants had a high motivation to follow the videos (mean=3.81/5, σ=1.08) and a very high motivation to perform well on the quizzes (mean=4.48/5, σ=0.75).

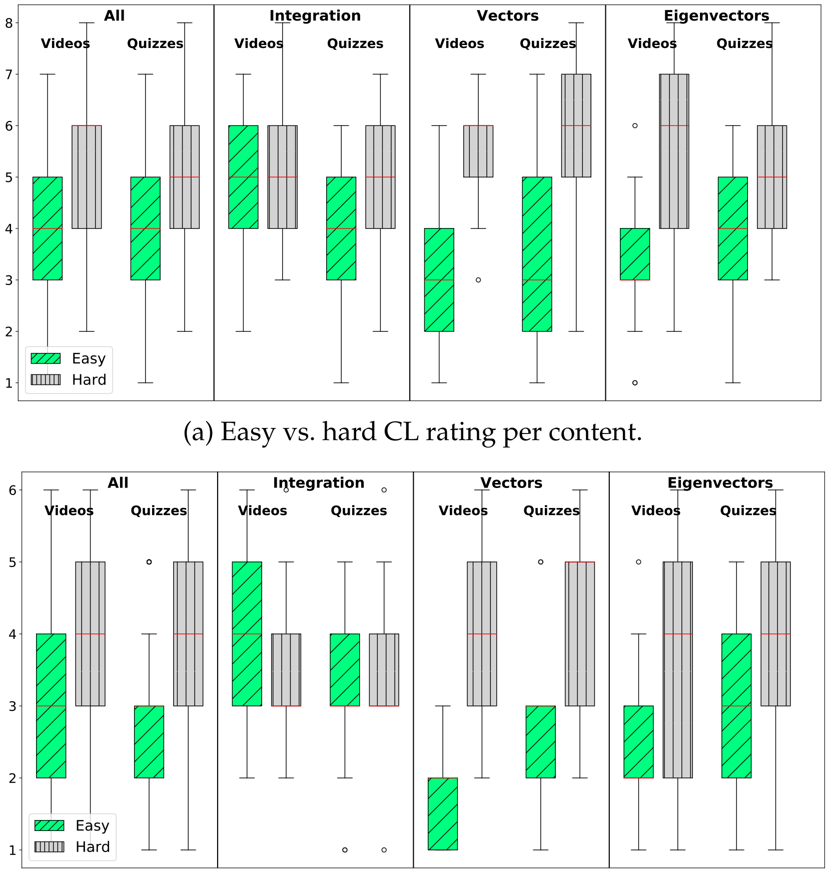
Content-Wise Ratings and Quiz Results

Content-Wise Subjective Ratings Figure 8.7 shows the CL and difficulty ratings for the quizzes and videos of each content item. While the differences across all topics, as well as within the topics vectors and eigenvectors, are clearly visible and significant (all $p < .01$), the integration content did not impose any statistically significant difference in perceived CL or difficulty.

Correlations of CL & Difficulty CL and difficulty ratings correlate significantly (all $p < .01$) and strongly for all contents, with Pearson correlation coefficients between .58 (for the easy integration videos) and .89 (for the easy integration quiz). Thus, participants considered the two constructs as highly similar.

Inter-Rater Differences We analyze inter-rater differences by considering the standard deviation across all ratings per video/quiz content. The results show that the average standard deviation is 1.34 (min=1.14, max=1.56) for video CL, 1.16 (min=0.73, max=1.71) for video difficulty, 1.47 (min=1.19, max=1.74) for quiz CL and 1.11 (min=1.00, max=1.25) for quiz difficulty. Thus, the inter-rater differences between videos and quizzes are very comparable, especially when considering the 9-point vs. 7-point rating scale.

Quiz Time & Performance Table 8.9 shows that strong differences in quiz times exist between the content items and that for all three topics, the quiz time was higher for the harder content. The average percentage of correct answers to



(a) Easy vs. hard CL rating per content.

(b) Easy vs. hard difficulty rating per content.

Figure 8.7: Subjective CL and difficulty ratings.

the quizzes was lowest for the eigenvector quizzes, while the other two topics were comparable. On all three topics students indeed performed better on quizzes corresponding to simpler videos; however, the differences are very small, showing that such performance measures themselves do not always work.

Discussion Overall our participants had comparable backgrounds in the explored mathematical topics and the data should not be substantially distorted by tiredness effects. Furthermore, and as anticipated, there is a significant difference between easy and hard content for both videos and quizzes in terms of CL and difficulty ratings except when considering the integration topic individually. Thus, we should also investigate each topic on its own. Furthermore, we see a strong correlation between CL and difficulty, indicating that participants perceive the two constructs as very related. While participants indeed required more time for the quizzes on hard content, the percentage of correct answers was rather comparable between easy and hard quizzes.

	Integration		Vector		Eigenvector	
	Easy	Hard	Easy	Hard	Easy	Hard
Time	88 (31)	129 (49)	119 (53)	199 (81)	93 (35)	174 (103)
Perf.	93 (18)	91 (15)	94 (16)	87 (20)	85 (18)	81 (24)

Table 8.9: Quiz time (in seconds) and performance (% of correct answers) mean and standard deviation (in brackets) for the different content items.

Predictive Models

We now aim to use the various captured measures to predict the demand imposed by the content as defined by our three measures IntrinDiff, SubjDiff, and SubjCL. After looking at correlations between features and these measures, we investigate how to best select a subset of the implemented features and which models are most suitable for these classification and regression tasks. Then we train the actual models and discuss their respective results. The regression problem of predicting SubjCL is thus identical to our previous two studies.

Correlation to Target Variables We analyze how strongly the individual features correlate with our target variables. Since we have lots of features and 3 different target variables, inspecting each individual correlation is unfeasible. Instead, we look at the highest correlations for both videos and quizzes, per topic and across topics.

Across videos, correlations are rather weak, where the maximum correlation coefficient of 0.2 was achieved for SubjCL (0.18 for SubjDiff and 0.14 for IntrinDiff). However, for the individual topics, we get much better results: for vectors the best correlations are within 0.38 and 0.44 for the 3 target measures, for integration within 0.39 and 0.43, and for eigenvectors between 0.30 and 0.36. For the quizzes, correlations are much higher, both across all topics (between 0.38 and 0.42) and within topics, with the highest correlation coefficients between 0.48 and 0.52 for the vectors, 0.39 to 0.52 for integration, and 0.42 to 0.48 for eigenvectors.

Feature Amount & Model Selection This section describes the experiments conducted to determine an appropriate model as well as an ideal number of features to use for training predictive models on our data. As a feature selection approach, we use recursive feature elimination with cross-validation (`RFECV` in `scikit-learn`) as it turned out to give better results than other feature selection approaches that we explored and makes results comparable to our previous study. As possible numbers of features, we test values ranging from 5 to 100 with an increment of 5. As for machine learning models, which also influence the number of features to select, we test the following models: linear models with different regularizers, namely a Logistic Regression, a Stochastic Gradient Descent regressor, a Lasso model, an Elastic Net, and a Ridge regressor, as well as a non-linear Random Forest regressor, all provided in the `scikit-learn` library. 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 differences by adding a random effect for subject and/or content[64]. These are again the same models we explored for CL estimation in PE, except that we added classification models for predicting IntrinDiff.

For each model and feature amount combination, we test different hyperparameter settings of the model to get its best performance on that number of features[65]. Note that this is not an exhaustive search in the hyperparameter space but rather a heuristic-based approach in searching for good hyperparameters. Finally, we plot all models' performances for CL rating, difficulty rating and intrinsic difficulty, once for videos and once for quizzes. This hyperparameter tuning step is thus an extension to our data analysis approach from the previous studies.

Across all 6 cases we get comparable results: For the classification case (IntrinDiff), the best results were achieved using Logistic Regression or LMEM models, especially for a small amount of features. For regression (SubjDiff, SubjCL), LMEM and Ridge performed best, showing that linear models seem to perform well on our data. Note that in the study with professional translators, LMEM, Ridge, and RandomForest were frequently chosen. Regarding feature amount, the range of 30 to 35 features gave good results across all 6 analyses. Since fewer features help interpret the results, we decided to use RFECV with 30 features in the following. While both linear models and LMEMs perform equally well in this preliminary analysis, Ridge and Logistic Regression are even simpler than LMEM, giving better generalization due to less participant dependence, which is why we use them in the remaining analyses.

To avoid overfitting, a 10-fold cross validation was used, and the best hyperparameters determined by grid search, namely Ridge regression with $\alpha = 2$ for regression, and Logistic Regression using L2 normalization and $C = 1$, were chosen. As features, we use all features presented in Section 8.3.2, with a few exceptions: Since the subjective measures are our target variables, we do not use them as predictors. Furthermore, we exclude the performance measures, as these exist only for the quizzes and not for the videos, resulting in a total of 202 features.

[64] Since the R package used for LMEMs does not support our feature selection approach, we instead perform feature selection with a Ridge model for regression and Logistic Regression for classification. For classification, we did not add a random effect for item (in our case the video/quiz) to the LMEM, as this would have trivially resulted in 100% accuracy. For the regression case, however, we did add a random effect for item as well. Thus, the LMEM approach uses the above measures to predict intrinsic difficulty, SubjCL, and SubjDiff, but additionally knows which participant (and content) the data comes from. In contrast, the `scikit-learn` models cannot react differently depending on participant or item.

[65] For SGD regressors, we explore L1 ratios of 0.15 and 0.5; for Lasso models alpha values 1, 2, and 10; for ElasticNet alpha values of 0.5 and 1 in combination with L1 ratios of 0.25, 0.5, 0.75; for Random Forests (both for regression and classification) we explore 10, 20, 30, and 50 for numbers of estimators and a maximum depth of None, 4, 8, and 12. For Ridge models we explored alpha values of 0.5, 1, 2, and 10; for SGD classifiers, alpha values 0.0001, and 0.01, with L1 and L2 regularization; for Logistic Regression C values of 0.5, 1, and 2, both with L1 and L2 regularization.

Furthermore, if every entry for a whole feature contains the same value, we drop it (which happened for 3 "minimum" features). If due to a sensor failure some data values of a feature are missing, we replace them by the participant's mean value for that feature (if available), or by the global mean (if no data exists for a particular feature for that participant), which happened 5 times. Furthermore, 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.

Classifying Intrinsic Difficulty Using the settings described above, we train models classifying IntrinDiff, i.e., binary classification. Figure 8.8 depicts the accuracy achieved by the Logistic Regression models in comparison to a simple baseline always predicting 'easy' (achieving 50% accuracy). As can be seen, distinguishing easy from hard quizzes based on the sensor data works very well for the quizzes (80-90% accuracy), both across topics and within topics. For the videos, however, only around 70-75% accuracy is achieved for the vector and eigenvector topics as well as across topics. A reason could be that the videos are consumed only passively, where sensor data might be less reliable. This difference is also visible in the correlations above (Section 8.3.4), where higher coefficients were found for quizzes than for videos. For the integration videos, very high classification results were achieved, probably due to some artifact in the data, for which we currently do not have a concrete explanation. One should note here that these results were achieved using feature selection on all available features; therefore, section 8.3.4 explores how results change when only single modalities are used.

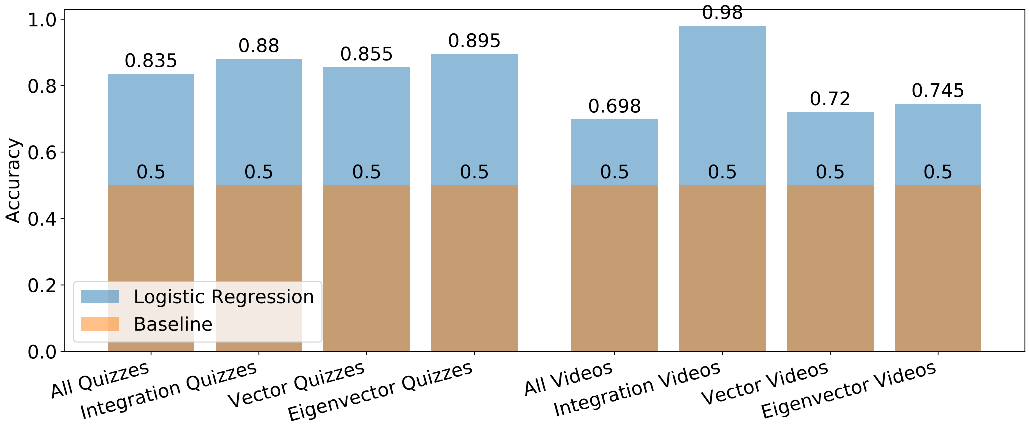
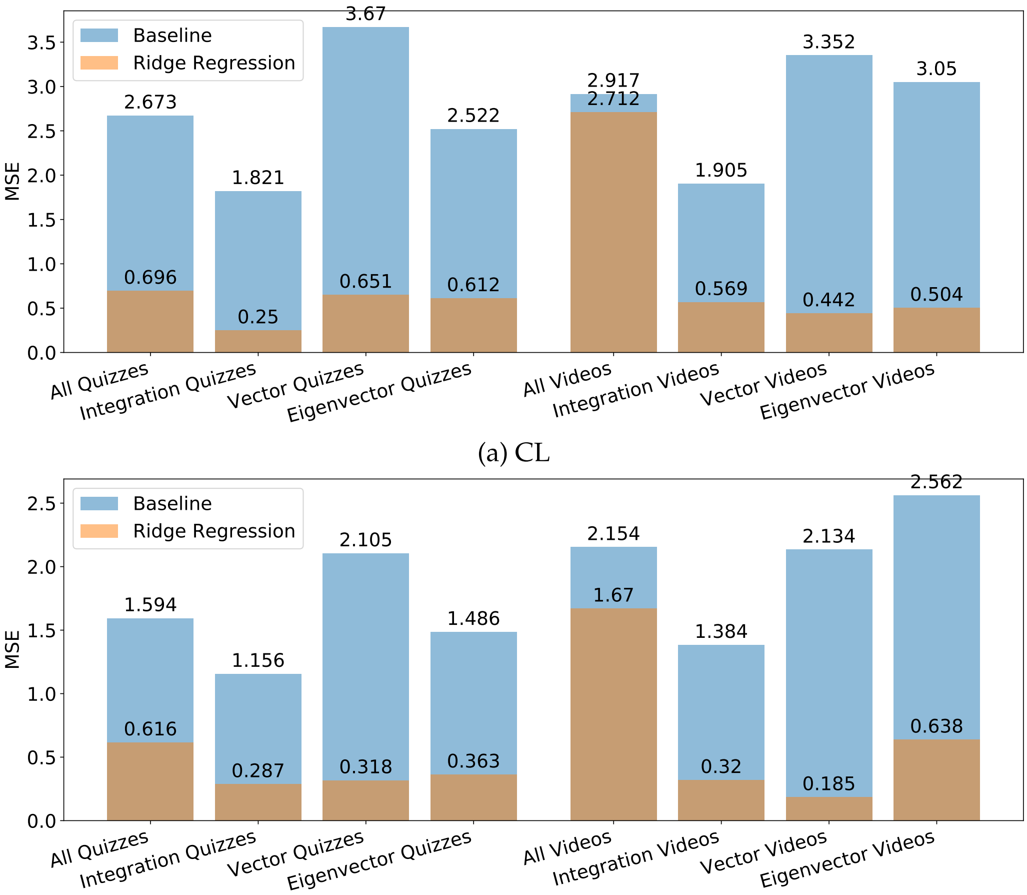


Figure 8.8: Classification of intrinsic difficulty.

Regression for Predicting Subjective Difficulty and Cognitive Load Next, we check the MSE for the regression models predicting SubjDiff and SubjCL for both quizzes and videos in comparison to simple baselines always predicting the mean value of the corresponding rating, as depicted in Figure 8.9. For each topic individually, very large percentage gains are achieved over the baseline; across quizzes the prediction also yields good results, while across videos, only marginal gains were achieved. A potential explanation might be that the differences when passively watching videos are less well represented in the physiological data than those that appear when actively working on the quizzes, especially when adding the variability of the different topics instead of comparing content within a topic. Section 8.3.4 also shows that the correlations across videos are much lower than within topics or across quizzes, explaining the bad results in this particular case. It is also interesting that the prediction of SubjCL and SubjDiff also works for the integration topic, where the subjective differences were not large, which again can be explained by the existing correlations presented in paragraph 8.3.4. Overall, the final MSEs found are very low (except for the across videos-case), indicating that within content topics one can estimate the imposed demand very well. Thus, we see that similar to our studies on CL estimation in PE of MT, SubjCL can also be estimated with low error in e-learning.



(a) CL

(b) Difficulty

Figure 8.9: Regression performance for SubjCL and SubjDiff.

Discussion Overall, the analysis shows that using roughly 30 features together with Ridge/Logistic Regression is a reasonable choice for our dataset. Furthermore, for both the classification and regression case, results for the quizzes are better than for the videos, which we believe comes from the higher degree of activity while solving quizzes than while watching videos. This in turn better differentiates the physiological data. Note that the PE setting also shows such a higher degree of activity and is thus closer to the quiz than to the video case here. Therefore, using feature selection on all modalities simultaneously proves to work very well for the quizzes for binary classification of IntrinDiff and for all regression cases except for predicting across videos. However, note that the limited amount of data might have introduced some bias, even though the results look consistent.

Modality Analysis

Now we have gained some insight into how well the three target measures can be predicted for the different content items. In this section, we aim to understand which feature modalities contribute how strongly to the models, thereby providing suggestions on how to implement cognition-aware e-learning systems in practice. This analysis thus has a similar goal as our previous studies in PE: investigating which modalities are more or less suitable and if a multi-modal approach is beneficial.

Modality Correlations to Target Variables To estimate the direct link between the modalities and the 3 target variables, we analyze the 10 highest correlating features within the 8 cases (across quizzes/videos, within each quiz/video). Of these in total 240 (3*10*8) features, 113 are eye features, 68 heart features, 53 skin features, and 6 body posture features, suggesting that eye features perform best, followed by heart and skin measures.

Naturally, this approach has some limitations: the same feature could count up to 24 times (to all target variables of all 8 cases), and there is a different number of features per modality. However, this initial analysis captures the direct link to the target measures independent of there being even more irrelevant features in the modality, and independent of linear dependency and thereby redundancy of multiple features. Nevertheless, to also investigate the predictive power, further analyses are presented in the following.

Modalities Selected through Feature Selection We analyze the features selected among all possible features to see if measures from some modality tend to be selected more or less often by our feature selection approach. We compare this among the total of 24 tasks (3 target measures times the 8 cases: across videos/quizzes and within the 3 topics each with video/quiz) for which we train our models. To better analyze the selected feature set, we count which individual features are selected most often. Then we check which modalities these highly selected features belong to. This shows that eye and heart features are selected most often (up to 17 out of 24 times), showing the importance of these features. In contrast, skin and body posture features are maximally selected 7 times, and can therefore be considered less important.

Intraclass Correlation Coefficient Apart from a high correlation to the target variables and being selected by the feature analysis, it is also interesting to analyze the degree to which participants resemble each other regarding a given feature. For this, we use the Intraclass Correlation Coefficient (ICC). It ranges between 0 (chance agreement) and 1 (perfect agreement) and gives "an indication of the extent to which different [participants] produce the same values of a given measure when exposed to the same [conditions]" (Vieira, 2016). It therefore indicates features that already generalize well from a small number of participants. According to Koo and Li (2016), values below 0.5 can be considered as 'poor', values between 0.5 and 0.75 as 'moderate', values between 0.75 and 0.9 as 'good', and everything above 0.9 as 'excellent'.

Overall, only 2.5% of the features can be considered 'excellent', 3.5% as 'good', 27% as 'moderate', and the majority of 67% as 'poor'. This is particularly interesting because the investigated measures were all proposed in the literature and used in CL studies, which usually do not report the ICC and mostly have a similar or even smaller amount of participants. Comparing the values to one of the few papers by Vieira (2016) that also reported ICC values on multiple modalities, we find that his 7 explored features were all in the range 0.25 to 0.6, and thus also 'poor' to 'moderate' according to Koo and Li (2016). The high amount of features in the 'poor' category can also be explained by the fact that not all of the 5 features calculated on top of continuous signals (min, max, mean, σ, range) consistently yield good ICC values. Furthermore, it can be seen as an explanation for why selecting 30 features already gives good results.

In terms of modalities, again eye features perform best, comprising 100% of the 'excellent', 57% of the 'good', and 11% of the 'moderate' features. Between heart and skin there does not seem to be a clear winner: heart features make up only 14% of the 'good' but 48% of the 'moderate' features, whereas skin yields 29% of the 'good' and only 11% of the 'moderate' features. All body posture features are within the 'poor' category.

Prediction Performance of Different Modalities To get more insight into the classification/regression performance that can be achieved through the different modalities (heart, skin, eyes) and combinations thereof, we train models on the combined modalities and on subsets of the modalities. Thus, as in our last studies on PE, we use the predictive power of the modalities as a proxy for their suitability of CL estimation. We ignore body posture features here, as the performance was very poor.

Figure 8.10 shows the results achieved compared to the baseline reported above, each plot containing a group for SubjCL, for SubjDiff (both MSE), and for IntrinDiff (as accuracy). The MSEs for SubjCL are higher than those for SubjDiff, as the subjective scales defining the target variable were 9- vs. 7-point scales. Note that this analysis was done across all quiz contents; the corresponding analysis across all video contents is omitted, because using all modalities across videos already resulted in only marginal gains compared to the baseline (see Figure 8.9). The resulting plot therefore did not provide further insights and is omitted for space reasons.

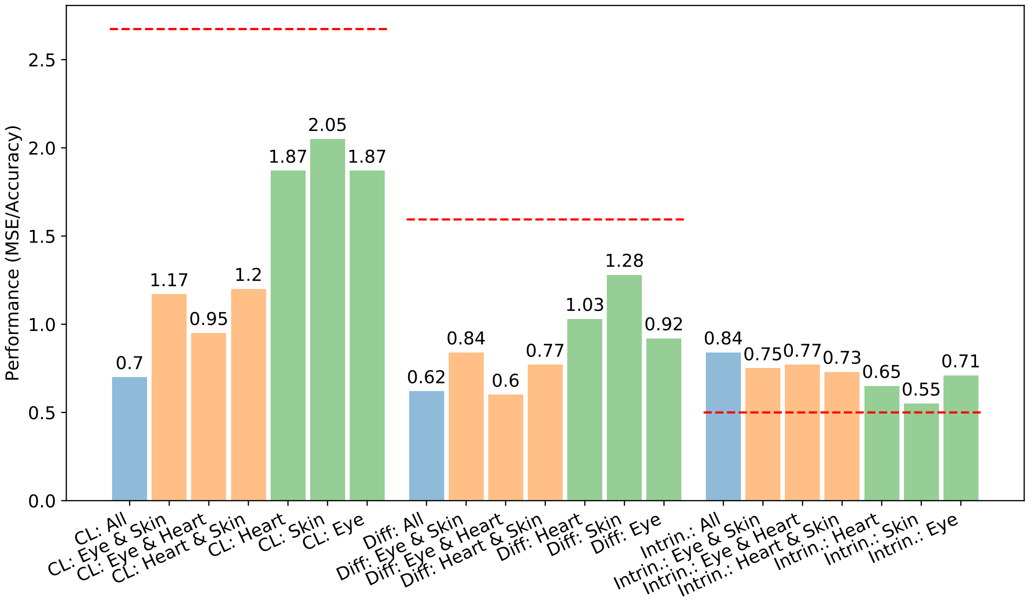


Figure 8.10: Regression performance in terms of MSE (SubjDiff and SubjCL) and classification accuracy (IntrinDiff) for prediction across all quiz contents when using only features from certain modalities.

As can be seen in the figure, the multi-modal approach is consistently better than single modalities; however, the combination of eye and heart features is also comparably good. Furthermore, we note that there is a trend that heart and eye features perform better than skin features, which can also explain why the combination of the two outperforms the modality pairs containing skin.

Discussion The various sub-analyses conducted to see which modalities perform better and worse, some focusing on correlations, others on predictive power, consistently show that eye features perform best, followed by heart, then skin, and last body posture. Combining two modalities improves results compared to single modalities, where eye and heart features combined performed best. We thereby extend the findings by Naismith and Cavalcanti (2015), who showed that eye features are more reliable than cardiovascular features in medical training, by additionally considering skin features and combinations of modalities.

8.3.5 Conclusion

This study with 21 participants has explored our CL measurement framework in e-learning, where it also comprised the most diverse set of features from a variety of modalities. We used a rather realistic e-learning setting relying on videos and quizzes instead of an unrealistic mental calculation task as done in many related works. With the data captured in this setting, we show that classifying intrinsic content difficulty works better for quizzes, where participants actively solve problems, than for videos, which they passively consume. This quiz case is thus also more related to PE of MT, where translators also actively engage with the text instead of passively consuming it. Our classification results are roughly comparable to Borys et al. (2017), who achieved up to 73% in a trinary classification task (no-task, low CL, high CL), but used a much less realistic setting where participants had to perform mental calculations instead of learning through videos and quizzes. It is also interesting that even though we did not use EEG measures, the combined power of multiple modalities gives comparable results. Regression analysis for predicting the subjectively reported level of CL and difficulty also works with very low error within content topics. Among the explored feature modalities, eye-based features yield the best results, followed by heart-based and then skin-based measures. Furthermore, combining multiple modalities results in better performance compared to using a single modality, which is an overall trend that we have also seen for PE of MT. The presented results can guide researchers and developers of cognition-aware e-learning environments by suggesting modalities and features that work particularly well for estimating difficulty and CL. Furthermore, the results suggest that adaptations like content recommendations, break proposals, or speed adaptations would be feasible using a multi-modal approach. One should however note that the data was captured from only 21 participants learning 6 mathematical contents, so further studies should be conducted in different learning domains.

This study also contributed to our second research question on the use of multi-modal sensing devices for CL estimation. However, instead of focusing on our main application area of PE, we explored in an in-depth analysis if our approach also works in another domain. We chose e-learning because CL theory originated in the educational context in which e-learning offers a similar setting to PE with one individual in front of a computer. Here, we see that our general framework also works in this domain, and further that multi-modal combinations reduce prediction errors, but also that eye, heart, and skin features already offer good results. We are thus optimistic that our CL estimation framework can indeed be used for measuring and in the long run reducing CL through adaptations during cognitively challenging tasks like PE or learning. Apart from a feasibility analysis, it is however also important to understand which of the various sensing devices users have reservations about for the purpose of CL adaptations. This is what we will investigate this in the next chapter.

Chapter 9
Privacy Concerns Regarding Cognitive Load Adaptations

The previous chapter explored our CL measurement framework in three studies and showed that cognition-aware PE or cognition-aware e-learning is overall feasible, even though further steps are required to implement the concept in practice. Similar to these studies, the related literature has mainly dealt with conceptual implications and technical possibilities to model the cognitive state, without considering whether users would also accept being monitored by the sensors for this purpose. Some works investigated privacy perceptions for wearables, however, no systematic evaluation of the users' privacy concerns regarding cognition-aware interfaces exists, which is what this chapter contributes. We mainly focus on the e-learning domain again, for which it is easier to find participants (compared to PE). However, the results are mostly not specific to e-learning, but rather general and thus also interesting for PE within adaptive CAT tools.

This chapter is based upon Herbig et al. (2019d).

9.1 Method

To gather feedback on the user acceptance of cognition-aware e-learning systems, we conduct an online survey with a variety of potential e-learning users. The evaluation has been approved by the university's ethical review board as well as the data protection officer and consists of the following blocks:

Demographics & Background The survey starts by asking about the participants' demographics and their usage of e-learning.

Willingness to Disclose Sensor Data (A) Then, we ask about participants' willingness to share sensor data with an unspecified application, i.e., without providing the context of e-learning. The 4-point scale asks whether they could imagine sharing the data, ranging from "not at all", to "would rather not", "probably would" and "completely". This question always appeared before B (see below), to receive replies that are not biased by the context of e-learning. Apart from a different participant sample, these results are of a general nature and therefore also practically applicable to PE.

Required Performance Improvements (B) After explaining that such data could be used to detect the cognitive state for adaptive e-learning, we let participants judge how big an improvement (in terms of faster learning or making fewer mistakes) would be necessary for them to disclose the individual sensor data, ranging from "none" to "small", "moderate", "strong" and "immense" improvements. This can be seen as a similar approach to Acquisti et al. (2013), who investigated the amount of money necessary to share otherwise private data in retailing. Note that the decision to formulate both questions positively (without artificially negating one), leads to the right-most value of B being related to the left-most value of A.

Sensors We use the following list of sensors, which covers a wide (yet incomplete) set of approaches for context awareness and cognitive load detection: heart rate, skin resistance, skin temperature, respiratory rate, body posture, blood pressure, typing/mouse/touch behavior, eye movements and blinks, pupil diameter, facial expressions, steps per day, mode of locomotion (e.g., in a vehicle), surrounding noises, ambient brightness, and location.

Adaptation Ideas & General Feedback Last, we ask participants for ideas on how e-learning tools could adapt to the users' states and what their general attitude towards this idea is. These open-ended answers are clustered based on manual coding. Even though the methodology is different (survey vs. interview), the goal is similar to chapter 6, namely to develop adaptation ideas and analyze users' perceptions of cognition-aware systems.

9.2 Results

On average, participants needed 9:09 minutes to complete the survey (σ=5:02). The following sections present the results for the different blocks of the study.

Demographics Overall, 50 participants, aged 19–48 (mean=28.7, σ=6.32, 19 female), were recruited using Academic Prolific[66], where we paid more than

[66] https://www.prolific.co/

the minimum wage. The only screening criterion we had was that their first language was German, since (a), the questionnaire was in German and (b), cultural differences might occur in such a privacy analysis. The participants had a rather high level of education, with only 7 participants not having general qualification for university entrance, 20 having this qualification but without a university degree, 15 having a Bachelor's degree, and 8 having a Master's degree or diploma. Furthermore, they had strong experience with technical products (mean=3.9 out of 4, σ=0.46). All of them used a computer and smartphone, half of them a tablet, and 22% a smartwatch or fitness tracker. 10 reported having no experience with e-learning systems, while the remaining 40 reported an average experience of 2.75 (σ=0.71), which tends towards "rather high" on a 4-point scale. On average they learn electronically for 3.25 (min=0, max=20, σ=5.13) hours per month and mainly use a PC or laptop for this (29/40). As e-learning platforms, participants mostly use Duolingo, Babbel, Moodle, Udemy, Codecademy, and Coursera. These platforms also make sense if one considers the main goals reported by the participants: learning of languages, learning programming, or using them in courses at the university.

General Feedback At the end of the survey, after having introduced the concept of cognition-aware e-learning, we asked participants about their opinions on this idea. Here, the answers were rather inconclusive, at 2.62/4 (σ=1.01), where 3 means "somewhat positive". Analyzing if there is a correlation between e-learning experience and the participant's opinion on the idea of cognition-aware e-learning, we find that numerically it exists, but the link is not significant, with p=0.101 for two-sided correlation, or respectively p=0.0505 if we assume a positive correlation in the first place.

When asked whether the participants have misgivings regarding the concept of cognition-aware e-learning, 22 participants reported that this was the case. The reasons stated by the participants are all of the form "data protection/surveillance/data theft". In contrast, 20 participants either had no misgivings, finding it "flexible/performance oriented", or saw no issues under the assumption that the data used would be communicated transparently, not sold, utilized only for this purpose, and that the user could self-define the individual sensors from which data is being used. Further comments were of the form "would be absolutely great/intriguing". 2 simply stated that they know their current state and which content is suitable for their situation themselves. Lastly, 6 participants provided no opinion on this question. These mixed feelings are in line with our findings in chapter 6, and CL-aware CAT tools would also benefit from the highlighted factors like opt-in mechanisms and data transparency.

Adaptation Ideas Interesting ideas on how to adapt e-learning tools towards the current situation and cognitive state of the user were provided. We clustered the participants' various proposals: 29 proposals were of the form *adapt content*, either by recommending the content itself, or by adapting duration, difficulty, speed, level of detail, or intensity. Furthermore, 17 participants suggest *varying the duration*, e.g., by proposing breaks in between, by changing the duration of learning intervals or by splitting learning content into parts of different length. 3 suggestions were to *recommend times for learning* where one could learn most efficiently. Another 3 proposals suggested *adapting the interface* to reduce strain through optical changes. 2 suggested relaxation exercises, e.g., some form of meditation, when high loads are detected. Among the other infrequent proposals was the idea to detect when the user only scans through text, to adapt the time limits, vibrate on attention loss, provide individual learning goals, simply use it for a quantified self-style motivation, provide individual feedback, or use the data to improve the learning content for the future. Even though the question was raised in a different domain, we see comparable replies to the answers given in chapter 6 on PE, e.g., to adapt content/show support, propose breaks, or change the interface based on detected CL.

Willingness to Disclose Sensor Data (A) When asking participants on a scale from 1 to 4 about their willingness to disclose data from different sensors, by simply assuming that *an application* would require this information, we get very indifferent results: the averages per sensor are in the range $[2.16, 2.92]$, where 2 is "somewhat disagree" and 3 is "somewhat agree"; standard deviations are within $[0.966, 1.216]$.

We test the results for each sensor with two-tailed t-tests for significance against 2.5, which is the mean of our four points, to get a clear understanding of the overall tendencies. The results in Table 9.1 show positive significant differences, meaning that there is a clear tendency to disclose the data, for typing, mouse and keyboard behavior (with $p < 0.05$). We find negative significant differences, meaning that they would rather not disclose it, for facial expressions, ambient noises, and pupil diameter ($p < 0.05$). For all other sensors, we do not find significant differences; however, we report the tendencies in the data: movement (number of steps), mode of motion, heart rate, breathing rate, surrounding brightness, and skin temperature showed a positive tendency (towards disclosure), while location, eye movement, skin resistance, body posture, and blood pressure showed a negative tendency (against disclosure). Note that these results were given independent of the e-learning context and thus also apply for PE.

Required Performance Improvements to Disclose Sensor Data (B) After having introduced the general idea of an e-learning system that can adapt to the user's current cognitive state based on sensor data, we ask participants how big the improvement gained would have to be, e.g., in terms of faster learning, or making fewer mistakes. We also told them to assume that the data is used only for this purpose. Here, we got different mean values and a greater spread than for A: the averages are in the range $[1.74, 3.00]$, and the standard deviations within $[1.258, 1.443]$, where our 5-point scale was from 0 ("no improvement at all required") to 4 ("immense improvement required").

	Sensor	**p-val**	**mean**	**SD**
Positively significant (tendency to disclose)	Typing/mouse/touch behavior	0.003	2.92	0.97
Negatively significant (tendency not to disclose)	Facial expressions	0.033	2.16	1.10
	Surrounding noises	0.042	2.16	1.15
	Pupil diameter	0.049	2.20	1.05
Positively insignificant (tendency to disclose)	Steps per day	0.061	2.80	1.11
	Mode of locomotion	0.438	2.62	1.09
	Heart rate	0.474	2.62	1.18
	Respiration rate	0.623	2.58	1.14
	Ambient brightness	0.699	2.56	1.09
	Skin temperature	0.797	2.54	1.09
Negatively insignificant (tendency not to disclose)	Location	0.394	2.38	0.99
	Eye movements and blinks	0.487	2.40	1.01
	Skin resistance	0.803	2.46	1.13
	Body posture	0.803	2.46	1.13
	Blood pressure	0.908	2.48	1.22

Table 9.1: Tendencies to disclose data depending on sensor (A).

Significance testing is conducted similarly to A, but against 2 ("moderate improvement required"), due to the different scale. Compared to A, we also have an inverted scale polarity: high values indicate greater skepticism. The results can be found in Table 9.2: We find positively significant differences, meaning strong improvements would be necessary, for facial expressions, ambient noises, pupil diameters, body posture, blood pressure (≤ 0.001), location, eye movements/blinking, breathing rate, and ambient brightness (< 0.05). No significant differences were found for the remaining sensors; however, we report the tendencies here: they were positive (meaning strong improvements are likely required), for skin temperature, skin resistance, heart rate, and mode of movement, and negative (meaning small improvements are possibly required) for typing/mouse/touch behavior and movement (steps).

Link between A and B We hypothesize that negative correlations exist between A and B, since a high willingness to disclose the data (A) should reduce the required improvement threshold (B), and a low willingness to disclose the data (A) should result in a high threshold for improvement (B).

Pearson correlation analyses show that this is the case, as all correlations for the sensor data are negative, strong (all $r < -0.5$), and significant (all $p < 0.01$). This can also be seen in Figure 9.1, where A and B are plotted against each other on a scale with equal polarity and ranges for both questions. For this, we linearly scaled the answers from A in the range [1,4] to the range [-2,+2], and mapped the answers from B such that 0 ("no improvement") corresponds to the highest value +2 and 4 ("immense improvement required") to the lowest value -2.

	Sensor	p-val	mean	σ
Positively significant (strong improvements required)	Facial expressions	< 0.001	3.00	1.28
	Surrounding noise	< 0.001	2.80	1.28
	Pupil diameter	< 0.001	2.82	1.32
	Body posture	< 0.001	2.64	1.26
	Blood pressure	0.001	2.64	1.34
	Location	0.012	2.52	1.40
	Eye movements/blinks	0.019	2.48	1.40
	Respiratory rate	0.023	2.44	1.33
	Ambient brightness	0.028	2.44	1.37
Negatively insignificant (small improvements required)	Typing/mouse/touch input	0.175	1.74	1.34
	Steps per day	0.764	1.94	1.41
Positively insignificant (strong improvements required)	Skin temperature	0.137	2.30	1.40
	Skin resistance	0.302	2.20	1.36
	Heart rate	0.317	2.20	1.40
	Mode of locomotion	0.332	2.20	1.44

Table 9.2: Tendencies for performance improvements necessary to disclose sensor data (B).

Participant and Sensor Group Differences We further test for group differences based on interesting sub-groups of our participants, which we defined according to their demographic data.

Analyzing the differences between *"smartwatch/fitness tracker users" vs. "everyone else"* using a t-test per sensor shows that for A, a significant difference ($p < 0.05$) for the steps per day and an almost significant difference for mode of locomotion ($p = 0.051$) exist, with the "smartwatch" group being more likely to disclose data.

Separating the participants into *"techies" vs. "non-techies"* (e.g., software developer vs. nurse) and *"teachers" vs. "non-teachers"* based on their job descriptions and using a t-test, as well as separating the education levels *"no high school graduation" vs. "high school graduation" vs. "college degree"* together with a multivariate ANOVA, does not lead to any significant differences, neither for A nor B.

Since we explicitly asked participants at the end of the survey whether they had misgivings regarding cognition-aware e-learning, we also clustered participants into the groups *"misgivings" and "no misgivings"*, expecting the "misgivings" group to be more concerned in A and B. For A we found, for all sensors except location (where both groups tend towards the middle), that the "no misgivings" group is more willing to disclose the data. Similarly for B, for all but three sensors, the "no misgivings" group requires significantly less improvement to disclose data. The exceptions are location, as well as ambient brightness and typing/mouse/touch behavior, where both groups have similar opinions.

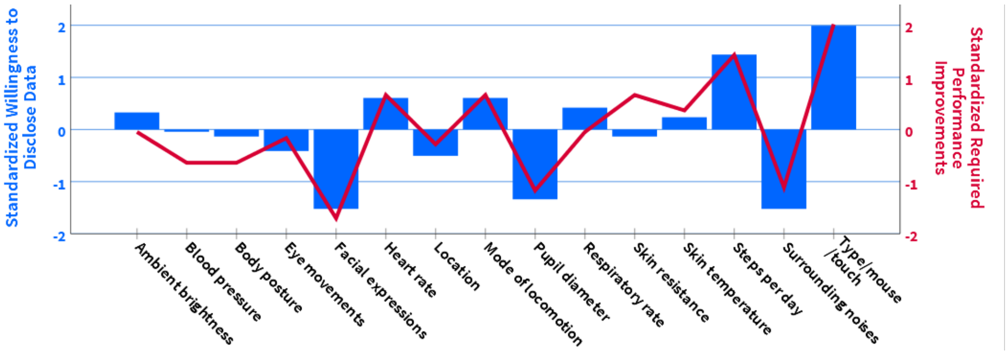


Figure 9.1: User acceptance regarding different sensors. Both analyses (A) and (B) are transformed to the same scale in the range $[-2, +2]$, where positive values indicate high willingness/low required improvements to disclose data.

Last, we group the sensors into *"consumer device sensors" vs. "non-consumer device sensors"*, where the "non-consumer device sensors" comprise blood pressure, body posture, eye movements and blinks, pupil diameter, and respiratory rate, which are not commonly built into smartphones, smartwatches, or fitness trackers. All other considered sensors are part of the "consumer device sensors" group. The results show that there is no significant difference for the general privacy concerns (A) with $t(49) = -1.33$ and $p = 0.190$, but that for the required improvements (B), there is a significant difference $t(49) = 2.90$ and $p = 0.006$, meaning that stronger improvements are required for sensors that are not commonly built into consumer devices than for the commonly built-in sensors.

9.3 Discussion

In this chapter, we investigated the perceptions users have towards data disclosure for cognition-aware systems. In general, our participants would most likely disclose typing/mouse/touch behavior, while pupil diameter, facial expressions, and surrounding noises would not likely be disclosed. We see this as two dimensions of intimacy, where users feel less unique in their typing/mouse/touch interactions, while observing someone's face or surrounding noises could feel more intimate. This is in line with Motti and Caine (2015), who found that sensors like cameras and microphones pose the most privacy concerns. Regarding required performance improvements for e-learning, such information about the surroundings and physiological data would only be disclosed by our participants in exchange for strong improvements. Interestingly, we found significant differences for improvement requirements between sensors integrated into consumer devices (e.g., heart rate or skin resistance) compared to less common measures (e.g., pupil diameter). This could mean that users are more concerned about the new or unknown sensors and that they might become less skeptical once sensors become more widespread. Interestingly within the learning domain, the level of education did not significantly influence the tendency to disclose data, nor the improvements required for disclosure.

While the averages for most sensors tend towards the middle, thereby making effects small, we still found several significant tendencies. A reason for this trend towards the middle could be that for the critical topic of privacy concerns, people do not claim to willingly disclose all personal data without any concerns, but at the same time they know from previous experience that they do share data for convenient services (Williams, 2018). This, combined with a fear of the unknown, might have led to the small differences in means. The fear of losing sensitive data, paired with the potential gains in learning success that our participants envisioned, might also explain the overall inconclusive judgements regarding the idea of cognition-aware e-learning systems.

We also see that a strong link between the data disclosure readiness (A) and the required performance improvements (B) exists, indicating that the concerns are more of a general nature than specific to the context of cognition-aware e-learning, thus, also relevant for cognition-aware PE in adaptive CAT environments.

Misgivings about the idea of cognition-aware e-learning exist mainly with regard to data protection; however, many users reported no concerns if topics like transparency, security and opt-ins are properly addressed. This indicates that these aspects should be of the highest priority when implementing the concept in practice, both in e-learning and PE. We further found plenty of interesting adaptation ideas, reflecting the interest in the topic that was also expressed in the form of approval/praise. There is also a (non-significant) tendency that users with more e-learning experience have more positive feelings towards the idea of cognition-aware e-learning.

The main limitation of this work is that we only asked about data disclosure in a survey, without having tested a cognition-aware e-learning system in practice. Furthermore, we only sampled German participants, so cultural differences might occur for different countries.

9.4 Conclusion

This chapter explored which of the numerous approaches to capture the cognitive state would be well accepted by users, whether some sensors are more concerning than others, and what benefits must be achieved through cognition-awareness to make data disclosure worthwile. To answer these questions, we conducted an online survey with 50 participants. The results show that people would provide access to behavioral data like keyboard input without major concerns; however, other sensors are considered much more sensitive. Participants also appear less concerned about sensors that are integrated into consumer devices than about less widespread ones. Our results can guide practitioners developing cognition-aware systems to achieve broader user acceptance and show researchers which measuring methods are perceived as intrusive. Due to the separation of the analysis into general findings (A), and findings specific to e-learning (B), we were able to see that most of our results in fact are of a general nature and therefore should also apply to PE of MT. This can also be seen in the general opinions regarding cognition-aware e-learning and ideas on how best to adapt to the cognitive state, which are both similar to what we already reported for the PE context in chapter 6.

After detailed analysis of the suitability of different sensing modalities for making CAT tools or e-learning platforms cognition-aware, this final chapter again involved users to understand which monitoring approaches they would find suitable and concerning. We have seen that even though physiological measures work very well, they lead to more privacy concerns than behavioral sensors. This highlights the importance of our multi-modal approach, which was particularly beneficial to these measures.

Part IV

Automatic Post-Editing

As discussed before, an efficient cooperation between human and machine becomes more and more important with better MT output. Part II discussed how multi-modal input in the form of e.g., pen or speech interactions can support the PE process. Part III then presented approaches to estimate cognitive load and discussed how this can be used to better adapt to the user's current state. Finally, this part discusses another form of support, namely Automatic Post-Editing (APE). As introduced in the literature review (section 2.5), APE is a method that aims to automatically correct errors made by MT systems before performing actual human PE (Knight and Chander, 1994), thereby reducing the translators' workload and increasing productivity (Pal et al., 2016a).

We first present two novel APE architectures: the Multi-Source Transformer (chapter 10), which we submitted to WMT 2018, as well as the Transference architecture from WMT 2019, for which we conducted several additional analyses (chapter 11). Both approaches are multi-source approaches, mapping $\{src, mt\} \rightarrow pe$ which has been shown to provide further benefits compared to single-source ($mt \rightarrow pe$) approaches (Bojar et al., 2017, 2016, 2015). As an architecture, both build upon the Transformer architecture (Vaswani et al., 2017), which itself is built upon attention mechanisms (Bahdanau et al., 2014) completely replacing recurrence and convolutions. These attention mechanisms can be used to provide awareness of errors in mt originating from src, because they model non-local dependencies in the input or output sequences, and importantly also global dependencies between them (in our case src, mt and pe).

Finally, chapter 12 provides a general discussion how APE can be used to improve the human-AI collaboration, e.g., by adapting an MT engine to a domain or translator-specific style from a limited amount of data and without long retraining times.

Part IV is based on publications Pal et al. (2018), Pal et al. (2019), and Pal et al. (2020).

Chapter 10
The Multi-Source Transformer for Automatic Post-Editing

As discussed in section 2.5, Automatic Post-Editing (APE) is a method that learns from human post-edits to avoid repetitive mistakes of the MT, which are one of the main reasons why many translators' dislike PE (O'Brien and Moorkens, 2014). This chapter presents our English–German APE system submitted to the APE Task organized at WMT 2018 (Chatterjee et al., 2018). The proposed model is an extension of the Transformer architecture: two separate self-attention-based encoders encode the machine translation output (mt) and the source (src), followed by a joint encoder that attends over a combination of these two encoded sequences (enc_{src} and enc_{mt}) for generating the post-edited sentence. We compare this multi-source architecture (i.e, $\{src, mt\} \rightarrow pe$) to a monolingual Transformer (i.e., $mt \rightarrow pe$) model and an ensemble combining the multi-source $\{src, mt\} \rightarrow pe$ and single-source $mt \rightarrow pe$ models. We will first present the model architecture in detail and then report on the experiments conducted to understand its benefits.

This chapter is based on Pal et al. (2018).

10.1 Architecture

MT errors originating from the input source sentences suggest that APE systems should leverage information from both the src and mt, instead of considering mt in isolation. This can help the model to disambiguate corrections applied at every time step. Overall, we thus explore ***single-source*** ($\mathbf{mt} \rightarrow \mathbf{pe}$), ***multi-source*** ($\{\mathbf{src}, \mathbf{mt}\} \rightarrow \mathbf{pe}$), and an **ensemble** of these two models for APE.

All our models are based on the Transformer architecture (Vaswani et al., 2017), which provided the new state-of-the-art in 2017 and continues to be a widely used model. We extend the Transformer architecture to investigate how efficient this approach is in a multi-source scenario. In MT tasks, it was already shown that a Transformer can learn long-range dependencies. Therefore, we explore if we can leverage information from src and mt via a joint encoder through self-attention (see Section 10.1.2) to provide dependencies between src–mt that are then projected to the pe.

10.1.1 Single-Source Transformer for APE ($\mathbf{mt} \rightarrow \mathbf{pe}$)

Our single-source model (SS) is basically an encoder-decoder-based Transformer architecture (Vaswani et al., 2017), however, learning to transform mt to pe instead of a standard MT task that aims to transform src into a reference translation. Transformer models can replace sequence-aligned recurrence entirely and use three types of multi-head attention: encoder-decoder attention (also known as vanilla attention), encoder self-attention, and masked decoder self-attention. Since for multi-head attention each head uses different linear transformations, it can learn these separate relationships in parallel.

10.1.2 Multi-Source Transformer for APE ($\{\mathbf{src}, \mathbf{mt}\} \rightarrow \mathbf{pe}$)

For our multi-source model (MS), we propose a novel joint Transformer model (see Figure 10.1), which combines the encodings of src and mt and attends over a combination of both sequences while generating the post-edited sentence. Apart from enc_{src} and enc_{mt}, each of which is equivalent to the original Transformer's encoder (Vaswani et al., 2017), we use a joint encoder with an equivalent architecture, to maintain the homogeneity of the Transformer model. For this, we extend Vaswani et al. (2017) by introducing an additional identical encoding block by which both the enc_{src} and the enc_{mt} encoders communicate with the decoder.

Our multi-source neural APE computes intermediate states $\mathbf{enc_{src}}$ and $\mathbf{enc_{mt}}$ for the two encoders, $\mathbf{enc_{src,mt}}$ for their combination, and $\mathbf{dec_{pe}}$ for the decoder in sequence-to-sequence modeling. One self-attention based encoder for src maps $\mathbf{s} = (s_1, s_2, ..., s_k)$ into a sequence of continuous representations, $\mathbf{enc_{src}} = (e_1, e_2, ..., e_k)$, and a second encoder for mt, $\mathbf{m} = (m_1, m_2, ..., m_l)$, returns another sequence of continuous representations, $\mathbf{enc_{mt}} = (e_1^{'}, e_2^{'}, ..., e_l^{'})$. The self-attention based joint encoder (see Figure 10.1) then receives the concatenation of $\mathbf{enc_{src}}$ and $\mathbf{enc_{mt}}$, $\mathbf{enc_{concat}} = [\mathbf{enc_{src}}, \mathbf{enc_{mt}}]$ as an input, and passes it through the stack of 6 layers, with residual connections, normalization, self-attention, and a position-wise fully connected feed-forward neural network. As a result, the joint encoder produces a final representation ($\mathbf{enc_{src,mt}}$) for both src and mt. Self-attention at this point provides the advantage of aggregating information from all of the words, including src and mt, and successively generates a new representation per word informed by the entire src and mt context. The decoder

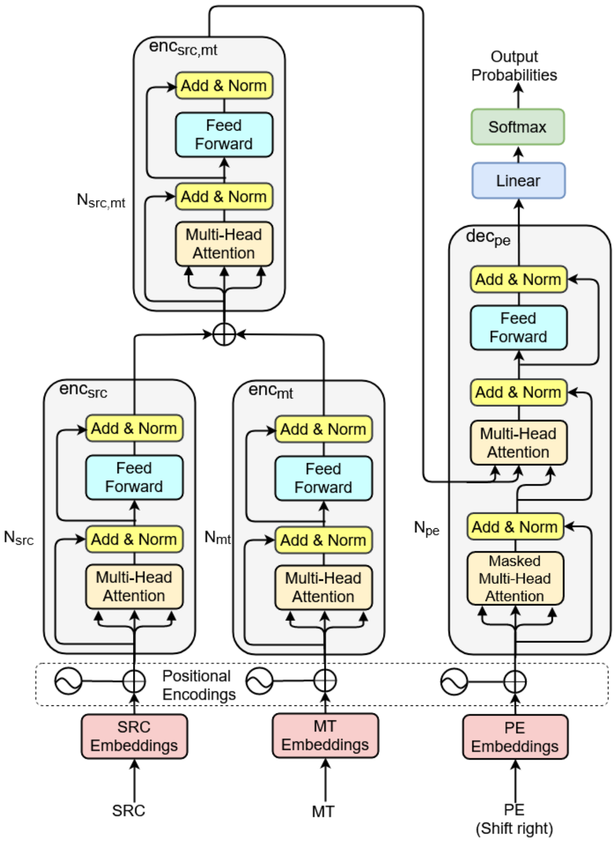


Figure 10.1: The *Multi-Source Transformer* model architecture for APE.

generates the pe output in sequence, $\mathbf{dec_{pe}} = (p_1, p_2, \ldots, p_n)$, one word at a time from left to right by attending previously generated words as well as the final representations ($\mathbf{enc_{src,mt}}$) generated by the encoder.

10.1.3 Ensemble

In order to leverage the network architecture for both single-source and multi-source APE as discussed above, we decided to **ensemble** several expert neural models. Each model is averaged using the 5 best saved checkpoints, which generate different translation outputs. Taking into account all these generated translation outputs, we implement an ensemble technique based on the frequency of occurrence of the output words: Corresponding to each input word, we calculate the most frequent occurrence of the output word from all the generated translation outputs. For the two different APE tasks (see section 10.2), we ensemble the following models:

- PBSMT task: We ensemble a SS ($mt \rightarrow pe$) and a MS ($\{src, mt\} \rightarrow pe$) average model.

- NMT task: We ensemble two average SS ($mt \rightarrow pe$) and MS ($\{src, mt\} \rightarrow pe$) models together with a SS and a MS model that are fine-tuned on a subset of the training set (see subsubsection 10.2.3).

10.2 Experiments

In our experiment we investigate (1) how well the Transformer-based APE architecture performs in general, (2) if our multi-source architecture using the additional joint encoder improves the performance over a single-source architecture, and (3) if ensembling of single-source and multi-source architectures facilitates APE even further.

10.2.1 Data

We explore our approach on both APE sub-tasks of WMT 2018 (Chatterjee et al., 2018), where the black box MT (we refer as 1st-stage MT) system to which APE is applied is either a Phrase-Based Statistical Machine Translation (PBSMT) or a Neural Machine Translation (NMT) model. For the PBSMT task, there is a total of 23K English–German APE data samples (11K from WMT 2016 and 12K from WMT 2017) (Bojar et al., 2017). For the NMT task, 13,442 samples of English–German APE data are provided.

All released APE data consists of English–German triplets containing source English text (src) from the IT domain, the corresponding German translations (mt) from a first stage MT system, and the corresponding human post-edited version (pe), all of them already tokenized. As this released APE dataset is small in size (see Table 10.1), additional resources are also available: first, the 'artificial training data' (Junczys-Dowmunt and Grundkiewicz, 2016) containing 4.5M sentences, 4M of which are weakly similar to the WMT 2016 training data, while 500K show very similar TER statistics; and second, the synthetic 'eSCAPE' APE corpus (Negri et al., 2018b), consisting of more than 7M triples for both NMT and PBSMT. More detail on these artifical datasets can be found in section 2.5.1.

Table 10.1 presents the statistics of the released data for the English–German APE task organized in WMT 2018. These datasets, except for the eSCAPE corpus, do not require any preprocessing in terms of encoding or alignment.

For cleaning the noisy eSCAPE dataset containing many unrelated language words (e.g., Chinese), we perform the following two steps: (i) we use the cleaning process described in Pal et al. (2015), and (ii) we execute the Moses (Koehn et al., 2007) corpus cleaning scripts with minimum and maximum number of tokens set to 1 and 80, respectively. After cleaning, we use the Moses tokenizer to tokenize the eSCAPE corpus. To handle out-of-vocabulary words, words are preprocessed into subword units (Sennrich et al., 2016) using Byte-Pair Encoding (BPE).

	Sentences				
	Corpus	**2016**	**2017**	**2018**	**Cleaning**
PBSMT	Train	12,000	11,000	-	-
	Dev	1,000	-	-	-
	Test	2,000	2,000	2,000	-
NMT	Train	-	-	13,442	-
	Dev	-	-	1,000	-
	Test	-	-	1,023	-
Additional Resources	Artificial	-	4M + 500K	-	-
	eSCAPE-PBSMT	-	-	7,258,533	6,521,736
	eSCAPE-NMT	-	-	7,258,533	6,485,507

Table 10.1: Statistics of the WMT 2018 APE shared task dataset.

10.2.2 Hyperparameter Settings

For the multi-source case ($\{\mathbf{src}, \mathbf{mt}\} \rightarrow \mathbf{pe}$), both the self-attended encoders, the joint encoder, and the decoder are composed of a stack of $N = 6$ identical layers. Each layer again consists of two sub-layers with normalization and a residual connection (He et al., 2016b) around each of the two sub-layers. During training, we employ label smoothing of value ϵ_{ls} = 0.1. The output dimension produced by all sub-layers and embedding layers is defined as $d_{model} = 256$. All dropout values in the network are set to 0.2. Each encoder and decoder contains a fully connected feed-forward network having dimensionality $d_{model} = 256$ for the input and output and dimensionality $d_{ff} = 1024$ for the inner layer. This is a similar setting to the original Transformer's (Vaswani et al., 2017) $C - model$. For the scaled dot-product attention, the input consists of queries and keys of dimension d_k, and values of dimension d_v. As multi-head attention parameters, we employ $h = 8$ for parallel attention layers, or heads. For each of these we use a dimensionality of $d_k = d_v = d_{model}/h = 32$. For optimization, we use the Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$. The learning rate is varied throughout the training process, first increasing linearly for the first training steps $warmup_{steps} = 4000$ and then adjusted as described in Vaswani et al. (2017).

At training time, the batch size is set to 32 samples, with a maximum sentence length of 80 subwords, and a vocabulary of the 50K most frequent subwords. After each epoch, the training data is shuffled. For encoding the word order, our model uses learned positional embeddings (Gehring et al., 2017), since Vaswani et al. (2017) reported nearly identical results to sinusoidal encodings. After finishing training, we save the 5 best checkpoints saved at each epoch. Finally, we use a single model obtained by averaging the last 5 checkpoints. During decoding, we perform greedy-search-based decoding.

We follow a similar hyperparameter setup for $mt \rightarrow pe$. The total number of parameters for our $\{src, mt\} \rightarrow pe$ and $mt \rightarrow pe$ model is 46×10^6 and 28×10^6.

10.2.3 Experiment Setup

This section presents the training process using the above datasets, to train single-source ($mt \rightarrow pe$), multi-source ($\{src, mt\} \rightarrow pe$), and ensemble models for both the PBSMT and NMT tasks.

PBSMT Task

For the PBSMT task, we first train both our SS and MS systems with the cleaned eSCAPE corpus for 3 epochs. We then perform transfer learning with 4M artificial data for 7 epochs. Afterwards, fine-tuning is performed using the 500K artificial and 23K real PE training data for another 20 epochs. Furthermore, we generate an ensemble model, by combining the 5-best-checkpoint model of SS with the 5-best-checkpoint model of MS.

We use the WMT 2016 development data (dev2016) containing 1,000 triplets to validate the model during training. To test our system performance, we use the WMT 2016 and 2017 test data (test2016, test2017), each containing 2,000 triplets. Furthermore, we report the results of the submitted ensemble model on test2018.

NMT Task

Initial tests for pre-training our NMT model on the NMT eSCAPE data showed no performance improvements. Therefore, we use the PBSMT SS and MS models as a basis for the NMT task. We use the PBSMT models after training them on the eSCAPE corpus, the 4M artificial data and the 500K + 23K train sets of WMT 16 and 17. These SMT-based models are then fine-tuned using the WMT 2018 NMT APE data (train18) for 60 epochs.

Afterwards, we perform an additional fine-tuning step towards the dev18/test18 dataset: For this, we extract sentences of train18 that are similar to the sentences contained in dev18/test18 and fine-train for another 15 epochs on this subset of train18, which we call fine-tune18. As a similarity measure we use the cosine similarity between the train src and mt segments and the test src and mt segments, respectively. These cosine similarities for src and mt are then simply multiplied to achieve an overall similarity measure. Our fine-tuning dataset contains only sentences with an overall similarity of at least 0.9.

Last, two separate ensemble models are created. One consists of only the non-fine-tuned SS and MS models, and one ensembles the SS and MS models in both fine-tuned and non-fine-tuned variants. Both ensembles are created by averaging over the 5 best checkpoints of each sub-model.

We report the results of all created models for the dev18 NMT dataset, and additionally those of the submitted overall ensemble model on test18.

10.3 Results & Discussion

This section presents and discusses the results on the PBSMT and NMT tasks.

10.3.1 Phrase-Based Statistical Machine Translation Task

Table 10.2 presents the results for the PBSMT APE task (see subsubsection 10.2.3), where two different Transformer-based models, one ensemble of these models and the baseline BLEU scores are shown. The baseline here refers to the original MT output evaluated with respect to the corresponding human PE translation. All models yield statistically significant results ($p < 0.001$) over this baseline. MS_{avg} also provides statistically significant improvement over SS_{avg}. For this and all following significance tests we employ the method by Clark et al. (2011)[67].

APE Systems	eScape	4M	500K	train16	train17	test16	test17	test18
Baseline	-					62.92	62.11	62.99
SS_{avg}	3 eps	7 eps	20 eps			66.27	66.60	-
MS_{avg}	3 eps	7 eps	20 eps			67.31	67.66	-
Ensemble	$MS_{avg\{5cps\}}$ + $SS_{avg\{5cps\}}$					**68.52**	**68.91**	**66.16**

Table 10.2: BLEU scores for all models on the WMT 2018 PBSMT task.

Generally, reasons for the good performance of our Transformer-based MS architecture in comparison to the SS approach for PBSMT-based APE could be the positional encoding that injects information about the relative or absolute position of the tokens in the sequence. This might help to handle word order errors in mt for a given src context. Another possible explanation lies in the self-attention mechanism, which handles local word dependencies for src, mt, and pe. After the individual dependencies are learned by the two encoders' self-attention mechanisms, another level of self-attention is performed that can jointly learn from both src and mt using our joint encoder, thereby informing the decoder about the long-range dependencies between the words within both src and mt. Compared to RNNs, we believe that this technique can better convey source information via mt to the decoder. The ensemble model then leverages the advantages of both our SS and MS approaches to further improve the results.

10.3.2 Neural Machine Translation Task

The results for our Transformer-based architecture for the NMT task are shown in Table 10.3. As can be seen, the baseline NMT system performs best, followed by the ensemble models, then the multi-source architectures and lastly the single-source approach. These differences between the three approaches, ensemble, MS, and SS, are all statistically significant. Fine-tuning provides some small, albeit insignificant, improvements over the non-fine-tuned versions.

[67] `https://github.com/jhclark/multeval`

APE Systems	Base Model	train18	fine-tune18	dev18	test18
Baseline	-	-	-	**76.66**	**74.73**
SS_{avg}	SS_{avg} (PBSMT)	60 eps	-	72.75	
MS_{avg}	MS_{avg} (PBSMT)	60 eps	-	74.84	-
SS_{ft}	SS_{avg} (NMT)	-	15 eps	73.17	-
MS_{ft}	MS_{avg} (NMT)	-	15 eps	75.05	-
$Ensemble$	$MS_{avg\{5cps\}}$ + $SS_{avg\{5cps\}}$			75.80	-
$Ensemble_{ft}$	$MS_{avg\{5cps\}}$ + $SS_{avg\{5cps\}}$ + $MS_{ft\{5cps\}}$ + $SS_{ft\{5cps\}}$			75.96	74.22

Table 10.3: BLEU scores for all models on the WMT 2018 NMT task.

Similar to many submissions to the NMT task of WMT 2018, our model did not improve over the strong NMT baseline (see subsubsection 2.5.2). Only two teams managed to statistically beat the baseline in terms of TER, and gained less than 0.4 TER point improvement. Reasons for those discouraging results as given by the task organizers were the very high MT quality as well the small data size of only about 13k training samples. Even though none of our architectures perform better than the baseline MT system for the NMT task, we clearly see that the multi-source approach helps, and that ensembling of different SS and MS models further increases the performance. These results are in line with our expectations, because intuitively, inspecting both src and mt should help detect and correct common errors. However, we are unsure why all of our models did not improve over the baseline, which could have been achieved by simply copying the mt. One reason might be the small amount of PE data, which comprises only 13K samples; this could also explain why the simple fine-tuning approach already leads to slightly higher BLEU scores. However, further human evaluation is necessary to better understand what our model is doing for the neural APE task and why it remains approximately 0.5 BLEU points below the baseline.

10.4 Conclusion

We investigated a novel Transformer-based multi-source APE approach that uses two encoders, a joint encoder, and a single decoder. Our model concatenates two separate self-attention-based encoders (enc_{src} and enc_{mt}) and passes this sequence through another self-attended joint encoder ($enc_{src,mt}$) to ensure capturing dependencies between src and mt. Finally, this joint encoder is fed to the decoder which follows a similar architecture as described in Vaswani et al. (2017). The entire model is optimized as a single end-to-end Transformer network.

This architecture yields statistically significant improvements over single-source Transformer-based models. An ensemble of both variants increases the performance further. For the PBSMT task, the baseline MT system was outperformed by 3.2 BLEU points, while the NMT baseline remains 0.51 BLEU points better than our APE approach on the 2018 test set. Thus, while our neural APE approach can improve over PBSMT baselines, it fails to improve (and even decreases) compared to state-of-the-art neural NMT.

This chapter contributed to our third research question by showing a practical implementation of a system that is able to learn from human post-edits to avoid repetitive mistakes, thereby making subsequent PE quicker and less repetitive. While strong improvements over the PBSMT baseline were shown, the inability to outperform the NMT system shows that further research is required. One option would be the exploration of different hyperparameter setups (e.g., the 'big' or the 'base' hyperparameter configuration in the original paper (Vaswani et al., 2017) or beam-search decoding, however, we instead focus on further improvements to the architecture itself, as discussed in the next chapter.

Chapter 11
The Transference Architecture for Automatic Post-Editing

In this chapter, we continue our search for APE architectures that best support the capture, preparation and provision of src and mt information and its integration with pe decisions by presenting a multi-source APE model, called *Transference*. Unlike previous approaches, it (i) uses a Transformer encoder block for src, (ii) followed by a decoder block, but without masking for self-attention on mt, which effectively acts as second encoder combining $src \rightarrow mt$, and (iii) feeds this representation into a final decoder block generating pe. We first submitted this architecture to WMT 2019, where we achieved 0.9 and 1.0 absolute BLEU points improvement on the development and test set. Our submission is on par with the winning approach while being simpler, as we do not use a BERT-based architecture. The results in comparison to other approaches can be found in the WMT 2019 findings (Barrault et al., 2019). After presenting the architecture and comparing it to the Multi-Source Transformer (see chapter 10), we conduct an experiment on both SMT and NMT data. Furthermore, we investigate the importance of our newly introduced second encoder and analyze the error types fixed by our model.

This chapter is based upon publications Pal et al. (2019) and Pal et al. (2020).

11.1 Architecture

As already argued for our previous architecture, errors in mt originating from src can be modelled using attention mechanisms (Bahdanau et al., 2014), which can capture non-local dependencies in the input or output sequences, and importantly also global dependencies between them (in our case src, mt and pe). We thus, again build upon the *Transformer* architecture (Vaswani et al., 2017), which uses positional encoding to encode the input and output sequences, and computes both self- and cross-attention through so-called multi-head attentions, which can be efficiently parallelized. Such multi-head attention allows to jointly attend to information at different positions from different representation subspaces, e.g., utilizing and combining information from src, mt, and pe.

We propose a multi-source Transformer model called *Transference* ($\{src, mt\}_{tr} \rightarrow pe$, Figure 11.1), which takes advantage of both the encodings of src and mt and attends over a combination of both sequences while generating the post-edited sentence. The second encoder, $enc_{src \rightarrow mt}$, makes use of the first encoder enc_{src} and a sub-encoder enc_{mt} for considering src and mt. Here, the enc_{src} encoder and the dec_{pe} decoder are equivalent to the original Transformer for neural MT (Vaswani et al., 2017). Our $enc_{src \rightarrow mt}$ follows an architecture similar to the Transformer's decoder, the difference being that no masked multi-head self-attention is used to process mt. We thus recombine the different blocks of the Transformer architecture and repurpose them for the APE task in a simple yet effective way. The name *Transference* was chosen as it describes how the second encoder conditionally learns the context dependencies from both src and mt and projects these to pe.

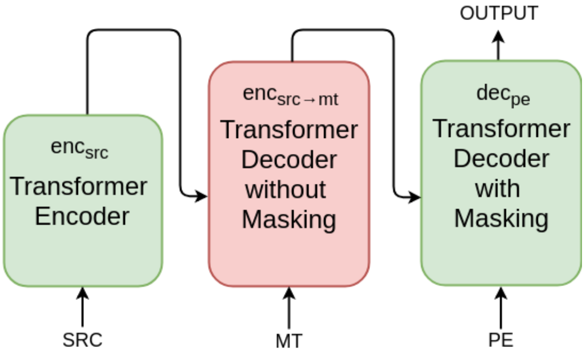

OUTPUT
enc_src
Transformer Encoder
enc_src→mt
Transformer Decoder without Masking
dec_pe
Transformer Decoder with Masking
SRC
MT
PE

Figure 11.1: Overview of the *Transference* model architecture for APE ($\{src, mt\}_{tr} \rightarrow pe$).

The suggested architecture is inspired by the two-step approach professional translators tend to use during PE: First, the source segment is compared to the corresponding translation suggestion (similar to what our $enc_{src \rightarrow mt}$ is doing), then corrections to the MT output are applied based on the encountered errors (in the same way that our dec_{pe} uses the encoded representation of $enc_{src \rightarrow mt}$ to produce the final translation).

Looking at the architecture in detail (see Figure 11.2), the self-attended encoder for src, s = $(s_1, s_2, \ldots, s_k)$ returns a sequence of continuous representations, enc_{src}, and the second self-attended sub-encoder for mt, $\mathbf{m} = (m_1, m_2, \ldots, m_l)$, returns another sequence of continuous representations, enc_{mt}. Self-attention at this point provides the advantage of aggregating information from all of the words, including src and mt. The internal enc_{mt} representation performs cross-attention over enc_{src} and prepares a final representation ($enc_{src \rightarrow mt}$) for the decoder (dec_{pe}). The decoder then generates the pe output in sequence, $\mathbf{p} = (p_1, p_2, \ldots, p_n)$, one word at a time from left to right by attending to previously generated words as well as the final encoder representation ($enc_{src \rightarrow mt}$).

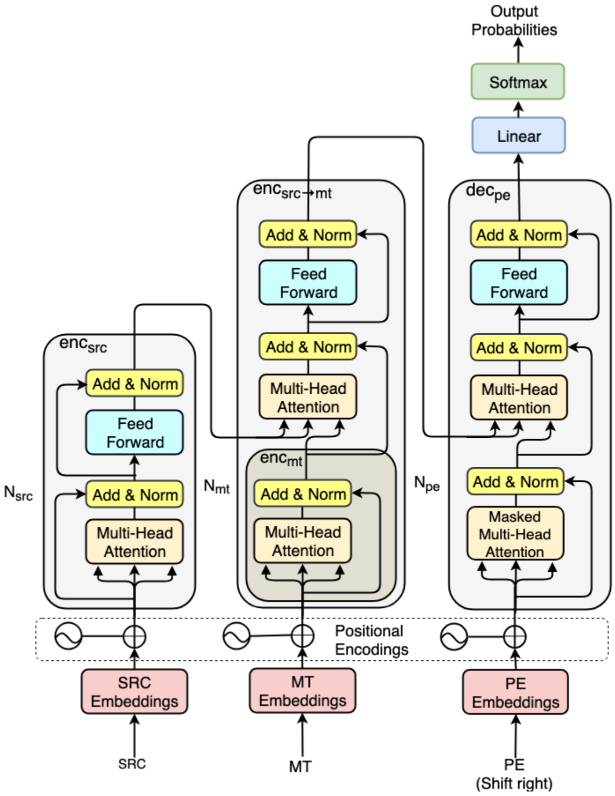


Figure 11.2: The *Transference* architecture for APE ($\{src, mt\}_{tr} \rightarrow pe$) in detail.

To summarize, our multi-source APE implementation extends Vaswani et al. (2017) by introducing an additional encoding block by which src and mt communicate with the decoder.

Comparison to Multi-Source Transformer The Multi-Source Transformer discussed in the last chapter uses 3 standard Transformer encoder blocks: one for src, one for mt, and one encoding the combined output of enc_{src} and enc_{mt} called $enc_{src,mt}$ (see Figure 10.1). In the Transference architecture, enc_{src} is identical to the Multi-Source Transformer, however, a Transformer decoder block (without masking) is used to encode mt and combine it with the output of enc_{src}, shown as $enc_{src \rightarrow mt}$ in Figure 11.2.

Comparison to $\mathbf{wmt18^{smt}_{best}}$ The Transference architecture also differs from the WMT 2018 PBSMT winning system ($wmt18^{smt}_{best}$, see Junczys-Dowmunt and Grundkiewicz (2018) in section 2.5.2) in several ways: (i) we use the original Transformer's decoder without modifications; (ii) one of our encoder blocks ($enc_{src \rightarrow mt}$) is identical to the Transformer's decoder block but uses no masking in the self-attention layer, thus, having one self-attention layer and an additional cross-attention for $src \rightarrow mt$; and (iii) in the decoder layer, the cross-attention is performed between the encoded representation from $enc_{src \rightarrow mt}$ and pe. Moreover, placing a cross-attention network within the $enc_{src \rightarrow mt}$ sub-layer rather than the dec_{pe} sub-layer as in $wmt18^{smt}_{best}$, during inference, $enc_{src \rightarrow mt}$ is forward propagated only once instead of multiple times i.e., once per decoding step.

Comparison to $\mathbf{wmt18^{nmt}_{best}}$ Our approach also differs from the WMT 2018 NMT winning system (Tebbifakhr et al., 2018) (see section 2.5.2): (i) $wmt18^{nmt}_{best}$ concatenates the encoded representation of two encoders and passes it as the key to the attention layer of the decoder, and (ii), the system additionally employs sequence-level loss functions based on maximum likelihood estimation and minimum risk training in order to avoid exposure bias during training.

Comparison to Libovický et al. In contrast to other multi-encoder based approaches and Libovický et al. (2018)'s approach (see section 2.5.2), where the authors focused on cross-attention of two encoders with respect to the Transformer decoder, we propose a novel architecture where the second encoder block is similar to the Transformer decoder block but without masking.

Comparison to $\mathbf{wmt19^{nmt}_{best}}$ Comparing with $wmt19^{nmt}_{best}$ (Lopes et al., 2019) (see section 2.5.2), the winning system of WMT 2019 uses a pre-trained deep bidirectional Transformer (multilingual BERT, Devlin et al. (2019)), while our model does not. $wmt19^{nmt}_{best}$ uses a single pre-trained BERT encoder that receives both the src and mt strings and applies a BERT-based encoder-decoder model. Additionally, they add a conservativeness penalty factor during beam decoding to avoid over-corrections in APE.

The main intuition is that our $enc_{src \rightarrow mt}$ attends over the src and mt and informs the pe to better capture, process, and share information between src-mt-pe, which efficiently models error patterns and the corresponding corrections. Our model performs better than past Transformer-based approaches and similar to the BERT-based approach ($wmt19^{nmt}_{best}$) without adding the overhead of the pre-trained model, as the experiment section will show.

11.2 Experiments

For the PBSMT task, we compare against four **baselines**: the **raw SMT** output provided by the 1st-stage PBSMT, the best-performing systems from WMT APE 2018 ($\mathbf{wmt18^{smt}_{best}}$), which are a single model and an ensemble model by Junczys-Dowmunt and Grundkiewicz (2018), as well as a Transformer directly translating from src to pe (**Transformer** ($\mathbf{src} \rightarrow \mathbf{pe}$)), thus, performing translation instead of APE. We evaluate the systems using BLEU (Papineni et al., 2002) and TER (Snover et al., 2006).

For the NMT task, we consider three **baselines**: the **raw NMT** output provided by the 1st-stage NMT system and the best-performing systems from WMT 2018 ($\mathbf{wmt18^{nmt}_{best}}$ by Tebbifakhr et al. (2018)) and WMT 2019 ($\mathbf{wmt19^{nmt}_{best}}$ by Lopes et al. (2019)).

Apart from the multi-encoder *Transference* architecture ($\{src, mt\}_{tr} \rightarrow pe$) and ensembling of this architecture, two simpler versions are also analyzed: first, a 'mono-lingual' ($\mathbf{mt} \rightarrow \mathbf{pe}$) APE model using only parallel mt–pe data and therefore only a single encoder, and second, an identical single-encoder architecture, however, using the concatenated src and mt text as input ($\{\mathbf{src} + \mathbf{mt}\} \rightarrow \mathbf{pe}$) (Niehues et al., 2016).

11.2.1 Data

As for the Multi-Source Transformer, we use the English–German WMT 2016 (Bojar et al., 2016), 2017 (Bojar et al., 2017), 2018 (Chatterjee et al., 2018) for our experiments. All these released APE datasets consist of English–German triplets containing source English text (src) from the IT domain, the corresponding German translations (mt) from a 1st-stage MT system, and the corresponding human post-edited version (pe). The 2019 version of the APE dataset released in WMT (Chatterjee et al., 2019) is the same as the WMT 2018 NMT data. As before, the dataset sizes are visualized in Table 10.1. Note that for WMT 2018 we only worked on the NMT sub-task and ignored the data for the PBSMT task.

Since the datasets are small in size, we again rely on the artificial datasets by Junczys-Dowmunt and Grundkiewicz (2016) containing 4.5M sentences as additional resources, 500K of which are very similar to the WMT 2016 training data according to TER statistics. For the NMT task, we further use the synthetic eScape APE corpus (Negri et al., 2018b), consisting of ~7M triples, which we clean using (i) the cleaning process described in Tebbifakhr et al. (2018), and (ii) we use the Moses (Koehn et al., 2007) corpus cleaning scripts with minimum and maximum number of tokens set to 1 and 100, respectively. After cleaning, we perform punctuation normalization, and then use the Moses tokenizer (Koehn et al., 2007) to tokenize the eScape corpus with 'no-escape' option. Finally, we apply true-casing. The cleaned version of the eScape corpus contains ~6.5M triplets. More information on these artificial datasets can be found in subsubsection 2.5.1.

11.2.2 Experiment Setup

To build models for the **PBSMT tasks** from 2016 and 2017, we first train a generic APE model using all the training data (4M + 500K + 12K + 11K). Afterwards, we fine-tune the trained model using the 500K artificial and 23K (12K + 11K) real PE training data. We use the WMT 2016 development data (dev2016) containing 1,000 triplets to validate the models during training. To test our system performance, we use the WMT 2016 and 2017 test data (test2016, test2017) as two sub-experiments, each containing 2,000 triplets (src, mt, and pe). We compare our model's performance with the four baseline systems described above: raw MT, $wmt18^{smt}_{best}$ single and ensemble, as well as Transformer ($src \rightarrow pe$).

Additionally, we check the performance of our model on the WMT 2018 **NMT APE task** (where unlike in previous tasks, the 1st-stage MT system is provided by NMT): For this, we explore two experimental setups: (i) we use the PBSMT task's APE model as a generic model which is then fine-tuned to a subset (12k) of the NMT data ($\{src, mt\}^{nmt}_{tr} \rightarrow pe^{gen,smt}$). One should note that it has been argued that the inclusion of SMT-specific data could be harmful when training NMT APE models (Junczys-Dowmunt and Grundkiewicz, 2018). (ii), we train a completely new generic model on the cleaned eScape data ($\sim$6.5M) along with a subset (12K) of the original training data released for the NMT task ($\{src, mt\}^{nmt}_{tr} \rightarrow pe^{gen,nmt}$). The aforementioned 12K NMT data are the first 12K of the overall 13.4K NMT data. The remaining 1.4K are used as validation data. The released development set (dev2018) is used as test data for our experiment, alongside the test2018, for which we could only obtain results for a few models by the WMT 2019 task organizers. We also explore an additional fine-tuning step of $\{src, mt\}^{nmt}_{tr} \rightarrow pe^{gen,nmt}$ towards the 12K NMT data (called $\{src, mt\}^{nmt}_{tr} \rightarrow pe^{ft}$), and a model averaging the 8 best checkpoints of $\{src, mt\}^{nmt}_{tr} \rightarrow pe^{ft}$, which we call $\{src, mt\}^{nmt}_{tr} \rightarrow pe^{ft}_{avg}$.

During PE, professional translators have to understand the source, and analyze if the MT correctly represents the source, which corresponds to our enc_{src} and $enc_{src \rightarrow mt}$. To investigate whether following this realistic understanding of the post-editing process is beneficial for APE, we compare the model to a version with **swapped inputs** (mt, src), called $\{mt, src\}^{smt}_{tr} \rightarrow pe^{gen}$. We carried out an experiment with the PBSMT task's APE dataset. Moreover, we fine-tune the $\{mt, src\}^{smt}_{tr} \rightarrow pe^{gen}$ model with 500K artificial and 23K real PE training data and compare the fine-tuned model ($\{mt, src\}^{smt}_{tr} \rightarrow pe^{ft}$) with $\{src, mt\}^{smt}_{tr} \rightarrow pe^{ft}$.

Last, we analyze the importance of our new second encoder ($enc_{src \rightarrow mt}$), compared to the source encoder (enc_{src}) and the decoder (dec_{pe}), by reducing and expanding the **amount of layers** in the encoders and the decoder. Our standard setup, which we use for fine-tuning, ensembling etc., is fixed to 6-6-6 for N_{src}-N_{mt}-N_{pe} (see Figure 11.2), where 6 is the value that was proposed by Vaswani et al. (2017) for the *base* model. We investigate what happens in terms of APE performance if we change this setting to 6-6-4 and 6-4-6.

As for the Multi-Source Transformer, our model operates on subword units (Sennrich et al., 2016) by using Byte-Pair Encoding (BPE), thereby avoiding out-of-vocabulary words and reducing the vocabulary size. In the preprocessing step, instead of learning an explicit mapping between BPEs in the src, mt and pe, we define BPE tokens by jointly processing all triplets. Thus, src, mt and pe derive a single BPE vocabulary. Since mt and pe belong to the same language (German) and src is a close language (English), they naturally share a good fraction of BPE tokens, which reduces the vocabulary size to 28k. We implemented our approach based on the Neutron implementation of the Transformer (Xu and Liu, 2019)[68].

11.2.3 Hyperparameter Settings

We follow a similar hyperparameter setup for all reported systems: All encoders (for $\{src, mt\}_{tr} \rightarrow pe$), and the decoder, are composed of a stack of $N_{src} = N_{mt} = N_{pe} = 6$ identical layers (except for the layer experiment) followed by layer normalization. The learning rate is varied throughout the training process, and increasing for the first training steps $warmup_{steps} = 8000$ and afterwards decreasing as described in Vaswani et al. (2017). All remaining hyperparameters are set analogously to those of the Transformer's *base* model. At training time, the batch size is set to 25K tokens, with a maximum sentence length of 256 subwords. After each epoch, the training data is shuffled. During decoding, we perform beam search with a beam size of 4. As described above, we use shared embeddings between mt and pe in all our experiments.

11.3 Results

For the PBSMT task, the results of our models, ***single-source*** ($\mathbf{mt} \rightarrow \mathbf{pe}$), ***multi-source single encoder*** ($\{\mathbf{src} + \mathbf{pe}\} \rightarrow \mathbf{pe}$), ***Transference*** model ($\{\mathbf{src}, \mathbf{mt}\}^{\mathbf{smt}}_{\mathbf{tr}} \rightarrow \mathbf{pe}$) and its alternative with ***swapped inputs*** ($\{\mathbf{mt}, \mathbf{src}\}^{\mathbf{smt}}_{\mathbf{tr}} \rightarrow \mathbf{pe}$), and ***ensemble***, in comparison to the four baselines, ***raw SMT***, $\mathbf{wmt18}^{\mathbf{smt}}_{\mathbf{best}}$ (Junczys-Dowmunt and Grundkiewicz, 2018) single and ensemble, as well as ***Transformer*** ($\mathbf{src} \rightarrow \mathbf{pe}$), are presented in Table 11.1 for test2016 and test2017.

For the NMT task, Table 11.2 reports the results obtained by our ***Transference*** model ($\{\mathbf{src}, \mathbf{mt}\}^{\mathbf{nmt}}_{\mathbf{tr}} \rightarrow \mathbf{pe}$) on the WMT 2018, 2019 NMT data for dev2018 (which we use as a test set) and test2018/2019 (where results were obtained by the organizers), compared to the baselines ***raw NMT***, $\mathbf{wmt18}^{\mathbf{nmt}}_{\mathbf{best}}$, and $\mathbf{wmt19}^{\mathbf{nmt}}_{\mathbf{best}}$.

11.3.1 Baselines

The **raw SMT** output in Table 11.1 is a strong black-box PBSMT system (i.e., 1st-stage MT). We report its performance observed with respect to the ground

[68] `https://github.com/anoidgit/transformer`.

#	Models	test2016		test2017	
		BLEU ↑	TER ↓	BLEU ↑	TER ↓
Baselines					
1.1	Raw SMT	62.11	24.76	62.49	24.48
1.2	Transformer ($src \rightarrow pe$)	56.59 (-5.52)	29.97 (+5.21)	53.06 (-9.43)	32.20 (+7.72)
1.3	$wmt18^{smt}_{best}$ (single)	70.86 (+8.75)	18.92 (-5.84)	69.72 (+7.23)	19.49 (-4.99)
1.4	$wmt18^{smt}_{best}$ (x4)	**71.04** (+8.93)	**18.86** (-5.9)	**70.46** (+7.97)	**19.03** (-5.45)
Baselines: Retrained $\mathbf{wmt18^{smt}_{best}}$ with our experimental setup					
1.5	$wmt18^{smt,gen}_{best}$ (single)	69.14 (+7.03)	20.41 (-4.35)	68.14 (+5.65)	20.98 (-3.5)
1.6	$wmt18^{smt,ft}_{best}$ (single)	70.12 (+8.01)	19.84 (-4.92)	69.16 (+6.67)	20.34 (-4.14)
General models trained on 23K+4.5M data					
2.1	$mt \rightarrow pe$	67.70 (+5.59)	21.90 (-2.86)	66.91 (+4.42)	22.32 (-2.16)
2.2	$\{src + mt\} \rightarrow pe$	69.32 (+7.21)	20.27 (-4.49)	68.26 (+5.77)	20.90 (-3.58)
2.3	$\{src, mt\}^{smt}_{tr} \rightarrow pe$	70.46 (+8.35)	19.21 (-5.55)	70.05 (+7.56)	19.46 (-5.02)
2.4	$\{mt, src\}^{smt}_{tr} \rightarrow pe$	70.26 (+8.15)	19.34 (-5.42)	69.34 (+6.85)	20.05 (-4.43)
Fine-tuning Exp. 2 models with 23K+500K data					
3.1	$mt \rightarrow pe$	68.43 (+6.32)	21.29 (-3.47)	67.78 (+5.29)	21.63 (-2.85)
3.2	$\{src + mt\} \rightarrow pe$	69.87 (+7.76)	19.94 (-4.82)	68.57 (+6.08)	20.68 (-3.8)
3.3	$\{src, mt\}^{smt}_{tr} \rightarrow pe^{ft}$	71.05 (+8.94)	19.05 (-5.71)	70.33 (+7.84)	19.23 (-5.25)
3.4	$\{mt, src\}^{smt}_{tr} \rightarrow pe^{ft}$	70.26 (+8.15)	19.40 (-5.36)	69.31 (+6.82)	19.91 (-4.57)
4.1	Exp3.3$^{smt}_{ens4ckpt}$	**71.59** (+9.48)	**18.78** (-5.98)	**70.89** (+8.4)	**18.91** (-5.57)
4.2	$ensemble^{smt}(x3)$	**72.19** (+10.08)	**18.39** (-6.37)	**71.58** (+9.09)	**18.58** (-5.90)
$\mathbf{\{src, mt\}^{smt}_{tr} \rightarrow pe}$ with different layer size					
5.1	$\{src, mt\}^{smt}_{tr} \rightarrow pe$ 6-6-4	70.85 (+8.74)	19.00 (-5.76)	69.82 (+7.33)	19.67 (-4.81)
5.2	$\{src, mt\}^{smt}_{tr} \rightarrow pe$ 6-4-6	69.93 (+7.82)	19.70 (-5.06)	69.61 (+7.12)	19.68 (-4.8)

Table 11.1: Evaluation results on the WMT APE test set 2016, and test set 2017 for the **PBSMT task**; ($\pm X$) value is the improvement over Raw SMT. The last section of the table shows the impact of increasing and decreasing the depth of the encoders and the decoder.

truth (pe), i.e., the post-edited version of mt. The original PBSMT system scores over 62 BLEU points and below 25 TER on test2016 and test2017.

Using a **Transformer (**$src \rightarrow pe$**)**, we test if APE is really useful, or if potential gains are only achieved due to the good performance of the Transformer architecture. While we cannot do a full training of the Transformer on the data that the raw MT engine was trained on due to the unavailability of the data, we use our PE datasets in an equivalent experimental setup as for all other models. The results of this system (Exp. 1.2 in Table 11.1) show that the performance is actually lower across both test sets, -5.52/-9.43 absolute points in BLEU and +5.21/+7.72 absolute in TER, compared to the raw SMT baseline.

We report four results from $\mathbf{wmt18^{smt}_{best}}$, (i) $wmt18^{smt}_{best}$ ($single$), which is the core multi-encoder implementation without ensembling but with checkpoint averaging, (ii) $wmt18^{smt}_{best}$ ($x4$) which is an ensemble of four identical 'single' models trained with different random initializations. The results of $wmt18^{smt}_{best}$ ($single$) and $wmt18^{smt}_{best}$ ($x4$) (Exp. 1.3 and 1.4) reported in Table 11.1 are from Junczys-Dowmunt and Grundkiewicz (2018). Since their training procedure slightly differs from ours, we also trained the $wmt18^{smt}_{best}$ system using exactly our experimental setup in order to make a fair comparison. This yields the baselines (iii) $wmt18^{smt,gen}_{best}$ ($single$) (Exp. 1.5), which is similar to $wmt18^{smt}_{best}$ ($single$), however, the training parameters and data are kept in line with our *Transference* general model (Exp. 2.3) and (iv) $wmt18^{smt,ft}_{best}$ ($single$) (Exp. 1.6), which is also trained maintaining the equivalent experimental setup compared to the fine tuned version of the *Transference* general model (Exp. 3.3). Compared to both raw SMT and Transformer ($src \rightarrow pe$) we see strong improvements for this state-of-the-art model, with BLEU scores of at least 68.14 and TER scores of at most 20.98 across the PBSMT testsets. $wmt18^{smt}_{best}$, however, performs better in its original setup (Exp. 1.3 and 1.4) compared to our experimental setup (Exp. 1.5 and 1.6).

The results on the WMT 2018 and 2019 NMT datasets (dev2018 and test2018) are presented in Table 11.2. The *raw NMT* system serves as one baseline against which we compare the performance of the different models. We evaluate the system hypotheses with respect to the ground truth (pe), i.e., the post-edited version of mt. The baseline original NMT system scores 76.76 BLEU points and 15.08 TER on dev2018, and 74.73 BLEU points and 16.80 TER on test2018.

#	Models	dev2018		test2018	
		BLEU ↑	TER ↓	BLEU ↑	TER ↓
Baselines					
6.1	Raw NMT	76.76	15.08	74.73	16.80
6.2	$wmt18^{nmt}_{best}$	**77.74 (+0.98)**	14.78 (-0.30)	75.53 (+0.80)	16.46 (-0.34)
6.3	$wmt19^{nmt}_{best}$	-	-	**75.96 (+1.23)**	**16.06 (-0.74)**
Fine-tuning Exp. 3.3 on 12k NMT data					
7	$\{src, mt\}^{nmt}_{tr} \rightarrow pe^{gen,smt}$	77.09 (+0.33)	14.94 (-0.14)	-	-
Transference model trained on eScape+ 12k NMT data					
8	$\{src, mt\}^{nmt}_{tr} \rightarrow pe^{gen,nmt}$	77.25 (+0.49)	14.87 (-0.21)	-	-
Fine-tuning model 8 on 12k NMT data					
9	$\{src, mt\}^{nmt}_{tr} \rightarrow pe^{ft}$	77.39 (+0.63)	14.71 (-0.37)	-	-
Averaging 8 checkpoints of Exp. 9					
10	$\{src, mt\}^{nmt}_{tr} \rightarrow pe^{ft}_{avg}$	77.67 (+0.91)	**14.52 (-0.56)**	75.75 (+1.02)	16.15 (-0.69)

Table 11.2: Evaluation results on the WMT APE 2018 development set for the **NMT task** (Exp. 6 and Exp. 10 results were obtained by the WMT 2019 task organizers). ($\pm X$) value is the improvement over Raw NMT.

11.3.2 Single-Encoder Transformer for APE

The two architectures $\mathbf{mt} \rightarrow \mathbf{pe}$ and $\{\mathbf{src} + \mathbf{mt}\} \rightarrow \mathbf{pe}$ use only a single encoder. Table 11.1 shows that $\mathbf{mt} \rightarrow \mathbf{pe}$ (Exp. 2.1) provides better performance (+4.42 absolute BLEU on test2017) compared to the original SMT, while $\{\mathbf{src} + \mathbf{mt}\} \rightarrow \mathbf{pe}$ (Exp. 2.2) provides further improvements by additionally using the src information. $\{\mathbf{src} + \mathbf{mt}\} \rightarrow \mathbf{pe}$ improves over $\mathbf{mt} \rightarrow \mathbf{pe}$ by +1.62/+1.35 absolute BLEU points on test2016/test2017. After fine-tuning, both single encoder Transformers (Exp. 3.1 and 3.2 in Table 11.1) show further improvements, +0.87 and +0.31 BLEU points, respectively, for test2017 and a similar improvement for test2016.

11.3.3 Transference Transformer for APE

SMT Results In contrast to the two models above, our *Transference* architecture uses multiple encoders. The fine-tuned version of the $\{src, mt\}_{tr}^{smt} \rightarrow pe$ model (Exp. 3.3 in Table 11.1) outperforms $wmt18_{best}^{smt}$ (single) (Exp. 1.3) in BLEU on both test sets, however, the TER score for test2016 increases. When ensembling the 4 best checkpoints of our $\{src, mt\}_{tr}^{smt} \rightarrow pe$ model (Exp. 4.1), the result beats the $wmt18_{best}^{smt}$ (x4) system, which is an ensemble of four different randomly initialized $wmt18_{best}^{smt}$ (single) systems. Our $\mathbf{ensemble^{smt}(x3)}$ combines two $\{src, mt\}_{tr}^{smt} \rightarrow pe$ (Exp. 2.3) models initialized with different random weights with the ensemble of the fine-tuned Transference model Exp3.3$_{ens4ckpt}^{smt}$(Exp. 4.1). This ensemble provides the best results for all datasets, providing roughly +1 BLEU point and -0.5 TER when comparing against $wmt18_{best}^{smt}$ (x4). In terms of the number of parameters, $wmt18_{best}^{smt}$ and our $\{src, mt\}_{tr}^{smt} \rightarrow pe$ model are the same. Moreover, our $\{src, mt\}_{tr}^{smt} \rightarrow pe$ model uses a single multi-head cross-attention in the decoder sub-layer, compared to two multi-head cross-attention mechanisms in $wmt18_{best}^{smt}$, therefore our model is 1.07 times faster for post-editing the testset of 2000 sentences. Furthermore, using more non-autoregressive encoder layers with fewer autoregressive decoder layers can significantly accelerate the inference (Xu et al., 2020). Instead of aggregating src and mt with the autoregressive pe decoder as proposed by Junczys-Dowmunt and Grundkiewicz (2018), our approach that aggregates src and mt with the non-autoregressive mt encoder is significantly faster than the $wmt18_{best}^{smt}$ in inference.

Swapping Inputs Additionally we compare our $\{src, mt\}_{tr}^{smt} \rightarrow pe$ model with $\{mt, src\}_{tr}^{smt} \rightarrow pe$, where we reverse the input order, i.e., enc_1 and enc_2 take mt and src, respectively, as input. Exp. 2.4 and Exp. 3.4 report $\{mt, src\}_{tr}^{smt} \rightarrow pe$ and $\{mt, src\}_{tr}^{smt} \rightarrow pe^{ft}$ respectively, which performed slightly worse than $\{src, mt\}_{tr}^{smt} \rightarrow pe$ and $\{src, mt\}_{tr}^{smt} \rightarrow pe^{ft}$. Surprisingly, fine-tuning does not help $\{mt, src\}_{tr}^{smt} \rightarrow pe^{ft}$ for the testset 2016, however, in case of testset 2017, fine-tuning shows small gain in performance. Moreover, the performance gain in fine-tuning for the case of $\{src, mt\}_{tr}^{smt} \rightarrow pe^{ft}$ over $\{src, mt\}_{tr}^{smt} \rightarrow pe$ is considerably stronger than the performance gain for $\{mt, src\}_{tr}^{smt} \rightarrow pe^{ft}$ over $\{mt, src\}_{tr}^{smt} \rightarrow pe$. Empirically, this confirms our hypothesis that our model ($\{src, mt\}_{tr}^{smt} \rightarrow pe$) following translators' two-step approach to PE is beneficial: first, the source segment is compared to the corresponding translation suggestion, then corrections to the MT output are applied based on the encountered errors.

NMT Results For the WMT 2018 NMT data we first test our $\{src, mt\}_{tr}^{nmt} \rightarrow pe^{gen,smt}$ model, which is the model from Exp. 3.3 fine-tuned towards NMT data as described in Section 11.2.2. Table 11.2 shows that our PBSMT APE model fine-tuned towards NMT (Exp. 7) can even slightly improve over the already very strong NMT system by about +0.3 BLEU and -0.1 TER, although these improvements are not statistically significant.

The overall results improve when we train our model on eScape and NMT data instead of using the PBSMT model as a basis. Our proposed generic *Transference* model (Exp. 8, $\{src, mt\}_{tr}^{nmt} \rightarrow pe^{gen,nmt}$) shows statistically significant improvements in terms of BLEU and TER compared to the baseline even before fine-tuning, and further improvements after fine-tuning (Exp. 9, $\{src, mt\}_{tr}^{nmt} \rightarrow pe^{ft}$). Finally, after averaging the 8 best checkpoints, our $\{src, mt\}_{tr}^{nmt} \rightarrow pe_{avg}^{ft}$ model (Exp. 10) also shows consistent improvements in comparison to the baseline and other experimental setups. Overall our fine-tuned model averaging the 8 best checkpoints achieves +1.02 absolute BLEU points and -0.69 absolute TER improvements over the baseline on test2018. Table 11.2 also shows the performance of our model compared to the winner system of WMT 2018 ($wmt18_{best}^{nmt}$) for the NMT task (Tebbifakhr et al., 2018). $wmt18_{best}^{nmt}$ scores 14.78 in TER and 77.74 in BLEU on the dev2018 and 16.46 in TER and 75.53 in BLEU on the test2018. In comparison to $wmt18_{best}^{nmt}$, our model (Exp. 10) achieves better scores in TER on both the dev2018 and test2018, however, in terms of BLEU our model scores slightly lower for dev2018, while some improvements are achieved on test2018. Compared to $wmt19_{best}^{nmt}$ (Exp. 6.3), our model scores slightly lower, however, the performance loss is not statistically significant. It is to be noted that the training strategy in $wmt19_{best}^{nmt}$ is different: (i) they used their own synthetic corpus prepared using the parallel data provided by the Quality Estimation shared task[69], (ii) they oversampled the APE training data 20 times, and (iii) they applied multilingual BERT.

Amount of Layers The number of layers (N_{src}-N_{mt}-N_{pe}) in all encoders and the decoder for these results is fixed to 6-6-6. In Exp. 5.1, and 5.2 in Table 11.1, we see the results of changing this setting to 6-6-4 and 6-4-6. This can be compared to the results of Exp. 2.3, since no fine-tuning or ensembling was performed for these three experiments. Exp. 5.1 shows that decreasing the number of layers on the decoder side does not hurt the performance. In fact, in the case of test2016, we got some improvement, while for test2017, the scores got slightly worse. In contrast, reducing the $enc_{src\rightarrow mt}$ encoder block's depth (Exp. 5.2) does reduce the performance for all four scores, showing the importance of this second encoder.

11.3.4 Analysis of Error Patterns

In Table 11.3, we analyze and compare the best performing SMT ($ensemble^{smt}(x3)$) and NMT ($\{src, mt\}_{tr}^{nmt} \rightarrow pe_{avg}^{ft}$) model outputs with the original MT outputs on the WMT 2017 (SMT) APE test set and on the WMT 2018 (NMT) devset. Improvements are measured in terms of number of words which need to be (i) inserted (*In*), (ii) deleted (*De*), (iii) substituted (*Su*), and (iv) shifted (*Sh*), as per TER (Snover et al., 2006), in order to turn the MT outputs into reference translations. Our model provides promising results by significantly reducing

[69] http://www.statmt.org/wmt19/qe-task.html

the required number of edits (24% overall for PBSMT task and 3.6% for NMT task) across all edit operations, thereby leading to reduced PE effort and hence improving human PE productivity.

	%*In*	**%*De***	**%*Su***	**%*Sh***
$ensemble^{smt}(x3)$ vs. *raw SMT*	+31	+29	+15	+32
$\{src, mt\}_{tr}^{nmt} \rightarrow pe_{avg}^{ft}$ vs. *raw NMT*	+6	+2	+4	-2

Table 11.3: % of error reduction in terms of different edit operations achieved by our best systems compared to the raw MT baselines.

When comparing PBSMT to NMT, we see that stronger improvements are achieved for PBSMT, probably because raw SMT is worse than raw NMT. For PBSMT, similar results are achieved for *In*, *De*, and *Sh*, while less gains are obtained in terms of *Su*. For NMT, *In* is improved most, followed by *Su*, *De*, and last *Sh*. For shifts in NMT, the APE system even creates further errors, instead of reducing them, which is an issue we aim to prevent in the future.

11.4 Discussion

The proposed multi-encoder based Transformer architecture ($\{src, mt\}_{tr}^{smt} \rightarrow pe$, Exp. 2.3) shows slightly worse results than $wmt18_{best}^{smt}$ (single) (Exp. 1.3) before fine-tuning, and roughly similar results after fine-tuning (Exp. 3.3). After ensembling, however, our *Transference* model (Exp. 4.2) shows consistent improvements when comparing against the best baseline ensemble $wmt18_{best}^{smt}$ (x4) (Exp. 1.4). Due to the unavailability of the sentence-level scores of $wmt18_{best}^{smt}$ (x4), we could not test if the improvements (roughly +1 BLEU, -0.5 TER) are statistically significant. Interestingly, our approach of taking the model optimized for PBSMT and fine-tuning it to the NMT task (Exp. 7) does not hurt the performance as was reported in the previous literature (Junczys-Dowmunt and Grundkiewicz, 2018). In contrast, some small, albeit statistically insignificant improvements over the raw NMT baseline were achieved. When we train the *Transference* architecture directly for the NMT task (Exp. 8), we get slightly better and statistically significant improvements compared to raw NMT. Fine-tuning this NMT model further towards the actual NMT data (Exp. 9), as well as performing checkpoint averaging using the 8 best checkpoints improves the results even further. Compared to $wmt18_{best}^{smt}$ and $wmt19_{best}^{nmt}$, our architecture is simpler, faster during inference, it follows the two-step approach of professional post-editors, and has no additional overhead like BERT.

The reasons for the effectiveness of our approach can be summarized as follows: (1) Our $enc_{src \rightarrow mt}$ contains two attention mechanisms: one is self-attention and another is cross-attention. The self-attention layer is not masked here; therefore, the cross-attention layer in $enc_{src \rightarrow mt}$ is informed by both previous and future time-steps from the self-attended representation of mt (enc_{mt}) and additionally from enc_{src}. As a result, each state representation of $enc_{src \rightarrow mt}$ is learned from the context of src and mt. This might produce better representations for dec_{pe} which can access the combined context. In contrast, in $wmt18^{smt}_{best}$, the dec_{pe} accesses representations from src and mt independently, first using the representation from mt and then using that of src. (2) The position-wise feed-forward layer in $enc_{src \rightarrow mt}$ of our model requires processing information from two attention modules, while in the case of $wmt18^{smt}_{best}$, the position-wise feed-forward layer in dec_{tgt} needs to process information from three attention modules, which may increase the learning difficulty of the feed-forward layer. (3) Since pe is a post-edited version of mt, sharing the same language, mt and pe are quite similar compared to src. Therefore, attending over a fine-tuned representation from mt along with src, which is what we have done in this work, might be a reason for the better results than those achieved by attending over src directly.

Evaluating the influence of the depth of our encoders and decoder shows that while the decoder depth appears to have limited importance, reducing the encoder depth indeed hurts performance, which is in line with Domhan (2018).

11.5 Conclusion

We presented a multi-encoder Transformer-based APE model that repurposes the standard Transformer blocks in a simple and effective way for the APE task: Our *Transference* architecture uses (i) a source encoder (enc_{src}) which encodes src information, followed by (ii) a second encoder ($enc_{src \rightarrow mt}$) which can also be viewed as a standard Transformer decoding block, however, without masking, and (iii) a decoder (dec_{pe}) which captures the final representation from $enc_{src \rightarrow mt}$ via cross-attention. The proposed model outperforms the best-performing system of WMT 2018 on the test2016, test2017, dev2018, and test2018 data. Moreover, our model is on par with but simpler than the WMT 2019 best system since our model does not apply BERT or any conservative factor during inference.

Taking a departure from traditional Transformer-based encoders, which perform self-attention only, our second encoder also performs cross-attention to produce representations for the decoder based on both src and mt. We also show that the encoder plays a more pivotal role than the decoder in Transformer-based APE, which could also be the case for Transformer-based generation tasks in general. Our architecture is generic and can be used for any multi-source task, e.g., (i) multi-source translation, (ii) document translation to model the associated context, (iii) question generation to generate question from given passage and a short answer text, (iv) question answering from given passage and question text, (v) summarization, etc.

Overall, this chapter showed a practical solution how errors in MT systems can be corrected by automatically learning from human post-edits. It thus contributes to our third research question, since the reduced amount of errors can improve productivity and the avoidance of repetitive errors can reduce feelings of assembly line work, thereby strengthening acceptance of PE. One should also note the ability of APE systems to learn from a rather small set of (real) post-edits compared to the training corpora used for MT systems. This shows that already a limited amount of data gathered in a certain domain or by a certain translator can be leveraged for APE. The next chapter will consider more practical usage of APE by going into detail on topics like continuous retraining or discussing how the amount of training data impacts APE results.

Chapter 12
Improving the Post-Editing Process Through Automatic Post-Editing

After having presented two state-of-the-art APE models, the question remains how APE can be best integrated to support the PE process. This chapter discusses studies how strong performance gains through APE are, reviews works exploring APE for domain adaptation, presents approaches to interactively learn from post-edits, and shows how the amount of training data impacts APE performance.

In contrast to previous chapters, this chapter is not based on own publications, but instead reviews recent works that help understand how APE systems like our Multi-Source Transformer or Tranference models can be employed to improve the PE process.

12.1 Practical Efficiency Gains Through Automatic Post-Editing

The main motivation for APE is to reduce human PE effort by automatically correcting errors before showing MT output to humans (Bojar et al., 2015). Recently, Wang et al. (2020) practically verified this with professional translators. The model used for this study is a hierarchical approach combining two APE algorithms (an atomic and a generative model), which are conditionally employed depending on the quality assumed by running the sample first through a QE model. Here, the atomic APE model is deployed on the presumably high-quality MT, while the generative approach, which can apply stronger edits like paraphrasing, is applied on the presumably lower quality MT. Their results of a between-subjects human evaluation comparing PE MT to PE APE shows that humans were on average 26.3% faster for PE APE. However, this finding is limited by the fact that this test was run on WMT 2017 data, stemming from a PBSMT system. Thus, similar investigations with state-of-the-art NMT and APE models should be conducted to analyze the extent to which error reductions achieved through APE yield PE efficiency gains.

12.2 Automatic Post-Editing for Domain Adaptation

Another potential advantage of APE as already pointed out by the WMT 2015 organizers (Bojar et al., 2015) is its ability to adapt MT output to a certain style as desired by a translator or required for a domain. However, all share tasks up to WMT 2019 used a domain-optimized MT system and explored if APE can improve this specialized output further, thereby making investigations of domain adaptation through APE impossible.

In contrast, the recent WMT APE shared tasks from 2020 and 2021 (Chatterjee et al., 2020; Akhbardeh et al., 2021) investigated how well adaptation of general purpose MTs towards a specific domain works with APE approaches. Even though the amount of real data provided was again rather limited (7000 triples in each task), participants of 2020 were able to significantly improve the general purpose MT on the domain data (both in terms of automatic and human evaluation), which shows how APE could be employed in the background of CAT tools to improve the PE experience. With the stronger baseline MT of 2021, all submissions were still able to significantly outperformed the MT, though only in terms of human evaluation. APE thus provides a suitable alternative to other MT adaptation techniques used by Language Service Providers (LSPs) to offer translators the best MT results for different language pairs and domains. It would further be interesting to explore the personalization aspect, i.e., translator adaptation (instead of domain adaptation) with APE, in the hope that a small amount of PEs gathered from an individual are sufficient to learn and apply stylistic preferences of that translator. While we know that LSPs now do initial tests with APE technology and MT market players are also exploring integration of APE (Crego et al., 2016), its usage for follow-up editing in CAT environments is far from wide-spread.

12.3 Online Automatic Post-Editing

Our two APE architectures as well as most APE models in the related works were analyzed in batch mode, which means they are trained on a batch of data and then applied to a whole test set. This makes sense for evaluating an architecture, as it enables a fair comparison to other approaches and is sufficient to test if APE reduces human PE effort (Bojar et al., 2015). For practical use cases, the batch mode would simply imply piping translation jobs first through MT and then through APE to reduce errors for human PE. Depending on the chosen first-stage MT system and the APE model, we have seen that significant improvements are indeed possible with this straight-forward implementation. For any new translation job, one could thus first choose the best MT model available in terms of domain-adaptation, and then start capturing human post-edits. Instead of retraining the full MT model (which is not even possible without access to the inner workings of the model), one could then use the captured human post-edits for training an APE model reducing MT errors. As we have seen, APE requires much less (real) data than MT systems, thereby making this feasible in practice.

However, the problem of correcting repetitive mistakes, which translators particularly dislike about PE (O'Brien and Moorkens, 2014), would not be completely solved, as capturing data and re-training in batch mode also requires time for APE. This is why not only APE research, but also CAT tools need to focus on continuous adaptations of the MT/APE output, and tools like Lilt indeed offer such a feature (Balashov, 2020). Dynamic adaptations were originally explored in academic projects like MateCAT to improve a PBSMT engine from post-edits (Bertoldi et al., 2013). For this, they proposed a cache-based adaptation technique that "dynamically stores target n-gram and phrase-pair features used by the translator". Using a recency decay, they reward features stored in the cache as well as similar occurrences in the same document. This way, they "mix the large global (static) model with a small local (dynamic) model estimated from recent items" in the history. The local model is then combined with the global model during decoding, by scoring matches in both global static and local dynamic phrase tables. The goal of this combination is to translate "more consistently with the user preferences". A technical evaluation showed that the cache-based adaptation is useful to improve the quality, but the gains were not statistically significant. Overall, the authors concluded that cache-based adaptation is effective with repetitive texts while not hurting with non-repetitive text. However, a user evaluation of the gains in actual PE was not conducted in this work.

Since then many researchers explored similar adaptive NMT approaches to avoid repetitive mistakes. Most interesting for APE is probably the investigation by Negri et al. (2018a), exploring an online APE approach that learns from each human post-edit, again targeting the problem of correcting repetitive mistakes in PE. The idea is to improve the APE model on-the-fly which is essential for its integration in a CAT tool, and stands in contrast to the shared task that only explored the batch mode. Their online adaptation works in several steps: (1) Before human PE, a knowledge base is queried for the (src, mt, pe) triples most similar to the (src, mt) of the current segment. (2) These similar triples are used to update the APE model for a few training iterations. (3) The updated APE model produces an ape proposal, which a human post-edits (pe). This new (src, mt, pe) triple is then used to again retrain the APE model and extend the knowledge base for the future. To gain experimental results, they simulate human post-edits and apply APE on generic and specialized as well as static and adaptive NMT models, representing different performance/cost trade-offs that LSPs can choose:

1. a ‘generic‘ model represents the case where a LSP simply uses a black-box off-the-shelf NMT engine, therefore requiring no MT expertise
2. a ‘generic online‘ model is an online version of the ‘generic‘ model learning from human post-edits, thus representing the case where the LSP has access to an online NMT system and the knowledge to adapt it
3. a ‘specialized‘ system is a fine-tuned version of the ‘generic‘ model, meaning a LSP with access to customer data and the inner working of a batch NMT
4. a ‘specialized online‘ system using in-domain data to fine-tune the model in an online approach, thus requiring the LSP to have customer data and know the inner workings of an online NMT engine

Similarly, the authors propose an ‘online APE‘ that is initially generic but learns from PE-data. Their results show that ‘online APE‘ significantly improves the ‘generic‘, ‘generic online‘, and ‘specialized‘ NMT approaches, however is unable to improve over the ‘specialized online‘ NMT approach. Especially the improvement over ‘generic online‘ is interesting as it shows that APE is more reactive to human corrections, as it can “leverage richer information in the form of (src, mt, pe) instances”. Not being able to improve over ‘specialized online‘ (on the explored data) shows that APE cannot outperform a competitive online NMT engine. However, APE has several advantages over ‘specialized online‘, namely it does not rely on fine-tuning on a large in-domain corpus but still achieves similar results. Furthermore, and as already pointed out by the share task organizers of WMT 2015 (Bojar et al., 2015), APE does not require access to the inner workings of the online NMT, but can be considered independent. Thus, the authors argue that even though competitive online NMT solutions perform similar to online APE, the bottlenecks of collecting enough parallel sentences, having access to the inner workings, and having the computational power for fine-tuning on the large data, impose a disadvantage of adaptive NMT compared to APE.

12.4 The Impact of Training Data Amount on Automatic Post-Editing

Based on these papers it seems like APE is useful in several situations, but cannot outperform a strong in-domain NMT system, indicating the limitations of at least using neural APE to improve strong in-domain NMT. Chollampatt et al. (2020) recently questioned this, as they found that even the most modern APE approaches rely on small datasets of actual human post-edits (combined with large artificial datasets), which is also common practice at WMT. In their work, they compile a new EN-DE corpus called SubEdits containing more than 161k triplets, consisting of actual human post-edits of movie subtitles from a strong in-domain NMT model. Using this corpus, they show that state-of-the-art neural

APE approaches indeed can improve over strong in-domain NMT models when trained on a larger number of actual PE data. Apart from TER and BLEU-based evaluation, their human evaluation (comparing the NMT, APE, and PE versions instead of direct assessment), also agreed that the APE output is better than the NMT output, and the fact that the human PE labels were still rated highest shows that the crowd-based human post-edits are of high quality.

An investigation of the amount of data required shows that "size does matter" and that the improvement starts to slow down after 100k samples (while still increasing). Thus, the authors conclude that previous APE studies have only plateaued due to limited data.

They further investigated if artificial data helps by adding portions of the eSCAPE data (as an out-domain artificial set) and a newly created SubEscape dataset created using the eSCAPE approach but on in-domain data. They find that (i) using only artificial data the NMT cannot be beaten, (ii) using SubEscape in combination with SubEdits however does result in small improvements over using SubEdits alone.

An analysis using the WMT data and SubEdits shows that similar to NMT, APE is very domain-dependent and thus in-domain data is crucial to improve in-domain NMT. A qualitative analysis further shows that APE (i) can fix incorrect named-entity translations, (ii) corrects undertranslations, (iii) has the same problems with context as sentence-level NMT and therefore may generate wrong pronouns.

12.5 Conclusion

To summarize, this chapter showed that the reduction in errors through APE indeed leads to performance gains in PE. Furthermore, the recent WMT 2020 shared task was able to verify that APE indeed is suitable to adapt general purpose MT models to a certain domain and achieves large performance gains already through limited amounts of data. We then discussed how APE can be used in an online setting, where the model is updated before and after each new post-edit. This approach outperforms even specialized NMT models, and is on par with specialized online NMT models, while requiring less domain-specific data, no access and knowledge about the internal workings of the NMT, and less computational cost, thus being a valuable extension for LSPs. We have also seen that given sufficiently many post-edits, it is also possible to improve over highly specialized MT systems by leveraging the specifics of the PE data.

In contrast to the previous chapters, this chapter did not present published new findings on APE. Instead, it shows how approaches like the Multi-Source Transformer or Transference model can improve productivity and avoid repetitive errors in practice by looking at related studies. The chapter thus addresses our third research question by viewing our APE contributions in a bigger picture and outlining next steps to integrate the models into tools like MMPE.

Part V

Overall Conclusion

After the three main parts on explicit multi-modal interactions for PE of MT (Part II) , implicit multi-modal sensor input for modeling CL during PE (Part III), and multi-source APE to automatically correct repetitive errors in MT output (Part IV), this final part provides an overall conclusion. We first summarize the whole dissertation, focusing on the contributions made (chapter 13). Then, we discuss various possibilities for future research (chapter 14).

Chapter 13
Summary & Contributions

This chapter summarizes the work presented and discusses how the individual parts can improve human-machine collaboration for translation. Apart from highlighting the various contributions, we also discuss potential negative implications of this work.

13.1 Summary

Overall, this dissertation aimed to simplify the complex and cognitively challenging PE task by exploring novel multi-modal interaction possibilities for CAT environments, by considering translators' CL, and by learning from previous post-edits to automatically correct repetitive mistakes.

We first introduced the Post-Editing (PE) task, discussing its advantages and disadvantages, thereby motivating our research. Furthermore, we provided an in-depth review of a broad range of related research upon which this dissertation builds (Part I). Afterwards, we investigated improvements to the PE process by following three research directions.

Motivated by the finding that interaction patterns in PE are very distinct from those in traditional translation from scratch, we explored whether modalities other than mouse and keyboard might be well-suited for the PE task (Part II, RQ1). We first conducted an elicitation study with professional translators, indicating that a combination of pen, touch, and speech could well support common PE tasks (Herbig et al., 2019a). Afterwards, we built MMPE, the first translation environment combining these input modalities, allowing users to directly cross out or hand-write new text, drag and drop words for reordering, or use spoken commands to update the text in place (Herbig et al., 2020c). An evaluation of MMPE with professional translators (Herbig et al., 2020b) suggest that pen and touch interaction are suitable for deletion and reordering tasks; however, they are of limited use for longer insertions. On the other hand, speech and multi-modal combinations of select and speech are considered suitable for replacements and insertions but offer less potential for deletion and reordering. Overall, participants were enthusiastic about the new modalities and saw them as good extensions to mouse and keyboard, but not as a complete substitute. We further leveraged the more detailed subjective feedback to refine MMPE by improving the interface layout and enhancing individual components (Herbig et al., 2020d). We also integrated eye tracking functionality as an input modality that can be used in combination with speech commands. Guided by another elicitation study, we proposed and implemented a set of mid-air hand gestures as a mouse replacement for PE (Jamara et al., 2021). In a practical evaluation, gesture-based PE turned out quite promising, especially when considering that our participants use mouse and keyboard every day but are novices to gesture interaction. Finally, we used MMPE to explore word-level QE during PE (Shenoy et al., 2021), showing that a QE quality level of at least 80% F1 sets the approximate boundary where word-level QE starts helping translators, and for these QE quality levels, a visualization showing the uncertainty of the model is preferred.

Apart from the interaction patters, we further explored the cognitive dimension of PE (Part III, RQ2). Initial interviews revealed that most translators would see value in automatically receiving additional resources when a high CL is detected during PE (Herbig et al., 2019a). Thus, robust approaches to automatically estimate CL during PE could help with a better management of cognitive resources. We therefore built a multi-modal CL estimation framework combining a wide range of physiological, behavioral, and performance measures, many of which have not been previously explored in the translation domain. Using this framework (at different stages of implementation) we conducted 3 studies to estimate perceived CL: one with translation students (Herbig et al., 2019c), one with professional translators (Herbig et al., 2021), and one with students in an e-learning task to explore how specific the measures are to PE (Herbig et al., 2020a). We formulated the CL estimation problem as a prediction task, mapping the various captured sensor data to the subjectively reported perceived CL. In all three experiments we saw that the multi-modal approach performs best, followed by eye, heart and skin measures, although the performance differences slightly differed between the studies. Apart from understanding which sensors are well suited for CL estimation, we surveyed potential users which data sources are more likely to be shared and which are seen as more critical for cognition-aware systems (Herbig et al., 2019d). The results show that behavioral data like keyboard input and widespread physiological sensors would be shared without major concerns, while microphone or camera recordings as well as less common physiological data would require strong improvements to make disclosure worthwhile.

Finally, we presented works on Automatic Post Editing to tackle the problem of repetitive mistakes by the MT engine (Part IV, RQ3). We proposed two new architectures that can adapt any black-box MT system to a set of captured post-edits: the Multi-Source Transformer (Pal et al., 2018), and the Transference Architecture for APE (Pal et al., 2019, 2020). Both architectures build upon the Transformer model, but modify it for the multi-source setting, where both src and mt are considered to create the pe hypothesis. We demonstrated that especially the Transference model achieves state-of-the-art performance while having a comparatively simple and efficient architecture. Finally, we showed that APE can indeed be used to adapt generic MT systems to certain domains based on a limited amount of data, that online APE can quickly react to newly captured post-edits, and that given sufficient training data, APE can even outperform highly domain-optimized MT engines.

Overall, these three research directions can enhance the PE process through multi-modal input for CAT tools, multi-modal modeling of and adaptation to CL, and multi-source APE to automatically correct repetitive errors in MT output.

13.2 Contributions

Our contributions lie at the intersection of HCI and NLP and use recent advances in both fields to improve the PE process. This section summarizes the main contributions, first grouped into the three research questions tackled by this dissertation (see section 1.3), followed by highlighting the main design, technical, and theoretical findings.

Regarding **explicit multi-modal input for PE (RQ1)**, the outcome of our initial elicitation study shows the design space of interaction modalities for different PE operations, as well as insights on hardware setup and interface design of CAT tools (chapter 3, Herbig et al. (2019a)). Based on this theoretical foundation, the main contribution is the development and testing of MMPE, a CAT tool focusing on PE that supports various input modalities including hand-writing, touch reordering, speech commands, mid-air hand gestures, and multi-modal combinations thereof (chapter 4, 5, Herbig et al. (2020c,d); Jamara et al. (2021)). The prototype development, the subjective feedback by participants, and the iterative refinements based on the feedback contribute general CAT tool design insights, and particularly outline how to integrate modalities other than mouse and keyboard into the PE process. Finally, the structured test of the modalities for each individual PE task (Herbig et al., 2020b; Jamara et al., 2021) clearly indicates the strengths and weaknesses of the explored modalities, thereby guiding future CAT developers and allowing them to focus on the most relevant ones.

Regarding the **cognitive dimension of PE through implicit multi-modal sensor input (RQ2)**, we contribute ideas towards cognition-aware CAT tools based on interviews with professional translators (chapter 6, Herbig et al. (2019a)). As steps towards these goals, our CL estimation framework offers unified access to a broad range of CL measures from different sensor modalities, and can therefore be seen as a technical contribution of its own (chapter 7, Herbig et al. (2019c, 2021, 2020a)). Exploring this framework in three studies shows which modalities and modality combinations perform better or worse for CL estimation (chapter 8, Herbig et al. (2019c, 2021, 2020a)). These findings, in combination with our final survey-based investigation of users' privacy concerns and willingness to share sensor data (chapter 9, Herbig et al. (2019d)), can guide developers of such cognition-aware CAT tools that aim at both high accuracy in estimating CL and high user acceptance.

Regarding **learning from post-edits in the form of Automatic Post-Editing (RQ3)**, we contribute two APE model architectures combining both source text and MT proposal to automatically correct errors before performing human PE (chapter 10, 11, Pal et al. (2018, 2019, 2020)). Their analyses on publicly available datasets further advance the search for suitable multi-source APE architectures. We also theoretically discuss how APE can be integrated into the PE process to avoid correcting repetitive mistakes, its suitability to adapt generic MT engines to certain domains, and its competitiveness in comparison to highly domain-optimized MT systems (chapter 12).

Overall, this dissertation provides the following **design contributions**: We analyzed the general design space of interaction modalities for PE in an elicitation study (chapter 3, Herbig et al. (2019a)) and contribute a range of insights on CAT interface design based on the iterative development of MMPE (chapter 4, 5, Herbig et al. (2020c,d); Jamara et al. (2021)). Furthermore, we contribute ideas how to design CAT tools that adapt to the users' CL based on interviews with professional translators (chapter 6, Herbig et al. (2019a)).

As **technical contributions** this dissertation offers MMPE, a multi-modal CAT tool that is available open-source (chapter 4, 5, Herbig et al. (2020c,d)). Another engineering contribution is our CL estimation framework, that provides unified access to a broad range of sensors for CL analyses (chapter 7, Herbig et al. (2019c, 2021, 2020a)). Finally, two novel APE architectures have been proposed and implemented, the latter showing state-of-the-art results in terms of automatically correcting repetitive MT errors (chapter 10, 11, Pal et al. (2018, 2019, 2020)).

Regarding **theoretical contributions**, this dissertation explores various concepts from HCI in the field of PE: We show in structured tests which interaction modalities are suitable for which PE tasks, first in elicitation studies (chapter 3, chapter 5, Herbig et al. (2019a); Jamara et al. (2021)), followed by practical experiments with the MMPE environment (chapter 4, chapter 5, Herbig et al. (2020b); Jamara et al. (2021)). Apart from that, we explore a broad range of CL measures in three studies, showing which features perform better and worse for estimating perceived CL (chapter 8, Herbig et al. (2019c, 2021, 2020a)). These studies are complemented by an investigation of privacy concerns regarding sensor usage (chapter 9, Herbig et al. (2019d)). Furthermore, our APE architectures are thoroughly analyzed and compared to the state of the art (chapter 10, 11, Pal et al. (2018, 2019, 2020)) and we theoretically discuss how such APE models can be leveraged to avoid repetitive corrections during PE (chapter 12).

13.3 Individual Contributions

Large parts of this thesis are based on previous publications as a result of joint work with researchers and students. This section provides a list of publications (in the same order as the initial list of publications) and specifies to which parts of the work this dissertation's author contributed.

(Herbig et al., 2019a) Developing the main idea, planning and conducting the elicitation study, data analysis, paper writing.

(Herbig et al., 2020b) Developing the main idea, implementing MMPE, planning and conducting the evaluation, data analysis, paper writing.

(Herbig et al., 2020a) Developing the main idea, implementing the data capturing prototype, planning the study, data analysis, paper writing.

(Pal et al., 2020) Planning and running evaluation analyses, paper writing.

(Jamara et al., 2021) Developing the main idea, guiding the elicitation study planning and analysis, guiding the implementation, prototype evaluation planning, supporting data analysis, paper writing.

(Shenoy et al., 2021) Developing the main idea, guiding the artificial data creation and study planning, guiding the integration into MMPE, supporting data analysis, paper writing.

(Herbig et al., 2019c) Developing the main idea, implementing the data capturing prototype, planning and conducting the study, data analysis, paper writing.

(Herbig et al., 2021) Developing the main idea, implementing the data capturing prototype, planning and conducting the study, data analysis, paper writing.

(Herbig et al., 2019d) Developing the main idea, designing the questionnaire, data analysis, paper writing.

(Herbig et al., 2020c) Developing the main idea, implementing MMPE, planning and conducting the study, data analysis, paper writing.

(Pal et al., 2018) Data selection for fine-tuning, planning and running evaluation analyses, paper writing.

(Pal et al., 2019) Planning and running evaluation analyses, paper writing.

(Herbig et al., 2020d) Developing the main idea, implementing improvements to MMPE including eye tracking component, paper writing.

(Herbig et al., 2019b) Developing the main idea, paper writing.

(Jamara, 2021) Developing the main idea, supervision of the Master's thesis including guidance on (1) planning and evaluating the studies, (2) integration of gestures into MMPE, and (3) thesis writing.

(Shenoy, 2021) Developing the main idea, supervision of the Master's thesis including guidance on (1) planning and evaluating the user study, (2) conceptualizing artificial QE generation, (3) integration into MMPE, and (4) thesis writing.

(Akmal, 2021) Developing the main idea, supervision of the Master's thesis including guidance on (1) planning and evaluating the user study, (2) conceptualizing interactive PE visualizations, (3) integration into MMPE, and (4) thesis writing.

(Wang, 2021) Developing the main idea, supervision of the Master's thesis including guidance on (1) planning and evaluating the user study, (2) outputting high-quality but diverse MT proposals, (3) integration into MMPE, and (4) thesis writing.

13.4 Possible Unintended Drawbacks and Negative Implications

While research in general is targeted towards knowledge gain and improvements in life and work conditions, it can cause negative unintended drawbacks. In our case, the contributions made aim to make PE quicker and less demanding. At the same time, however, any performance improvements gained could lead to higher expectations in terms of translation volume per day, which could be coupled with lower payment per word, thereby increasing instead of reducing pressure on the human translator. As with any technology, it is therefore crucial to ensure appropriate use: the technology itself is neither good nor bad, the way it is used defines whether its strengths or weaknesses come to life.

Furthermore, this dissertation proposed considerable data capturing, e.g., for analyzing interactions within MMPE, for estimating CL, or for gathering training material for domain-adaptations through APE. When deploying such systems in practice, one should revisit these data capturing possibilities and ensure data sparsity, as well as security measures against inappropriate tracking or surveillance. Most critically in this regard is the tracking of speech input used for voice commands, as well as the physiological and behavioral data for CL estimation. For the former, organizations should aim for secure local ASR and video capture systems with very restricted access and ideally even without data logging. For the latter, our study on privacy concerns and willingness to share data gives insights into the sensitivity of the different CL measurement approaches, but naturally only necessary data should be stored, with restricted access and ideally only in anonymized form.

Chapter 14
Future Work

This chapter closes the thesis by outlining further possible follow-up studies to better understand the presented technologies, discussing how our contributions can be combined, and by proposing the integration of further interaction modalities or language technologies to continue improving the collaboration between human translators and MT systems.

14.1 Exploring Eye Tracking for Multi-Modal Input

As discussed in chapter 5, MMPE now also supports eye tracking as an additional input modality. Even though participants in our initial elicitation study (chapter 3) did not expect eye tracking to be suitable for PE, their comments showed that they seldomly considered it in combination with speech commands. At the same time, our multi-modal input in the form of cursor placement followed by a speech command lead to comments that one would "have to do two things at once", which could be resolved by eye input. Thus, our implementation of eye and speech should be tested in practice once the pandemic allows this. Furthermore, the combination of eye tracking and keyboard should be explored, e.g., in a similar approach to ReType (Sindhwani et al., 2019) but modified towards the PE setting. We are also interested to see if the visualization of the last fixations on source and target segments that we implemented similar to the GazeMarks (Kern et al., 2010) approach can further facilitate PE.

14.2 Further Studies on Modality Usage

Our presented studies on multi-modal interaction for PE MT output tested the different modalities in a very structured way (chapter 4). This allowed us to fairly compare the different modalities for the different tasks. At the same time, the guided nature of the test made it impossible to see which interaction modalities professional translators would choose for which editing task when having the freedom of choice. Would they stick to certain modalities for certain tasks? Would they correct multiple error types in a segment with the same or different modalities? Would certain modalities turn out better or worse than initially thought after a longer usage? Investigating these questions can yield further knowledge about modality usage during PE, so we hope that researchers and practitioners explore these ideas with our open-source MMPE prototype.

Of course, the individual modalities could also be further adapted and re-evaluated based on the latest publications: E.g., Fan et al. (2021) presented an algorithm for eyes-free speech interfaces based on large pre-trained language models that can (1) automatically remove colloquial inserts and (2) estimate whether a user wants to replace or insert text based on the target words, which might also improve speech input for PE.

14.3 Automatic Adaptations to Estimated Cognitive Load

Throughout this thesis, we have seen that PE is a cognitively demanding task and discussed approaches to estimate CL during PE. Future research should extend our behavioral measures to also include pen-, touch-, and speech-based features which could enhance CL estimation within MMPE. Furthermore, while several of our studies (chapter 8) showed the general feasibility of CL adaptations, translators' proposals for cognition-aware CAT tools discussed in chapter 6 should be implemented and explored in practice. Predictions of CL on shorter periods of time could further allow quicker adaptations to the user state as proposed by Schultheis and Jameson (2004). Overall, this would yield a system supporting multi-modal explicit input for text editing, as well as multi-sensory input for CL adaptations.

14.4 Practical Studies on Automatic Post-Editing

Part IV presented state-of-the-art APE systems and discussed how these can be leveraged for efficient PE. These models should be integrated into MMPE to explore their suitability to avoid repetitive mistakes and domain-/translator-adaptations in practice. While research on APE itself is growing year after year, investigating its practical use for the discussed benefits remains underexplored.

14.5 Integrating Multiple Machine Translation Proposals

In this thesis, we limited the concept of PE to editing the single best MT proposal. MT engines can however easily generate a variety of translation proposals for a sentence. As Balashov (2020) states, human brains are precision- and not recall-oriented, meaning that we are "much better at selecting the best candidate from a list of suggestions than at generating new suggestions from scratch". Thus, providing high quality and diverse options to chose from (e.g., using Diverse Beam Search by Vijayakumar et al. (2016)), a translator could find a proposal that is exactly or very close to a satisfying translation. Of course, the "diverse" aspect is essential, as seeing 10 very similar proposals most likely increases CL through additional information processing. Overall, we hypothesize that especially for short sentences, offering high quality but diverse translations with proper change highlighting could be beneficial for post-editors, whereas for longer sentences too many proposals could overwhelm the human. Wang (2021) presents our initial implementation and pre-study on the topic, which should be used as a basis to improve the prototype and run a full-scale study with professional translators.

14.6 Supporting Interactive Post-Editing

Instead of only varying the amount of proposals that the translator can choose from for PE, one should also explore interactive PE using our multi-modal setting. So instead of having the MT work first and the human afterwards, the two could work interleaved, with the MT proposing a translation, the human correcting a part therein, the MT adapting to this change and re-proposing another translation. Some tools like DeepL already do that by showing alternatives for the current word from the beam search process. However, DeepL's current visualization only presents alternatives for the clicked position, without showing if choosing this alternative will only replace the word by a synonym, or if the whole remainder of the sentence adapts. Better visualization, e.g., clustering the alternatives by what they will change, or showing the MT output in a graph form, could offer a lot of potential. Some first investigations of that topic have been pursued in Akmal (2021), however, proper studies with professional translators and combining such interactive PE with pen and finger touch input should be explored in the future, as they might make the keyboard almost superfluous.

14.7 Mobile Post-Editing

Furthermore, as MT gets better and better, and fewer and fewer mistakes need to be corrected, we hypothesize that the desktop setting becomes less relevant, wheras the tablet setting will gain popularity. While we believe that many of our findings will be directly applicable to the tablet mode (touch/pen/speech/multi-modal input), further studies investigating if additional changes are required should be conducted. Furthermore, the integration of interactive PE or multiple MT outputs, leading to less required manual changing (at the cost of more selection) might make tablets even more relevant. So far, apart from few studies on the Kanjingo app (O'Brien et al., 2014), most CAT studies and products ignore the mobile setting.

14.8 Adaptations for Context-Aware Machine Translation

A hot topic in MT research is also to move away from systems that translate sentence by sentence towards context-aware document-level MT. These models can generate better MT proposals for PE, as coreference chains etc. can be considered. Future research should therefore explore changes to CAT interfaces that help translators analyze context or visualizations of the internal workings of document-level MT systems (e.g., which part of the previous sentence lead to the choice of word form in this sentence). This interface aspect is so far a rather unresearched field, which might however offer a lot of potential as MT systems are becoming better and better at understanding context.

14.9 Exploration in Related Contexts

Furthermore, studies in very related contexts like text review and proofreading should be conducted to explore which of our findings are also generalizable to these fields. Other interesting paths of research include CAT tools for sign language translation or audiovisual data, which are gaining importance. E.g., one could adapt MMPE to support YouTube subtitle creation and translation by showing the video alongside the automatically generated ASR output that requires correction.

14.10 Closing Remarks

This thesis holistically explored multi-modality for the task of Post-Editing Machine Translation. We considered both multi-modal interaction possibilities to correct the MT output more efficiently, and multi-modal sensor usage to model the translators' cognitive load while post-editing. Enhanced with interview and questionnaire results, we further provide lots of expert feedback on how the collaboration of human and machine for the translation task can be further enhanced. We also looked at supportive language technologies, namely APE and QE, and how such AI components can be integrated into the process. By open sourcing the main prototype and large parts of the data, we hope to simplify future research that builds upon our findings. Given the various ideas for additional studies, extensions to the systems, and a wider, namely more interactive, consideration of PE, we hope that some of these ideas will be picked up and explored in future research. Naturally, future trends in both HCI and NLP need to be considered to make sure that the latest technologies are investigated when trying to further advance the human-AI collaboration for translation.

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
APE	Automatic Post-Editing
APR	Average Pause Ratio
ASR	Automatic Speech Recognition
BLEU	BiLingual Evaluation Understudy
BPE	Byte-Pair Encoding
CAT	Computer-Aided Translation
CL	Cognitive Load
GSR	Galvanic Skin Response
HCI	Human-Computer Interaction
HRV	Heart Rate Variability
HT	Human Translation
ICA	Index of Cognitive Activity
IMT	Interactive Machine Translation
MMPE	Multi-Modal Post-Editing
MT	Machine Translation
NLP	Natural Language Processing
NMT	Neural Machine Translation
PBSMT	Phrase-Based Statistical Machine Translation
PE	Post-Editing
PWR	Pause to Word Ratio
QA	Quality Assurance
QE	Quality Estimation
RBMT	Rule-Based Machine Translation
SMT	Statistical Machine Translation
TB	Term Base
TER	Translation Edit Rate
TM	Translation Memory
WMT	Conference/Workshop on Machine Translation

Bibliography

Accuray Research LLP. 2017. Global e-learning market analysis & trends – industry forecast to 2025.

Alessandro Acquisti, Leslie K John, and George Loewenstein. 2013. What is privacy worth? *Journal of Legal Studies*, 42(2):249–274.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation. In *Conference on Machine Translation*, pages 1–88. Association for Computational Linguistics.

Atika Akmal. 2021. IPE: enhancing the visualization of multiple alternatives in interactive post-editing. Master's thesis, Saarland University.

Vicent Alabau, Ragnar Bonk, Christian Buck, Michael Carl, Francisco Casacuberta, Mercedes García-Martínez, Jesús González, Philipp Koehn, Luis Leiva, Bartolomé Mesa-Lao, Daniel Ortiz, Herve Saint-Amand, Germán Sanchis, and Chara Tsoukala. 2013a. CASMACAT: an open source workbench for advanced computer aided translation. *The Prague Bulletin of Mathematical Linguistics*, 100:101–112.

Vicent Alabau, Ragnar Bonk, Christian Buck, Michael Carl, Francisco Casacuberta, Mercedes García-Martínez, Jesús González, Luis Leiva, Bartolomé Mesa-Lao, Daniel Ortiz, Herve Saint-Amand, Germán Sanchin, and Chara Tsoukala. 2013b. Advanced computer aided translation with a web-based workbench. In *Workshop on Post-Editing Technologies and Practice*, pages 55–62. Association for Computational Linguistics.

Vicent Alabau, Christian Buck, Michael Carl, Francisco Casacuberta, Mercedes García-Martínez, Ulrich Germann, Jesús González-Rubio, Robin Hill, Philipp Koehn, Luis A Leiva, Bartolomé Mesa-Lao, Daniel Ortiz, Herve Saint-Amand, Germán Sanchis, and Chara Tsoukala. 2014a. CASMACAT: a computer-assisted translation workbench. In *Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 25–28. Association for Computational Linguistics.

Vicent Alabau and Francisco Casacuberta. 2012. Study of electronic pen commands for interactive-predictive machine translation. In *International Workshop on Expertise in Translation and Post-Editing*, pages 17–18.

Vicent Alabau and Luis A Leiva. 2014. Proofreading human translations with an e-pen. In *EACL Workshop on Humans and Computer-assisted Translation*, pages 10–15. Association for Computational Linguistics.

Vicent Alabau, Luis A Leiva, Daniel Ortiz-Martínez, and Francisco Casacuberta. 2012. User evaluation of interactive machine translation systems. In *Conference of the European Association for Machine Translation*, pages 20–23. European Association for Machine Translation.

Vicent Alabau, Alberto Sanchis, and Francisco Casacuberta. 2014b. Improving on-line handwritten recognition in interactive machine translation. *Pattern Recognition*, 47(3):1217–1228.

Fabio Alves, Arlene Koglin, Bartolomé Mesa-Lao, Mercedes García Martínez, Norma B de Lima Fonseca, Arthur de Melo Sá, José Luiz Gonçalves, Karina Sarto Szpak, Kyoko Sekino, and Marceli Aquino. 2016. Analysing the impact of interactive machine translation on post-editing effort. In *New Directions in Empirical Translation Process Research*, pages 77–94. Springer.

Nora Aranberri, Gorka Labaka, A Diaz de Ilarraza, and Kepa Sarasola. 2014. Comparison of post-editing productivity between professional translators and lay users. In *AMTA Workshop on Post-editing Technology and Practice*, pages 20–33. Association for Machine Translation in the Americas.

Syed Arshad, Yang Wang, and Fang Chen. 2013. Analysing mouse activity for cognitive load detection. In *Australian Computer-Human Interaction Conference*, pages 115–118. Association for Computing Machinery.

Stylianos Asteriadis, Paraskevi Tzouveli, Kostas Karpouzis, and Stefanos Kollias. 2009. Estimation of behavioral user state based on eye gaze and head pose – application in an e-learning environment. *Multimedia Tools and Applications*, 41(3):469–493.

Automatic Language Processing Advisory Committee. 1966. *Language and Machines: Computers in Translation and Linguistics*. The National Academies Press.

Wilker Aziz, Sheila Castilho, and Lucia Specia. 2012. PET: a tool for post-editing and assessing machine translation. In *International Conference on Language Resources and Evaluation*, pages 3982–3987. European Language Resources Association.

Alan D Baddeley and Robert H Logie. 1999. Working memory: the multiple-component model. In *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*, page 28–61. Cambridge University Press.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, pages 1–15. OpenReview.net.

Kiavash Bahreini, Rob Nadolski, and Wim Westera. 2016. Towards multimodal emotion recognition in e-learning environments. *Interactive Learning Environments*, 24(3):590–605.

Yuri Balashov. 2020. The translator's extended mind. *Minds and Machines*, 30(3):349–383.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *ACL Workshop on Iintrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72. Association for Computational Linguistics.

Sergio Barrachina, Oliver Bender, Francisco Casacuberta, Jorge Civera, Elsa Cubel, Shahram Khadivi, Antonio Lagarda, Hermann Ney, Jesús Tomás, Enrique Vidal, and Juan-Miguel Vilar. 2009. Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28.

Loïc Barrault, Ondřej Bojar, Marta R Costa-Jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation. In *Conference on Machine Translation*, pages 1–61. Association for Computational Linguistics.

Mathias Bauer, Cassandra Bräuer, Jacqueline Schuldt, and Heidi Krömker. 2018. Adaptive e-learning technologies for sustained learning motivation in engineering science - acquisition of motivation through self-reports and wearable technology. In *International Conference on Computer Supported Education*, pages 418–425. SciTePress.

Jackson Beatty. 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, 91(2):276–292.

Hanna Béchara, Yanjun Ma, and Josef van Genabith. 2011. Statistical post-editing for a statistical MT system. In *Machine Translation Summit*, pages 308–315. Association for Computational Linguistics.

Moritz Becker, Christian Matt, Thomas Widjaja, and Thomas Hess. 2017. Understanding privacy risk perceptions of consumer health wearables – an empirical taxonomy. In *International Conference on Information Systems*, pages 1–21. AIS.

Mathias Benedek and Christian Kaernbach. 2010. A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods*, 190(1):80–91.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Conference on Empirical Methods in Natural Language Processing*, pages 257–267. Association for Computational Linguistics.

Marten E van den Berg, Peter R Rijnbeek, Maartje N Niemeijer, Albert Hofman, Gerard van Herpen, Michiel L Bots, Hans Hillege, Cees A Swenne, Mark Eijgelsheim, Bruno H Stricker, and Jan A Kors. 2018. Normal values of corrected heart-rate variability in 10-second electrocardiograms for all ages. *Frontiers in Physiology*, 9:1–9.

Jan Van den Bergh, Eva Geurts, Donald Degraen, Mieke Haesen, Iulianna Van der Lek-Ciudin, and Karin Coninx. 2015. Recommendations for translation environments to improve translators' workflows. *Translating and the Computer*, 37:106–119.

Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2013. Cache-based online adaptation for machine translation enhanced computer assisted translation. In *Machine Translation Summit*, pages 35–42. Association for Machine Translation in the Americas.

Ergun Biçici, Declan Groves, and Josef van Genabith. 2013. Predicting sentence translation quality using extrinsic and language independent features. *Machine Translation*, 27(3-4):171–192.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Conference on Machine Translation*, pages 12–58. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation. In *Conference on Machine Translation*, pages 169–214. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Conference on Machine Translation*, pages 131–198. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Conference on Machine Translation*, pages 1–46. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation. In *Conference on Machine Translation*, pages 272–307. Association for Computational Linguistics.

Magdalena Borys, Małgorzata Plechawska-Wójcik, Martyna Wawrzyk, and Kinga Wesołowska. 2017. Classifying cognitive workload using eye activity and EEG features in arithmetic tasks. In *International Conference on Information and Software Technologies*, pages 90–105. Springer.

Lynne Bowker and Jairo Buitrago Ciro. 2015. Investigating the usefulness of machine translation for newcomers at the public library. *Translation and Interpreting Studies*, 10(2):165–186.

Julie Brousseau, Caroline Drouin, George Foster, Pierre Isabelle, Roland Kuhn, Yves Normandin, and Pierre Plamondon. 1995. French speech recognition in an automatic dictation system for translators: the TransTalk project. In *European Conference on Speech Communication and Technology*, pages 193–196. International Speech Communication Association.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, John Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, Robert L Mercer, and Paul Roossin. 1988. A statistical approach to language translation. In *International Conference on Computational Linguistics*, pages 71–76. The COLING Organizing Committee.

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Annual Conference on Neural Information Processing Systems*, pages 1–25. Curran Associates, Inc.

Jody Byrne. 2006. *Technical Translation: Usability Strategies for Translating Technical Documentation*. Springer.

Chris Callison-Burch, Cameron Shaw Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Conference on Machine Translation*, pages 136–158. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar F Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Joint Workshop on Statistical Machine Translation and Metrics*, pages 17–53. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Conference on Machine Translation*, pages 22–64. Association for Computational Linguistics.

Michael Carl. 2012. Translog-II: a program for recording user activity data for empirical reading and writing research. In *International Conference on Language Resources and Evaluation*, pages 4108–4112. European Language Resources Association.

Michael Carl, Barbara Dragsted, Jakob Elming, Daniel Hardt, and Arnt Lykke Jakobsen. 2011. The process of post-editing: a pilot study. *Copenhagen Studies in Language*, 41:131–142.

Michael Carl, Martin Kay, and Kristian TH Jensen. 2010. Long distance revisions in drafting and post-editing. *Natural Language Processing and its Applications*, pages 193–204.

Sheila Castilho, Federico Gaspari, Joss Moorkens, Maja Popović, and Antonio Toral. 2019. Editors' foreword to the special issue on human factors in neural machine translation. *Machine Translation*, 33(1-2):1–7.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, 108:109–120.

Guillaume Chanel, Cyril Rebetez, Mireille Bétrancourt, and Thierry Pun. 2008. Boredom, engagement and anxiety as indicators for adaptation to difficulty in games. In *International Conference on Entertainment and Media in the Ubiquitous Era*, pages 13–17. Association for Computing Machinery.

Rajen Chatterjee, M Amin Farajian, Matteo Negri, Marco Turchi, Ankit Srivastava, and Santanu Pal. 2017a. Multi-source neural automatic post-editing: FBK's participation in the WMT 2017 APE shared task. In *Conference on Machine Translation*, pages 630–638. Association for Computational Linguistics.

Rajen Chatterjee, Christian Federmann, Matteo Negri, and Marco Turchi. 2019. Findings of the WMT 2019 shared task on automatic post-editing. In *Conference on Machine Translation*, pages 13–30. Association for Computational Linguistics.

Rajen Chatterjee, Markus Freitag, Matteo Negri, and Marco Turchi. 2020. Findings of the WMT 2020 shared task on automatic post-editing. In *Conference on Machine Translation*, pages 644–657. Association for Computational Linguistics.

Rajen Chatterjee, Gebremedhen Gebremelak, Matteo Negri, and Marco Turchi. 2017b. Online automatic post-editing for MT in a multi-domain translation environment. In *Conference of the European Chapter of the Association for Computational Linguistics*, pages 525–535. Association for Computational Linguistics.

Rajen Chatterjee, Matteo Negri, Raphael Rubino, and Marco Turchi. 2018. Findings of the WMT 2018 shared task on automatic post-editing. In *Conference on Machine Translation*, pages 710–725. Association for Computational Linguistics.

Rajen Chatterjee, Marion Weller, Matteo Negri, and Marco Turchi. 2015. Exploring the planet of the APEs: a comparative study of state-of-the-art methods for MT automatic post-editing. In *Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing: Short Papers*, pages 156–161. The Association for Computer Linguistics.

Fang Chen, Jianlong Zhou, Yang Wang, Kun Yu, Syed Z Arshad, Ahmad Khawaji, and Dan Conway. 2016. *Robust Multimodal Cognitive Load Measurement*. Springer.

Jingjing Chen. 2016. *Enhancing student engagement and interaction in e-learning environments through learning analytics and wearable sensing*. Ph.D. thesis, Hong Kong Baptist University.

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

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: encoder–decoder approaches. In *Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111. Association for Computational Linguistics.

Shamil Chollampatt, Raymond Hendy Susanto, Liling Tan, and Ewa Szymanska. 2020. Can automatic post-editing improve NMT? In *Conference on Empirical Methods in Natural Language Processing*, pages 2736–2746. Association for Computational Linguistics.

Kenneth W Church and Eduard H Hovy. 1993. Good applications for crummy machine translation. *Machine Translation*, 8(4):239–258.

Jonathan H Clark, Chris Dyer, Alon Lavie, and Noah A Smith. 2011. Better hypothesis testing for statistical machine translation: controlling for optimizer instability. In *Annual Meeting of the Association for Computational Linguistics: Short Papers*, pages 176–181. Association for Computational Linguistics.

Sven Coppers, Jan Van den Bergh, Kris Luyten, Karin Coninx, Iulianna van der Lek-Ciudin, Tom Vanallemeersch, and Vincent Vandeghinste. 2018. Intellingo: an intelligible translation environment. In *Conference on Human Factors in Computing Systems*, pages 1–13. Association for Computing Machinery.

Gregory W Corder and Dale I Foreman. 2009. *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach*. Wiley.

Josep Crego, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurelien Coquard, Yongchao Deng, Satoshi Enoue, Chiyo Geiss, Joshua Johanson, Ardas Khalsa, Raoum Khiari, Byeongil Ko, Catherine Kobus, Jean Lorieux, Leidiana Martins, Dang-Chuan Nguyen, Alexandra Priori, Thomas Riccardi, Natalia Segal, Christophe Servan, Cyril Tiquet, Bo Wang, Jin Yang, Zhang Dakun, Jing Zhou, and Peter Zoldan. 2016. Systran's pure neural machine translation systems. *CoRR*, abs/1610.05540:1–23.

Oliver Čulo and Jean Nitzke. 2016. Patterns of terminological variation in post-editing and of cognate use in machine translation in contrast to human translation. In *Conference of the European Association for Machine Translation*, pages 106–114. European Association for Machine Translation.

Cristina Cumbreno and Nora Aranberri. 2019. Comparison of temporal, technical and cognitive dimension measurements for post-editing effort. In *MEMENTO Workshop on Modelling Parameters of Cognitive Effort in Translation Production*, pages 5–6. European Association for Machine Translation.

Joke Daems. 2016. *A translation robot for each translator? A comparative study of manual translation and post-editing of machine translations: process, quality and translator attitude*. Ph.D. thesis, Ghent University.

Joke Daems, Orphée De Clercq, and Lieve Macken. 2017. Translationese and post-editese: how comparable is comparable quality? *Linguistica Antverpiensia New Series-Themes in Translation Studies*, 16:89–103.

Joke Daems and Lieve Macken. 2019. Interactive adaptive SMT versus interactive adaptive NMT: a user experience evaluation. *Machine Translation*, 33(1-2):117–134.

Aswarth Abhilash Dara, Josef van Genabith, Qun Liu, John Judge, and Antonio Toral. 2014. Active learning for post-editing based incrementally retrained MT. In *Conference of the European Chapter of the Association for Computational Linguistics: Short Papers*, pages 185–189. Association for Computational Linguistics.

Prerit Datta, Akbar Siami Namin, and Moitrayee Chatterjee. 2018. A survey of privacy concerns in wearable devices. In *International Conference on Big Data*, pages 4549–4553. IEEE.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *International Conference on Machine Learning*, pages 933–941. International Machine Learning Society.

Giselle De Almeida. 2013. *Translating the post-editor: an investigation of post-editing changes and correlations with professional experience across two Romance languages*. Ph.D. thesis, Dublin City University.

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

Michael Denkowski, Chris Dyer, and Alon Lavie. 2014. Learning from post-editing: online model adaptation for statistical machine translation. In *Conference of the European Chapter of the Association for Computational Linguistics*, pages 395–404. Association for Computational Linguistics.

Franck Dernoncourt. 2014. Replacing the computer mouse. *CoRR*, abs/1410.5907:1–9.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186. Association for Computational Linguistics.

Denise Díaz, James Cross, Vishrav Chaudhary, Ahmed Kishky, and Philipp Koehn. 2020. A survey of qualitative error analysis for neural machine translation systems. In *Conference of the Association for Machine Translation in the Americas*, pages 48–77. Association for Machine Translation in the Americas.

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

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *International Conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc.

Stephen Doherty, Sharon O'Brien, and Michael Carl. 2010. Eye tracking as an MT evaluation technique. *Machine Translation*, 24(1):1–13.

Tobias Domhan. 2018. How much attention do you need? A granular analysis of neural machine translation architectures. In *Annual Meeting of the Association for Computational Linguistics*, pages 1799–1808. Association for Computational Linguistics.

Barbara Dragsted, Inger Margrethe Mees, and Inge Gorm Hansen. 2011. Speaking your translation: students' first encounter with speech recognition technology. *Translation & Interpreting*, 3(1):10–43.

Marc Dymetman, Julie Brousseau, George F Foster, Pierre Isabelle, Yves Normandin, and Pierre Plamondon. 1994. Towards an automatic dictation system for translators: the TransTalk project. In *International Conference on Spoken Language Processing*. International Speech Communication Association.

José Esteban, José Lorenzo, Antonio S Valderrábanos, and Guy Lapalme. 2004. TransType2 - an innovative computer-assisted translation system. In *Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 94–97. Association for Computational Linguistics.

European Commission representation in the UK and the Institute of Translation and Interpreting. 2016. UK translator survey: final report. Technical report, Chartered Institute of Linguists.

Jiayue Fan, Chenning Xu, Chun Yu, and Yuanchun Shi. 2021. Just speak it: minimize cognitive load for eyes-free text editing with a smart voice assistant. In *Symposium on User Interface Software and Technology*, pages 910–921. Association for Computing Machinery.

Michael Farrell. 2018. Machine translation markers in post-edited machine translation output. *Translating and the Computer*, 40:50–59.

Marcello Federico, Nicola Bertoldi, Mauro Cettolo, Matteo Negri, Marco Turchi, Marco Trombetti, Alessandro Cattelan, Antonio Farina, Domenico Lupinetti, Andrea Martines, Alberto Massidda, Holger Schwenk, Loïc Barrault, Frederic Blain, Philipp Koehn, Christian Buck, and Ulrich Germann. 2014. The MateCat tool. In *International Conference on Computational Linguistics: System Demonstrations*, pages 129–132. The COLING Organizing Committee.

Rune Fensli, PE Pedersen, T Gundersen, and O Hejlesen. 2008. Sensor acceptance model – measuring patient acceptance of wearable sensors. *Methods of Information in Medicine*, 47(1):89–95.

Rebecca Fiederer and Sharon O'Brien. 2009. Quality and machine translation: a realistic objective. *The Journal of Specialised Translation*, 11:52–74.

Leah Findlater, Ben Lee, and Jacob Wobbrock. 2012. Beyond QWERTY: augmenting touch screen keyboards with multi-touch gestures for non-alphanumeric input. In *Conference on Human Factors in Computing Systems*, page 2679–2682. Association for Computing Machinery.

Federico Cirett Galán and Carole R Beal. 2012. EEG estimates of engagement and cognitive workload predict math problem solving outcomes. In *Conference on User Modeling, Adaptation, and Personalization*, page 51–62. Springer.

Alejandro García-Aragón and Clara Inés López-Rodríguez. 2017. Translators' needs and preferences in the design of specialized termino-lexicographic tools. In *Human Issues in Translation Technology*, pages 80–108. Routledge.

Ankush Garg and Mayank Agarwal. 2019. Machine translation: a literature review. *CoRR*, abs/1901.01122:1–17.

Federico Gaspari, Antonio Toral, Sudip Kumar Naskar, Declan Groves, and Andy Way. 2014. Perception vs. reality: measuring machine translation post-editing productivity. In *AMTA Workshop on Post-Editing Technology and Practice*, pages 60–72. Association for Machine Translation in the Americas.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, page 1243–1252. International Machine Learning Society.

Giuseppe Ghiani, Marco Manca, and Fabio Paternò. 2015. Dynamic user interface adaptation driven by physiological parameters to support learning. In *Symposium on Engineering Interactive Computing Systems*, page 158–163. Association for Computing Machinery.

Joseph H Goldberg and Xerxes P Kotval. 1999. Computer interface evaluation using eye movements: methods and constructs. *International Journal of Industrial Ergonomics*, 24(6):631–645.

Michael D Good, John A Whiteside, Dennis R Wixon, and Sandra J Jones. 1984. Building a user-derived interface. *Communications of the ACM*, 27(10):1032–1043.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Spence Green, Jason Chuang, Jeffrey Heer, and Christopher D Manning. 2014a. Predictive translation memory: a mixed-initiative system for human language translation. In *Symposium on User Interface Software and Technology*, page 177–187. Association for Computing Machinery.

Spence Green, Jeffrey Heer, and Christopher D Manning. 2013. The efficacy of human post-editing for language translation. In *Conference on Human Factors in Computing Systems*, page 439–448. Association for Computing Machinery.

Spence Green, Sida I Wang, Jason Chuang, Jeffrey Heer, Sebastian Schuster, and Christopher D Manning. 2014b. Human effort and machine learnability in computer aided translation. In *Conference on Empirical Methods in Natural Language Processing*, pages 1225–1236. Association for Computational Linguistics.

Ana Guerberof. 2008. *Productivity and quality in the post-editing of outputs from translation memories and machine translation: a pilot study*. Ph.D. thesis, Tarragona: Universitat Rovira i Virgili.

Ana Guerberof. 2009. Productivity and quality in the post-editing of outputs from translation memories and machine translation. *International Journal of Localization*, 7(1):11–21.

Ana Guerberof-Arenas. 2013. What do professional translators think about post-editing. *The Journal of Specialised Translation*, 19:75–95.

Ana Guerberof-Arenas and Antonio Toral. 2020. The impact of post-editing and machine translation on creativity and reading experience. *Translation Spaces*, 9(2):255–282.

Markus Guhe, Wayne D Gray, Michael J Schoelles, Wenhui Liao, Zhiwei Zhu, and Qiang Ji. 2005. Non-intrusive measurement of workload in real-time. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 49(12):1157–1161.

Eija Haapalainen, SeungJun Kim, Jodi F Forlizzi, and Anind K Dey. 2010. Psycho-physiological measures for assessing cognitive load. In *International Conference on Ubiquitous Computing*, pages 301–310. Association for Computing Machinery.

Aaron Li-Feng Han, Derek Fai Wong, and Lidia S Chao. 2016. Machine translation evaluation: a survey. *CoRR*, abs/1605.04515:1–16.

Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In *Advances in Psychology*, pages 139–183. Elsevier.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic Chinese to English news translation. *CoRR*, abs/1803.05567:1–25.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016a. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, pages 770–778. IEEE.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016b. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, pages 770–778. IEEE.

Yifan He, Yanjun Ma, Josef van Genabith, and Andy Way. 2010. Bridging SMT and TM with translation recommendation. In *Annual Meeting of the Association for Computational Linguistics*, pages 622–630. Association for Computational Linguistics.

Nico Herbig, Tim Düwel, Mossad Helali, Lea Eckhart, Patrick Schuck, Subhabrata Choudhury, and Antonio Krüger. 2020a. Investigating multi-modal measures for cognitive load detection in e-learning. In *Conference on User Modeling, Adaptation and Personalization*, pages 88–97. Association for Computing Machinery.

Nico Herbig, Tim Düwel, Santanu Pal, Kalliopi Meladaki, Mahsa Monshizadeh, Antonio Krüger, and Josef van Genabith. 2020b. MMPE: a multi-modal interface for post-editing machine translation. In *Annual Meeting of the Association for Computational Linguistics*, pages 1691–1702. Association for Computational Linguistics.

Nico Herbig, Santanu Pal, Tim Düwel, Kalliopi Meladaki, Mahsa Monshizadeh, Vladislav Hnatovskiy, Antonio Krüger, and Josef van Genabith. 2020c. MMPE: a multi-modal interface using handwriting, touch reordering, and speech commands for post-editing machine translation. In *Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 327–334. Association for Computational Linguistics.

Nico Herbig, Santanu Pal, Tim Düwel, Raksha Shenoy, Antonio Krüger, and Josef van Genabith. 2020d. Improving the multi-modal post-editing (MMPE) CAT environment based on professional translators' feedback. In *AMTA Workshop on Post-Editing in Modern-Day Translation*, pages 93–108. Association for Machine Translation in the Americas.

Nico Herbig, Santanu Pal, Josef van Genabith, and Antonio Krüger. 2019a. Multimodal approaches for post-editing machine translation. In *Conference on Human Factors in Computing Systems*, pages 1–11. Association for Computing Machinery.

Nico Herbig, Santanu Pal, Antonio Krüger, and Josef van Genabith. 2021. Multimodal estimation of cognitive load in post-editing of machine translation. In *Translation, Interpreting, Cognition: The Way Out of the Box*, pages 1–32. Language Science Press.

Nico Herbig, Santanu Pal, Josef van Genabith, and Antonio Krüger. 2019b. Integrating artificial and human intelligence for efficient translation. *CoRR*, abs/1903.02978:1–4.

Nico Herbig, Santanu Pal, Mihaela Vela, Antonio Krüger, and Josef van Genabith. 2019c. Multi-modal indicators for estimating perceived cognitive load in post-editing of machine translation. *Machine Translation*, 33(1–2):91–115.

Nico Herbig, Patrick Schuck, and Antonio Krüger. 2019d. User acceptance of cognition-aware e-learning: an online survey. In *International Conference on Mobile and Ubiquitous Multimedia*, pages 1–6. Association for Computing Machinery.

G Robert J Hockey. 1997. Compensatory control in the regulation of human performance under stress and high workload: a cognitive-energetical framework. *Biological Psychology*, 45(1):73–93.

Chris Hokamp. 2015. A component-centric design framework for translation interfaces. In *EXPERT Scientific and Technological Workshop*, pages 91–102. Tradulex.

Chris Hokamp and Qun Liu. 2015. HandyCAT - an open-source platform for CAT tool research. In *Conference of the European Association for Machine Translation*, page 216. European Association for Machine Translation.

Gahangir Hossain and Mohammed Yeasin. 2014. Understanding effects of cognitive load from pupillary responses using Hilbert analytic phase. In *Conference on Computer Vision and Pattern Recognition Workshops*, pages 375–380. IEEE.

Seyyed Abed Hosseini and Mohammad Ali Khalilzadeh. 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*, pages 1–6. IEEE.

John Hutchins. 2000. Warren Weaver and the launching of MT. In John Hutchins, editor, *Early Years in Machine Translation: Memoirs and Biographies of Pioneers*, pages 17–20. John Benjamins Publishing Company.

John Hutchins. 2007. Machine translation: a concise history. *Computer Aided Translation: Theory and Practice*, 13(29-70):11.

John Hutchins and Evgenii Lovtskii. 2000. Petr Petrovich Troyanskii (1894–1950): a forgotten pioneer of mechanical translation. *Machine Translation*, 15(3):187–221.

Cristina Iani, Daniel Gopher, and Peretz Lavie. 2004. Effects of task difficulty and invested mental effort on peripheral vasoconstriction. *Psychophysiology*, 41(5):789–798.

Shamsi T Iqbal, Xianjun Sam Zheng, and Brian P Bailey. 2004. Task-evoked pupillary response to mental workload in human-computer interaction. In *CHI Extended Abstracts on Human Factors in Computing Systems*, pages 1477–1480. Association for Computing Machinery.

Shoya Ishimaru, Soumy Jacob, Apurba Roy, Syed Saqib Bukhari, Carina Heisel, Nicolas Großmann, Michael Thees, Jochen Kuhn, and Andreas Dengel. 2017. Cognitive state measurement on learning materials by utilizing eye tracker and thermal camera. In *International Conference on Document Analysis and Recognition*, pages 32–36. IEEE.

Rashad Albo Jamara. 2021. Using hand gestures for text editing tasks in post-editing of machine translation. Master's thesis, Saarland University.

Rashad Albo Jamara, Nico Herbig, Antonio Krüger, and Josef van Genabith. 2021. Mid-air hand gestures for post-editing of machine translation. In *Annual Meeting of the Association for Computational Linguistics*, pages 6763–6773. Association for Computational Linguistics.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing. In *Conference on Machine Translation*, pages 751–758. Association for Computational Linguistics.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2018. MS-UEdin submission to the WMT2018 APE shared task: dual-source transformer for automatic post-editing. In *Conference on Machine Translation*, pages 835–839. Association for Computational Linguistics.

Samira Ebrahimi Kahou, Xavier Bouthillier, Pascal Lamblin, Caglar Gulcehre, Vincent Michalski, Kishore Konda, Sébastien Jean, Pierre Froumenty, Yann Dauphin, Nicolas Boulanger-Lewandowski, Raul Chandias Ferrari, Mehdi Mirza, David Warde-Farley, Aaron Courville, Pascal Vincent, Roland Memisevic, Christopher Pal, and Yoshua Bengio. 2016. EmoNets: multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, 10(2):99–111.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709. Association for Computational Linguistics.

Slava Kalyuga, Paul Chandler, and John Sweller. 1998. Levels of expertise and instructional design. *Human Factors*, 40(1):1–17.

Xiaomian Kang, Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2020. Dynamic context selection for document-level neural machine translation via reinforcement learning. In *Conference on Empirical Methods in Natural Language Processing*, pages 2242–2254. Association for Computational Linguistics.

Martin Kay. 1997. The proper place of men and machines in language translation. *Machine Translation*, 12(1-2):3–23.

Dorothy Kenny. 2017. *Human Issues in Translation Technology: The IATIS Yearbook*. Taylor & Francis.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André FT Martins. 2019. OpenKiwi: an open source framework for quality estimation. In *Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122. Association for Computational Linguistics.

Dagmar Kern, Paul Marshall, and Albrecht Schmidt. 2010. Gazemarks: gaze-based visual placeholders to ease attention switching. In *Conference on Human Factors in Computing Systems*, pages 2093–2102. Association for Computing Machinery.

Shahram Khadivi and Hermann Ney. 2008. Integration of speech recognition and machine translation in computer-assisted translation. *Transactions on Audio, Speech, and Language Processing*, 16(8):1551–1564.

M Asif Khawaja, Fang Chen, and Nadine Marcus. 2014. Measuring cognitive load using linguistic features: implications for usability evaluation and adaptive interaction design. *International Journal of Human-Computer Interaction*, 30(5):343–368.

M Asif Khawaja, Natalie Ruiz, and Fang Chen. 2007. Potential speech features for cognitive load measurement. In *Australasian Conference on Computer-Human Interaction*, pages 57–60. Association for Computing Machinery.

Diederik P Kingma and Jimmy Ba. 2015. Adam: a method for stochastic optimization. In *International Conference on Learning Representations*, pages 1–15. OpenReview.net.

Wolfgang Klimesch. 1999. EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Research Reviews*, 29(2-3):169–195.

Kevin Knight. 1999. Decoding complexity in word-replacement translation models. *Computational Linguistics*, 25(4):607–615.

Kevin Knight and Ishwar Chander. 1994. Automated postediting of documents. In *Conference on Artificial Intelligence*, pages 779–784. AAAI Press / The MIT Press.

Rebecca Knowles, Marina Sanchez-Torron, and Philipp Koehn. 2019. A user study of neural interactive translation prediction. *Machine Translation*, 33(1-2):135–154.

Philipp Koehn. 2009a. A process study of computer-aided translation. *Machine Translation*, 23(4):241–263.

Philipp Koehn. 2009b. A web-based interactive computer aided translation tool. In *Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 17–20.

Philipp Koehn and Barry Haddow. 2009. Interactive assistance to human translators using statistical machine translation methods. In *Machine Translation Summit*, pages 1–8. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 177–180. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 48–54. Association for Computational Linguistics.

Philipp Koehn and Jean Senellart. 2010. Convergence of translation memory and statistical machine translation. In *AMTA Workshop on MT Research and the Translation Industry*, pages 21–31. Association for Machine Translation in the Americas.

Arlene Koglin. 2015. An empirical investigation of cognitive effort required to post-edit machine translated metaphors compared to the translation of metaphors. *Translation & Interpreting*, 7(1):126–141.

Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163.

Maarit Koponen. 2012. Comparing human perceptions of post-editing effort with post-editing operations. In *Conference on Machine Translation*, pages 181–190. Association for Computational Linguistics.

Maarit Koponen. 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.

Maarit Koponen, Wilker Aziz, Luciana Ramos, and Lucia Specia. 2012. Post-editing time as a measure of cognitive effort. In *AMTA Workshop on Post-Editing Technology and Practice*, pages 11–20. Association for Machine Translation in the Americas.

Maarit Koponen, Leena Salmi, and Markku Nikulin. 2019. A product and process analysis of post-editor corrections on neural, statistical and rule-based machine translation output. *Machine Translation*, 33(1-2):61–90.

Kaisa Koskinen and Minna Ruokonen. 2017. Love letters or hate mail? Translators' technology acceptance in the light of their emotional narratives. In *Human Issues in Translation Technology*, pages 8–24. Routledge.

Arthur F Kramer. 1991. Physiological metrics of mental workload: a review of recent progress. *Multiple-Task Performance*, pages 279–328.

Hans P Krings. 2001. *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*. Kent State University Press.

Jan-Louis Kruger and Stephen Doherty. 2016. Measuring cognitive load in the presence of educational video: towards a multimodal methodology. *Australasian Journal of Educational Technology*, 32(6):19–31.

Jan-Louis Kruger, Stephen Doherty, Wendy Fox, and Peter De Lissa. 2018. Multimodal measurement of cognitive load during subtitle processing. In *Innovation and Expansion in Translation Process Research*, pages 267–294. John Benjamins Publishing Company.

A Kumaran, K Saravanan, and Sandor Maurice. 2008. wikiBABEL: community creation of multilingual data. In *International Symposium on Wikis*, pages 1–11. Association for Computing Machinery.

Fan-Ray Kuo, Gwo-Jen Hwang, Yen-Jung Chen, and Shu-Ling Wang. 2007. Standards and tools for context-aware ubiquitous learning. In *Conference on Advanced Learning Technologies*, pages 704–705. IEEE.

Isabel Lacruz and Gregory M Shreve. 2014. Pauses and cognitive effort in post-editing. In *Post-Editing of Machine Translation: Processes and Applications*, pages 246–273. Cambridge Scholars Publishing.

Isabel Lacruz, Gregory M Shreve, and Erik Angelone. 2012. Average pause ratio as an indicator of cognitive effort in post-editing: a case study. In *AMTA Workshop on Post-Editing Technology and Practice*, pages 21–30. Association for Machine Translation in the Americas.

John D Lafferty, Andrew McCallum, and Fernando C N Pereira. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*, pages 282–289. International Machine Learning Society.

Antonio Lagarda, Vicent Alabau, Francisco Casacuberta, Roberto Silva, and Enrique Díaz-de Liaño. 2009. Statistical post-editing of a rule-based machine translation system. In *Conference of the North American Chapter of the Association of Computational Linguistics*, pages 217–220. The Association for Computational Linguistics.

Elina Lagoudaki. 2006. Translation memories survey 2006: user's perceptions around TM usage. *Translating and the Computer*, 28.

Elina Lagoudaki. 2009a. Translation editing environments. In *Machine Translation Summit Workshop on Beyond Translation Memories*, pages 1–9. Association for Computational Linguistics.

Pelagia Maria Lagoudaki. 2009b. *Expanding the possibilities of translation memory systems: from the translators wishlist to the developers design*. Ph.D. thesis, Imperial College London.

Denis Lalanne, Laurence Nigay, Philippe Palanque, Peter Robinson, Jean Vanderdonckt, and Jean-François Ladry. 2009. Fusion engines for multimodal input: a survey. In *International Conference on Multimodal Interfaces*, pages 153–160. Association for Computing Machinery.

Philippe Langlais and Guy Lapalme. 2002. TransType: development-evaluation cycles to boost translator's productivity. *Machine Translation*, 17(2):77–98.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research*, 67:653–672.

Samuel Läubli, Mark Fishel, Gary Massey, Maureen Ehrensberger-Dow, and Martin Volk. 2013. Assessing post-editing efficiency in a realistic translation environment. In *Machine Translation Summit Workshop on Post-Editing Technology and Practice*, pages 83–91. Association for Computational Linguistics.

Samuel Läubli, Patrick Simianer, Joern Wuebker, Geza Kovacs, Rico Sennrich, and Spence Green. 2021. The impact of text presentation on translator performance. *Target: International Journal of Translation Studies*.

Veronica Lawson. 1982. *Practical Experience of Machine Translation*. North-Holland Amsterdam.

Matthieu LeBlanc. 2017. 'I can't get no satisfaction!' Should we blame translation technologies or shifting business practices? In *Human Issues in Translation Technology*, pages 63–80. Routledge.

Dongjun Lee. 2020. Two-phase cross-lingual language model fine-tuning for machine translation quality estimation. In *Conference on Machine Translation*, pages 1024–1028. Association for Computational Linguistics.

Dongjun Lee, Junhyeong Ahn, Heesoo Park, and Jaemin Jo. 2021. IntelliCAT: intelligent machine translation post-editing with quality estimation and translation suggestion. In *Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing: System Demonstrations*, pages 11–19. Association for Computational Linguistics.

WonKee Lee, Jaehun Shin, and Jong-Hyeok Lee. 2019. Transformer-based automatic post-editing model with joint encoder and multi-source attention of decoder. In *Conference on Machine Translation*, pages 114–119. Association for Computational Linguistics.

Pierre-Majorique Léger, Fred D Davis, Julien Perret, and Mary Dunaway. 2010. Psychophysiological measures of cognitive absorption. In *Special Interest Group on Human-Computer Interaction*, pages 1–5. Association for Information Systems.

Miikael Lehto and Martti Lehto. 2017. Health information privacy of activity trackers. In *European Conference on Cyber Warfare and Security*, pages 243–251. Academic Conferences International Limited.

Luis A Leiva, Vicent Alabau, and Enrique Vidal. 2013. Error-proof, high-performance, and context-aware gestures for interactive text edition. In *CHI Extended Abstracts on Human Factors in Computing Systems*, page 1227–1232. Association for Computing Machinery.

Derick Leony, Abelardo Pardo Sánchez, Hugo A Parada Gélvez, and Carlos Delgado Kloos. 2012. A widget to recommend learning resources based on the learner affective state. In *International Workshop on Motivational and Affective Aspects in Technology Enhanced Learning*, pages 1–4. CEUR-Workshop Proceedings.

Jimmie Leppink. 2017. Cognitive load theory: practical implications and an important challenge. *Journal of Taibah University Medical Sciences*, 12(5):385–391.

Jindřich Libovický, Jindřich Helcl, and David Mareček. 2018. Input combination strategies for multi-source transformer decoder. In *Conference on Machine Translation*, pages 253–260. Association for Computational Linguistics.

Jindřich Libovický, Jindřich Helcl, Marek Tlustý, Ondřej Bojar, and Pavel Pecina. 2016. CUNI system for WMT16 automatic post-editing and multimodal translation tasks. In *Conference on Machine Translation*, pages 646–654. Association for Computational Linguistics.

Chantal Lidynia, Philipp Brauner, and Martina Ziefle. 2017. A step in the right direction – understanding privacy concerns and perceived sensitivity of fitness trackers. In *International Conference on Applied Human Factors and Ergonomics*, pages 42–53. Springer.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Annual Meeting of the Association for Computational Linguistics*, pages 605–612. Association for Computational Linguistics.

Arle R Lommel and Donald A DePalma. 2016. Europe's leading role in machine translation. *Common Sense Advisory*.

Robyn Longhurst. 2003. Semi-structured interviews and focus groups. In *Key Methods in Geography*, pages 143–156. SAGE Publications Sage CA.

António V Lopes, M Amin Farajian, Gonçalo M Correia, Jonay Trénous, and André F T Martins. 2019. Unbabel's submission to the WMT2019 APE shared task: BERT-based encoder-decoder for automatic post-editing. In *Conference on Machine Translation*, pages 120–125. Association for Computational Linguistics.

Byron Lowens, Vivian Genaro Motti, and Kelly Caine. 2017. Wearable privacy: skeletons in the data closet. In *International Conference on Healthcare Informatics*, pages 295–304. IEEE.

Yu Lu, Sen Zhang, Zhiqiang Zhang, Wendong Xiao, and Shengquan Yu. 2017. A framework for learning analytics using commodity wearable devices. *Sensors*, 17(6):1382.

Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *International Workshop on Spoken Language Translation*, pages 76–79. Association for Computational Linguistics.

Pedro Macizo and M Teresa Bajo. 2006. Reading for repetition and reading for translation: do they involve the same processes? *Cognition*, 99(1):1–34.

David J Mack, Sandro Belfanti, and Urs Schwarz. 2017. The effect of sampling rate and lowpass filters on saccades – a modeling approach. *Behavior Research Methods*, 49(6):2146–2162.

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, Jindrich Helcl, and Ankit Srivastava. 2017. Machine translation: phrase-based, rule-based and neural approaches with linguistic evaluation. *Cybernetics and Information Technologies*, 17(2):28–43.

Sandra P Marshall. 2002. The index of cognitive activity: measuring cognitive workload. In *Conference on Human Factors and Power Plants*, pages 5–9. IEEE.

Mercedes Garcia Martinez, Karan Singla, Aniruddha Tammewar, Bartolomé Mesa-Lao, Ankita Thakur, MA Anusuya, Banglore Srinivas, and Michael Carl. 2014. SEECAT: ASR & eye-tracking enabled computer assisted translation. In *Conference of the European Association for Machine Translation*, pages 81–88. European Association for Machine Translation.

Christopher Davey Mellinger. 2014. *Computer-Assisted Translation: An Empirical Investigation of Cognitive Effort*. Kent State University.

Bartolomé Mesa-Lao. 2012. The next generation translator's workbench: post-editing in CASMACAT v. 1.0. *Translating and the Computer*, 34:1–16.

Bartolomé Mesa-Lao. 2014. Speech-enabled computer-aided translation: a satisfaction survey with post-editor trainees. In *EACL Workshop on Humans and Computer-assisted Translation*, pages 99–103. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119. Curran Associates, Inc.

Philipp Mock, Peter Gerjets, Maike Tibus, Ulrich Trautwein, Korbinian Möller, and Wolfgang Rosenstiel. 2016. Using touchscreen interaction data to predict cognitive workload. In *International Conference on Multimodal Interaction*, pages 349–356. Association for Computing Machinery.

Barbara Moissa, Geoffray Bonnin, and Anne Boyer. 2019. Exploiting wearable technologies to measure and predict students' effort. In *Perspectives on Wearable Enhanced Learning*, pages 411–431. Springer.

Joss Moorkens. 2018. What to expect from neural machine translation: a practical in-class translation evaluation exercise. *The Interpreter and Translator Trainer*, 12(4):375–387.

Joss Moorkens and Sharon O'Brien. 2013. User attitudes to the post-editing interface. In *Machine Translation Summit Workshop on Post-Editing Technology and Practice*, pages 19–25. Association for Computational Linguistics.

Joss Moorkens and Sharon O'Brien. 2015. Post-editing evaluations: trade-offs between novice and professional participants. In *Conference of the European Association for Machine Translation*, pages 75–81. European Association for Machine Translation.

Joss Moorkens and Sharon O'Brien. 2017. Assessing user interface needs of post-editors of machine translation. In *Human Issues in Translation Technology*, pages 127–148. Routledge.

Joss Moorkens, Sharon O'Brien, Igor AL da Silva, Norma B de Lima Fonseca, and Fabio Alves. 2015. Correlations of perceived post-editing effort with measurements of actual effort. *Machine Translation*, 29(3-4):267–284.

Joss Moorkens, Sharon O'Brien, and Joris Vreeke. 2016. Developing and testing Kanjingo: a mobile app for post-editing. *Revista Tradumàtica*, 14:58–65.

Meredith Ringel Morris. 2012. Web on the wall: insights from a multimodal interaction elicitation study. In *International Conference on Interactive Tabletops and Surfaces*, pages 95–104. Association for Computing Machinery.

Meredith Ringel Morris, Andreea Danielescu, Steven Drucker, Danyel Fisher, Bongshin Lee, and Jacob O Wobbrock. 2014. Reducing legacy bias in gesture elicitation studies. *Interactions*, 21(3):40–45.

Brian Mossop. 2007. Empirical studies of revision: what we know and need to know. *The Journal of Specialised Translation*, 8:5–20.

Vivian Genaro Motti and Kelly Caine. 2015. Users' privacy concerns about wearables. In *International Conference on Financial Cryptography and Data Security*, pages 231–244. Springer.

LJM Mulder. 1992. Measurement and analysis methods of heart rate and respiration for use in applied environments. *Biological Psychology*, 34(2):205–236.

Makoto Nagao. 1984. A framework of a mechanical translation between Japanese and English by analogy principle. In *International NATO Symposium on Artificial and Human Intelligence*, page 173–180. Elsevier.

Laura M Naismith and Rodrigo B Cavalcanti. 2015. Validity of cognitive load measures in simulation-based training: a systematic review. *Academic Medicine*, 90(11):24–35.

Ángel Navarro and Francisco Casacuberta. 2021. Introducing mouse actions into interactive-predictive neural machine translation. In *Machine Translation Summit*, pages 270–281. Association for Machine Translation in the Americas.

Tapas Nayek, Sudip Kumar Naskar, Santanu Pal, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. 2015. CATaLog: new approaches to TM and post editing interfaces. In *Workshop on Natural Language Processing for Translation Memories*, pages 36–42. Association for Computational Linguistics.

Matteo Negri, Marco Turchi, Nicola Bertoldi, and Marcello Federico. 2018a. Online neural automatic post-editing for neural machine translation. In *Italian Conference on Computational Linguistics*, pages 1–6. CEUR-Workshop Proceedings.

Matteo Negri, Marco Turchi, Rajen Chatterjee, and Nicola Bertoldi. 2018b. ESCAPE: a large-scale synthetic corpus for automatic post-editing. In *International Conference on Language Resources and Evaluation*, pages 24–30. European Language Resources Association.

Jan Niehues, Eunah Cho, Thanh-Le Ha, and Alex Waibel. 2016. Pre-translation for neural machine translation. In *International Conference on Computational Linguistics*, pages 1828–1836. The COLING Organizing Committee.

Michael Nielsen, Moritz Störring, Thomas B Moeslund, and Erik Granum. 2003. A procedure for developing intuitive and ergonomic gesture interfaces for HCI. In *International Gesture Workshop*, pages 409–420. Springer.

Sharon O'Brien. 2005. Methodologies for measuring the correlations between post-editing effort and machine translatability. *Machine Translation*, 19(1):37–58.

Sharon O'Brien. 2006a. Eye-tracking and translation memory matches. *Perspectives: Studies in Translatology*, 14(3):185–205.

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

Sharon O'Brien and Joss Moorkens. 2014. Towards intelligent post-editing interfaces. In *Proceedings of the FIT World Congress*, pages 131–137. BDÜ Fachverlag.

Sharon O'Brien, Joss Moorkens, and Joris Vreeke. 2014. Kanjingo – a mobile app for post-editing. In *Conference of the European Association for Machine Translation*, pages 137–141. European Association for Machine Translation.

Sharon O'Brien, Minako O'Hagan, and Marian Flanagan. 2010. Keeping an eye on the UI design of translation memory: how do translators use the "concordance" feature? In *European Conference on Cognitive Ergonomics*, pages 187–190. Association for Computing Machinery.

Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28. Association for Computational Linguistics.

Franz Josef Och, Nicola Ueffing, and Hermann Ney. 2001. An efficient A* search algorithm for statistical machine translation. In *ACL Workshop on Data-Driven Methods in Machine Translation*, pages 1–8. Association for Computational Linguistics.

Michael Ortega and Laurence Nigay. 2009. AirMouse: finger gesture for 2D and 3D interaction. In *IFIP Conference on Human-Computer Interaction*, pages 214–227. Springer.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: a fast, extensible toolkit for sequence modeling. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 48–53. Association for Computational Linguistics.

Sharon Oviatt. 2003. Advances in robust multimodal interface design. *Computer Graphics and Applications*, 23(5):62–68.

Sharon Oviatt, Bjorn Schuller, Philip Cohen, Daniel Sonntag, Gerasimos Potamianos, and Antonio Kruger. 2017. Introduction: scope, trends, and paradigm shift in the field of computer interfaces. In *The Handbook of Multimodal-Multisensor Interfaces, Volume 1: Foundations, User Modeling, and Common Modality Combinations*, page 1–15. Association for Computing Machinery.

Fred Paas, Juhani E Tuovinen, Huib Tabbers, and Pascal WM Van Gerven. 2003. Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, 38(1):63–71.

Fred GWC Paas and Jeroen JG Van Merriënboer. 1994. Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review*, 6(4):351–371.

Santanu Pal. 2018. *A hybrid machine translation framework for an improved translation workflow*. Ph.D. thesis, Saarland University.

Santanu Pal, Nico Herbig, Antonio Krüger, and Josef van Genabith. 2018. A transformer-based multi-source automatic post-editing system. In *Conference on Machine Translation*, pages 827–835. Association for Computational Linguistics.

Santanu Pal, Sudip Naskar, and Josef van Genabith. 2015. UdS-Sant: English–German hybrid machine translation system. In *Conference on Machine Translation*, pages 152–157. Association for Computational Linguistics.

Santanu Pal, Sudip Kumar Naskar, and Josef van Genabith. 2016a. Multi-engine and multi-alignment based automatic post-editing and its impact on translation productivity. In *International Conference on Computational Linguistics*, pages 2559–2570. The COLING Organizing Committee.

Santanu Pal, Sudip Kumar Naskar, Mihaela Vela, and Josef van Genabith. 2016b. A neural network based approach to automatic post-editing. In *Annual Meeting of the Association for Computational Linguistics: Short Papers*, pages 281–286. Association for Computational Linguistics.

Santanu Pal, Sudip Kumar Naskar, Marcos Zampieri, Tapas Nayak, and Josef van Genabith. 2016c. CATaLog Online: a web-based CAT tool for distributed translation with data capture for APE and translation process research. In *International Conference on Computational Linguistics: System Demonstrations*, pages 98–102. The COLING Organizing Committee.

Santanu Pal, Hongfei Xu, Nico Herbig, Antonio Krüger, and Josef van Genabith. 2019. USAAR-DFKI – the Transference architecture for English–German automatic post-editing. In *Conference on Machine Translation*, pages 124–131. Association for Computational Linguistics.

Santanu Pal, Hongfei Xu, Nico Herbig, Sudip Kumar Naskar, Antonio Krüger, and Josef van Genabith. 2020. The Transference architecture for automatic post-editing. In *International Conference on Computational Linguistics*, pages 5963–5974. The COLING Organizing Committee.

Santanu Pal, Marcos Zampieri, and Josef van Genabith. 2016d. USAAR: an operation sequential model for automatic statistical post-editing. In *Conference on Machine Translation*, pages 759–763. Association for Computational Linguistics.

Santanu Pal, Marcos Zampieri, Sudip Kumar Naskar, Tapas Nayak, Mihaela Vela, and Josef van Genabith. 2016e. CATaLog Online: porting a post-editing tool to the web. In *International Conference on Language Resources and Evaluation*, pages 599–604. European Language Resources Association.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Carla Parra Escartín and Manuel Arcedillo. 2015a. Living on the edge: productivity gain thresholds in machine translation evaluation metrics. In *Workshop on Post-editing Technology and Practice*, pages 46–56. Association for Machine Translation in the Americas.

Carla Parra Escartín and Manuel Arcedillo. 2015b. Machine translation evaluation made fuzzier: a study on post-editing productivity and evaluation metrics in commercial settings. In *Machine Translation Summit*, pages 131–141. Association for Computational Linguistics.

Carla Parra Escartín, Hanna Béchara, and Constantin Orasan. 2017. Questing for quality estimation: A user study. *The Prague Bulletin of Mathematical Linguistics*, 108:343–354.

Alfredo J Perez and Sherali Zeadally. 2018. Privacy issues and solutions for consumer wearables. *IT Professional*, 20(4):46–56.

Ana M Pernas, Adenauer C Yamin, João LB Lopes, and Jose P M de Oliveira. 2014. A semantic approach for learning situation detection. In *Conference on Advanced Information Networking and Applications*, pages 1119–1126. IEEE.

Bastian Pfleging, Drea K Fekety, Albrecht Schmidt, and Andrew L Kun. 2016. A model relating pupil diameter to mental workload and lighting conditions. In *Conference on Human Factors in Computing Systems*, pages 5776–5788. Association for Computing Machinery.

Hélène Pielmeier and Paul Daniel O'Mara. 2010. The state of the linguist supply chain. *Common Sense Advisory*.

Martin Popel. 2018. CUNI transformer neural MT system for WMT18. In *Conference on Machine Translation*, pages 482–487. Association for Computational Linguistics.

Maja Popovic, Arle Lommel, Aljoscha Burchardt, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Relations between different types of post-editing operations, cognitive effort and temporal effort. In *Conference of the European Association for Machine Translation*, pages 191–198. European Association for Machine Translation.

Alexander Prange, Michael Barz, and Daniel Sonntag. 2018. A categorisation and implementation of digital pen features for behaviour characterisation. *CoRR*, abs/1810.03970:1–42.

Andrew Raij, Animikh Ghosh, Santosh Kumar, and Mani Srivastava. 2011. Privacy risks emerging from the adoption of innocuous wearable sensors in the mobile environment. In *Conference on Human Factors in Computing Systems*, pages 11–20. Association for Computing Machinery.

Christopher M Rives, Craig T Brown, Dustin L Hoffman, and Peter M On. 2014. Gesture based edit mode. US Patent 8,707,170.

Manuel Rodrigues, Sérgio Gonçalves, Davide Carneiro, Paulo Novais, and Florentino Fdez-Riverola. 2013. Keystrokes and clicks: measuring stress on e-learning students. In *Management Intelligent Systems*, pages 119–126. Springer.

Rudolf Rosa, David Mareček, and Ondřej Dušek. 2012. DEPFIX: a system for automatic correction of Czech MT outputs. In *Conference on Machine Translation*, pages 362–368. Association for Computational Linguistics.

Dennis W Rowe, John Sibert, and Don Irwin. 1998. Heart rate variability: indicator of user state as an aid to human-computer interaction. In *Conference on Human Factors in Computing Systems*, pages 480–487. Association for Computing Machinery.

Pilar Sánchez-Gijón, Joss Moorkens, and Andy Way. 2019. Post-editing neural machine translation versus translation memory segments. *Machine Translation*, 33(1-2):31–59.

Germán Sanchis-Trilles, Vicent Alabau, Christian Buck, Michael Carl, Francisco Casacuberta, Mercedes García-Martínez, Ulrich Germann, Jesús González-Rubio, Robin L Hill, Philipp Koehn, Luis A Leiva, Bartolomé Mesa-Lao, Daniel Ortiz-Martínez, Herve Saint-Amand, Chara Tsoukala, and Enrique Vidal. 2014. Interactive translation prediction versus conventional post-editing in practice: a study with the CASMACAT workbench. *Machine Translation*, 28(3-4):217–235.

Eva-Maria Schomakers, Chantal Lidynia, and Martina Ziefle. 2018. Exploring the acceptance of mHealth applications – do acceptance patterns vary depending on context? In *International Conference on Applied Human Factors and Ergonomics*, pages 53–64. Springer.

Lasse Schou, Barbara Dragsted, and Michael Carl. 2009. Ten years of Translog. *Copenhagen Studies in Language*, 37:37–51.

Douglas Schuler and Aki Namioka. 1993. *Participatory Design: Principles and Practices*. CRC Press.

Holger Schultheis and Anthony Jameson. 2004. Assessing cognitive load in adaptive hypermedia systems: physiological and behavioral methods. In *International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, pages 225–234. Springer.

Lane Schwartz, Isabel Lacruz, and Tatyana Bystrova. 2015. Effects of word alignment visualization on post-editing quality & speed. In *Machine Translation Summit*, pages 186–199. Association for Computational Linguistics.

Benjamin Alun Screen. 2016. What does translation memory do to translation? The effect of translation memory output on specific aspects of the translation process. *Translation & Interpreting*, 8(1):1–18.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The University of Edinburgh's neural MT systems for WMT17. In *Conference on Machine Translation*, pages 389–399. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725. Association for Computational Linguistics.

Fred Shaffer and JP Ginsberg. 2017. An overview of heart rate variability metrics and norms. *Frontiers in Public Health*, 5:1–17.

Ram Pratap Sharma and Gyanendra K Verma. 2015. Human computer interaction using hand gesture. *Procedia Computer Science*, 54:721–727.

Liping Shen, Victor Callaghan, and Ruimin Shen. 2008. Affective e-learning in residential and pervasive computing environments. *Information Systems Frontiers*, 10(4):461–472.

Liping Shen, Minjuan Wang, and Ruimin Shen. 2009. Affective e-learning: using "emotional" data to improve learning in pervasive learning environment. *Journal of Educational Technology & Society*, 12(2):176–189.

Raksha Shenoy. 2021. ImpoWord MTQE: impact of word-level machine translation quality estimation on post-editing effort. Master's thesis, Saarland University.

Raksha Shenoy, Nico Herbig, Antonio Krüger, and Josef van Genabith. 2021. Investigating the helpfulness of word-level quality estimation for post-editing machine translation output. In *Conference on Empirical Methods in Natural Language Processing*, pages 10173–10185. Association for Computational Linguistics.

Yu Shi, Natalie Ruiz, Ronnie Taib, Eric Choi, and Fang Chen. 2007. Galvanic skin response (GSR) as an index of cognitive load. In *CHI Extended Abstracts on Human Factors in Computing Systems*, pages 2651–2656. Association for Computing Machinery.

Jaehun Shin and Jong-Hyeok Lee. 2018. Multi-encoder transformer network for automatic post-editing. In *Conference on Machine Translation*, pages 853–858. Association for Computational Linguistics.

Dimitar Shterionov, Félix Do Carmo, Joss Moorkens, Eric Paquin, Dag Schmidtke, Declan Groves, and Andy Way. 2019. When less is more in neural quality estimation of machine translation. an industry case study. In *Machine Translation Summit*, pages 228–235. European Association for Machine Translation.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007a. Statistical phrase-based post-editing. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 508–515. Association for Computational Linguistics.

Michel Simard and Pierre Isabelle. 2009. Phrase-based machine translation in a computer-assisted translation environment. In *Machine Translation Summit*, pages 120–127. Association for Computational Linguistics.

Michel Simard, Nicola Ueffing, Pierre Isabelle, and Roland Kuhn. 2007b. Rule-based translation with statistical phrase-based post-editing. In *Conference on Machine Translation*, pages 203–206. Association for Computational Linguistics.

Shyamli Sindhwani, Christof Lutteroth, and Gerald Weber. 2019. ReType: quick text editing with keyboard and gaze. In *Conference on Human Factors in Computing Systems*, pages 1–13. Association for Computing Machinery.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Conference of the Association for Machine Translation in the Americas*, pages 223–231. Association for Machine Translation in the Americas.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Conference on Machine Translation*, pages 259–268. Association for Computational Linguistics.

Erin Solovey, Paul Schermerhorn, Matthias Scheutz, Angelo Sassaroli, Sergio Fantini, and Robert Jacob. 2012. Brainput: enhancinge interactive systems with streaming fNIRS brain input. In *Conference on Human Factors in Computing Systems*, pages 2193–2202. Association for Computing Machinery.

T Soukupova and Jan Cech. 2016. Real-time eye blink detection using facial landmarks. In *Computer Vision Winter Workshop*, pages 1–8.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André FT Martins. 2020. Findings of the WMT 2020 shared task on quality estimation. In *Conference on Machine Translation*, pages 743–764. Association for Computational Linguistics.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine Translation*, 24(1):39–50.

Lucia Specia and Kashif Shah. 2018. Machine translation quality estimation: applications and future perspectives. In *Translation Quality Assessment*, pages 201–235. Springer.

Lucia Specia, Marco Turchi, Nicola Cancedda, Marc Dymetman, and Nello Cristianini. 2009. Estimating the sentence-level quality of machine translation systems. In *Conference of the European Association for Machine Translation*, pages 28–37. European Association for Machine Translation.

Martin Spüler, Carina Walter, Wolfgang Rosenstiel, Peter Gerjets, Korbinian Moeller, and Elise Klein. 2016. EEG-based prediction of cognitive workload induced by arithmetic: a step towards online adaptation in numerical learning. *ZDM*, 48(3):267–278.

Els Stuyven, Koen Van der Goten, André Vandierendonck, Kristl Claeys, and Luc Crevits. 2000. The effect of cognitive load on saccadic eye movements. *Acta Psychologica*, 104(1):69–85.

Keh-Yih Su, Ming-Wen Wu, and Jing-Shin Chang. 1992. A new quantitative quality measure for machine translation systems. In *International Conference on Computational Linguistics*, pages 433–439. The COLING Organizing Committee.

John Sweller. 1988. Cognitive load during problem solving: effects on learning. *Cognitive Science*, 12(2):257–285.

John Sweller, Jeroen JG Van Merrienboer, and Fred GWC Paas. 1998. Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3):251–296.

Midori Tatsumi. 2009. Correlation between automatic evaluation metric scores, post-editing speed, and some other factors. In *Machine Translation Summit*, pages 332–339. Association for Computational Linguistics.

Amirhossein Tebbifakhr, Ruchit Agrawal, Rajen Chatterjee, Matteo Negri, and Marco Turchi. 2018. Multi-source transformer with combined losses for automatic post editing. In *Conference on Machine Translation*, pages 859–865. Association for Computational Linguistics.

Carlos Teixeira and Sharon O'Brien. 2017. The impact of MT quality estimation on post-editing effort. In *Machine Translation Summit*, pages 217–239. Asia-Pacific Association for Machine Translation.

Carlos SC Teixeira, Joss Moorkens, Daniel Turner, Joris Vreeke, and Andy Way. 2019. Creating a multimodal translation tool and testing machine translation integration using touch and voice. *Informatics*, 6(13):1–21.

Irina Temnikova. 2010. Cognitive evaluation approach for a controlled language post-editing experiment. In *International Conference on Language Resources and Evaluation*, pages 3485–3490. European Language Resources Association.

Dimitri Theologitis. 1998. Language tools at the EC translation service: the theory and the practice. *Translating and the Computer*, 18:1–16.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Alexander Zubiaga, and Hassan Sawaf. 1997. Accelerated DP based search for statistical translation. In *European Conference on Speech Communication and Technology*, pages 2667–2670. International Speech Communication Association.

Antonio Toral. 2019. Post-editese: an exacerbated translationese. In *Machine Translation Summit*, pages 273–281. European Association for Machine Translation.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018a. Attaining the unattainable? Reassessing claims of human parity in neural machine translation. In *Conference on Machine Translation*, pages 113–123. Association for Computational Linguistics.

Antonio Toral and Víctor M Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Conference of the European Chapter of the Association for Computational Linguistics*, pages 1063–1073. Association for Computational Linguistics.

Antonio Toral, Martijn Wieling, and Andy Way. 2018b. Post-editing effort of a novel with statistical and neural machine translation. *Frontiers in Digital Humanities*, 5(1):1–11.

Olga Torres-Hostench, Joss Moorkens, Sharon O'Brien, and Joris Vreeke. 2017. Testing interaction with a mobile MT post-editing app. *Translation & Interpreting*, 9(2):138–150.

Marco Turchi, Matteo Negri, and Marcello Federico. 2014. Data-driven annotation of binary MT quality estimation corpora based on human post-editions. *Machine Translation*, 28(3-4):281–308.

Marco Turchi, Matteo Negri, and Marcello Federico. 2015. MT quality estimation for computer-assisted translation: does it really help? In *Annual Meeting of the Association for Computational Linguistics*, pages 530–535. Association for Computational Linguistics.

Emmanuel Sebastian Udoh and Abdulwahab Alkharashi. 2016. Privacy risk awareness and the behavior of smartwatch users: a case study of Indiana University students. In *Future Technologies Conference*, pages 926–931. IEEE.

Masao Utiyama, Graham Neubig, Takashi Onishi, and Eiichiro Sumita. 2011. Searching translation memories for paraphrases. In *Machine Translation Summit*, pages 325–331. Association for Computational Linguistics.

Tamara Van Gog, Femke Kirschner, Liesbeth Kester, and Fred Paas. 2012. Timing and frequency of mental effort measurement: evidence in favour of repeated measures. *Applied Cognitive Psychology*, 26(6):833–839.

Karl F Van Orden, Wendy Limbert, Scott Makeig, and Tzyy-Ping Jung. 2001. Eye activity correlates of workload during a visuospatial memory task. *Human Factors*, 43(1):111–121.

Tom Vanallemeersch and Vincent Vandeghinste. 2014. Improving fuzzy matching through syntactic knowledge. *Translating and the Computer*, 36:217–227.

Vincent Vandeghinste, Tom Vanallemeersch, Liesbeth Augustinus, Bram Bulté, Frank Van Eynde, Joris Pelemans, Lyan Verwimp, Patrick Wambacq, Geert Heyman, Marie-Francine Moens, Iulianna van der Lek-Ciudin, Frieda Steurs, Ayla Rigouts Terryn, Els Lefever, Arda Tezcan, Lieve Macken, Véronique Hoste, Joke Daems, Joost Buysschaert, Sven Coppers, Jan Van den Bergh, and Kris Luyten. 2019. Improving the translation environment for professional translators. *Informatics*, 6(2):1–36.

Vincent Vandeghinste, Tom Vanallemeersch, Liesbeth Augustinus, Joris Pelemans, Geert Heyman, Iulianna van der Lek-Ciudin, Arda Tezcan, Donald Degraen, Jan van den Bergh, Lieve Macken, Els Lefever, Marie-Francine Moens, Patrick Wambacq, Frieda Steurs, and Frank Coninx, Karin andVan Eynde. 2016. SCATE – Smart Computer-Aided Translation Environment. *Baltic Journal of Modern Computing*, 4(2):382–382.

Dusan Varis and Ondřej Bojar. 2017. CUNI system for WMT17 automatic post-editing task. In *Conference on Machine Translation*, pages 661–666. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Annual Conference on Neural Information Processing Systems*, pages 5998–6008. Curran Associates, Inc.

Radu-Daniel Vatavu and Jacob O Wobbrock. 2015. Formalizing agreement analysis for elicitation studies: new measures, significance test, and toolkit. In *Conference on Human Factors in Computing Systems*, pages 1325–1334. Association for Computing Machinery.

Mihaela Vela, Santanu Pal, Marcos Zampieri, Sudip Kumar Naskar, and Josef van Genabith. 2019. Improving CAT tools in the translation workflow: new approaches and evaluation. In *Machine Translation Summit*, pages 8–15. European Association for Machine Translation.

Philip A Vernon. 1993. Der Zahlen-Verbindungs-Test and other trail-making correlates of general intelligence. *Personality and Individual Differences*, 14(1):35–40.

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

Lucas Nunes Vieira. 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.

Lucas Nunes Vieira and Lucia Specia. 2011. A review of translation tools from a post-editing perspective. In *Joint EM+/CNGL Workshop Bringing MT to the User*, pages 33–42. Associtaion for Computational Linguistics.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse Beam Search: decoding diverse solutions from neural sequence models. *CoRR*, abs/1610.02424:1–16.

María Viqueira Villarejo, Begoña García Zapirain, and Amaia Méndez Zorrilla. 2012. A stress sensor based on galvanic skin response (GSR) controlled by ZigBee. *Sensors*, 12(5):6075–6101.

Jessica Vitak, Yuting Liao, Priya Kumar, Michael Zimmer, and Katherine Kritikos. 2018. Privacy attitudes and data valuation among fitness tracker users. In *International Conference on Information*, pages 229–239. Springer.

Minh Tue Vo and Alex Waibel. 1993. A multi-modal human-computer interface: combination of gesture and speech recognition. In *Conference Companion on Human Factors in Computing Systems*, page 69–70. Association for Computing Machinery.

Susanne Vogel and Lars Schwabe. 2016. Learning and memory under stress: implications for the classroom. *npj Science of Learning*, 1:1–10.

Karel Vredenburg, Ji-Ye Mao, Paul W Smith, and Tom Carey. 2002. A survey of user-centered design practice. In *Conference on Human Factors in Computing Systems*, pages 471–478. Association for Computing Machinery.

Juan Pablo Wachs, Mathias Kölsch, Helman Stern, and Yael Edan. 2011. Vision-based hand-gesture applications. *Communications of the ACM*, 54(2):60–71.

Julian Wallis. 2006. *Interactive translation vs. pre-translation in the context of translation memory systems: Investigating the effects of translation method on productivity, quality and translator satisfaction*. Ph.D. thesis, University of Ottawa.

Carina Walter, Wolfgang Rosenstiel, Martin Bogdan, Peter Gerjets, and Martin Spüler. 2017. Online EEG-based workload adaptation of an arithmetic learning environment. *Frontiers in Human Neuroscience*, 11(1):1–11.

Jiayi Wang, Kai Fan, Bo Li, Fengming Zhou, Boxing Chen, Yangbin Shi, and Luo Si. 2018. Alibaba submission for WMT18 quality estimation task. In *Conference on Machine Translation*, pages 809–815. Association for Computational Linguistics.

Ke Wang, Jiayi Wang, Niyu Ge, Yangbin Shi, Yu Zhao, and Kai Fan. 2020. Computer assisted translation with neural quality estimation and auotmatic post-editing. In *Conference on Empirical Methods in Natural Language Processing*, pages 2175–2186. Association for Computational Linguistics.

Shiya Wang. 2021. The impact of multiple machine translation proposals in human post-editing. Master's thesis, Saarland University.

Warren Weaver. 1953. Recent contributions to the mathematical theory of communication. *ETC: A Review of General Semantics*, 11(1):261–281.

Warren Weaver. 1999. Warren Weaver Memorandum, July 1949. *MT News International*, 22:5–6.

Frank Weichert, Daniel Bachmann, Bartholomäus Rudak, and Denis Fisseler. 2013. Analysis of the accuracy and robustness of the leap motion controller. *Sensors*, 13(5):6380–6393.

Susan Welsh and Marc Prior. 2014. OmegaT for CAT beginners. *Recuperat el*, 27:1–30.

Rongxiang Weng, Hao Zhou, Shujian Huang, Lei Li, Yifan Xia, and Jiajun Chen. 2019. Correct-and-Memorize: learning to translate from interactive revisions. In *International Joint Conference on Artificial Intelligence*, pages 5255–5263. ijcai.org.

Meredydd Williams. 2018. *Exploring the influence of privacy awareness on the privacy paradox on smartwatches*. Ph.D. thesis, University of Oxford.

Jacob O Wobbrock, Htet Htet Aung, Brandon Rothrock, and Brad A Myers. 2005. Maximizing the guessability of symbolic input. In *CHI Extended Abstracts on Human Factors in Computing Systems*, pages 1869–1872. Association for Computing Machinery.

Jacob O Wobbrock, Meredith Ringel Morris, and Andrew D Wilson. 2009. User-defined gestures for surface computing. In *Conference on Human Factors in Computing Systems*, pages 1083–1092. Association for Computing Machinery.

Chih-Hung Wu, Yueh-Min Huang, and Jan-Pan Hwang. 2016a. Review of affective computing in education/learning: trends and challenges. *British Journal of Educational Technology*, 47(6):1304–1323.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016b. Google's neural machine translation system: bridging the gap between human and machine translation. *CoRR*, abs/1609.08144:1–23.

Hongfei Xu, Josef van Genabith, Deyi Xiong, and Qiuhui Liu. 2020. Analyzing word translation of transformer layers. *CoRR*, abs/2003.09586:1–12.

Hongfei Xu and Qiuhui Liu. 2019. Neutron: an implementation of the transformer translation model and its variants. *CoRR*, abs/1903.07402:1–6.

Hongfei Xu, Qiuhui Liu, and Josef van Genabith. 2019. UdS submission for the WMT 19 automatic post-editing task. In *Conference on Machine Translation*, pages 147–152. Association for Computational Linguistics.

Masaru Yamada. 2015. Can college students be post-editors? An investigation into employing language learners in machine translation plus post-editing settings. *Machine Translation*, 29(1):49–67.

Takehiro Yamakoshi, K Yamakoshi, S Tanaka, M Nogawa, Sang-Bum Park, Mariko Shibata, Y Sawada, P Rolfe, and Yasuo Hirose. 2008. Feasibility study on driver's stress detection from differential skin temperature measurement. In *Engineering in Medicine and Biology Society*, pages 1076–1079. IEEE.

Zhiwen Yu, Yuichi Nakamura, Seiie Jang, Shoji Kajita, and Kenji Mase. 2007. Ontology-based semantic recommendation for context-aware e-learning. In *International Conference on Ubiquitous Intelligence and Computing*, pages 898–907. Springer.

Marcos Zampieri and Mihaela Vela. 2014a. Quantifying the influence of MT output in the translators' performance: a case study in technical translation. In *EACL Workshop on Humans and Computer-assisted Translation*, pages 93–98. Association for Computational Linguistics.

Marcos Zampieri and Mihaela Vela. 2014b. Quantifying the influence of MT output in the translators' performance: a case study in technical translation. In *EACL Workshop on Humans and Computer-assisted Translation*, pages 93–98. Association for Computational Linguistics.

Julián Zapata. 2016. Translating on the go? Investigating the potential of multimodal mobile devices for interactive translation dictation. *Revista Tradumàtica*, 14(1):66–74.

Julián Zapata, Sheila Castilho, and Joss Moorkens. 2017. Translation dictation vs. post-editing with cloud-based voice recognition: a pilot experiment. In *Machine Translation Summit*, pages 123–136. Association for Computational Linguistics.

Anna Zaretskaya, Gloria Corpas Pastor, and Miriam Seghiri. 2015. Translators' requirements for translation technologies: results of a user survey. In *Conference New Horizons is Translation and Interpreting Studies*, pages 1–8.

Anna Zaretskaya and Míriam Seghiri. 2018. User perspective on translation tools: findings of a user survey. In Gloria Corpas and Isabel Durán, editors, *Trends in E-Tools and Resources for Translators and Interpreters*, pages 37–56. Brill.

Anna Zaretskaya, Mihaela Vela, Gloria Corpas Pastor, and Miriam Seghiri. 2016. Comparing post-editing difficulty of different machine translation errors in Spanish and German translations from English. *International Journal of Language and Linguistics*, 3(3):91–100.

Andre Zenner and Antonio Krüger. 2017. Shifty: a weight-shifting dynamic passive haptic proxy to enhance object perception in virtual reality. *Transactions on Visualization and Computer Graphics*, 23(4):1285–1294.

Michael Zimmer, Priya Kumar, Jessica Vitak, Yuting Liao, and Katie Chamberlain Kritikos. 2020. "There's nothing really they can do with this information": unpacking how users manage privacy boundaries for personal fitness information. *Information, Communication & Society*, 23(7):1020–1037.