MMPE: A Multi-Modal Interface for Post-Editing Machine Translation

Nico Herbig[1], Tim Düwel[1], Santanu Pal[1,2], Kalliopi Meladaki[1],
Mahsa Monshizadeh[2], Antonio Krüger[1], Josef van Genabith[1,2]
[1]German Research Center for Artificial Intelligence (DFKI),
Saarland Informatics Campus, Germany
[2]Department of Language Science and Technology,
Saarland University, Germany
{firstname.lastname}@dfki.de
{firstname.lastname}@uni-saarland.de

Abstract

Current advances in machine translation (MT) increase the need for translators to switch from traditional translation to post-editing (PE) of machine-translated text, a process that saves time and reduces errors. This affects the design of translation interfaces, as the task changes from mainly generating text to correcting errors within otherwise helpful translation proposals. Since this paradigm shift offers potential for modalities other than mouse and keyboard, we present MMPE, the first prototype to combine traditional input modes with pen, touch, and speech modalities for PE of MT. The results of an evaluation with professional translators suggest that pen and touch interaction are suitable for deletion and reordering tasks, while they are of limited use for longer insertions. On the other hand, speech and multi-modal combinations of select & 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 & keyboard, but not as a complete substitute.

1 Introduction

As machine translation (MT) has been making substantial improvements in recent years[1], more and more professional translators are integrating this technology into their translation workflows (Zaretskaya et al., 2016; Zaretskaya and Seghiri, 2018). The process of using a pre-translated text as a basis and improving it to create the final translation is called post-editing (PE). Older research showed a strong dislike of translators towards PE (Lagoudaki, 2009; Wallis, 2006), and more recent studies agree that translators are still cautious about PE and question its benefits (Gaspari et al., 2014; Koponen, 2012), partly because they see it as a threat to their profession (Moorkens, 2018). Experienced translators in particular exhibit rather negative attitudes (Moorkens and O'Brien, 2015). Conversely, novice translators have been shown to have more positive views on PE (Yamada, 2015). Green et al. (2013) demonstrated that some translators actually strongly prefer PE and argue that "users might have dated perceptions of MT quality".

Apart from translators' preference, productivity gains of 36% when using modern neural MT for PE (Toral et al., 2018) already result in substantial changes in translation workflows (Zaretskaya and Seghiri, 2018) and will probably continue to do so the better MT becomes. Thus, PE requires thorough investigation in terms of interface design, since the task changes from mostly text production to comparing and adapting MT and translation memory (TM) proposals, or put differently, from control to supervision. Previous elicitation-based research (Herbig et al., 2019a) investigated how translation environments could better support the PE process and found that translators envision PE interfaces relying on touch, pen, and speech input combined with mouse and keyboard as particularly useful. A small number of prototypes exploring some of these modalities also showed promising results (Teixeira et al., 2019).

This paper presents MMPE, the first translation environment combining standard mouse & keyboard input with touch, pen, and speech interactions for PE of MT. The results of a study with 11 professional translators show that participants are enthusiastic about having these alternatives, even though time measurements and subjective ratings do not always agree. Overall, pen and touch modalities are well suited for deletion and reordering operations, while speech and multi-modal interaction are suitable for insertions and replacements.

[1] WMT 2019 translation task: http://matrix.statmt.org/, accessed 16/04/2020

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1691–1702
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

2 Related Work

In this section, we present related research on translation environments and particularly focus on existing multi-modal approaches to PE.

2.1 CAT and Post-Editing

Most professional translators nowadays use so-called CAT (computer-aided translation) tools (van den Bergh et al., 2015). These provide features like MT and TM together with quality estimation and concordance functionality (Federico et al., 2014), alignments between source and MT (Schwartz et al., 2015), interactive MT offering assistance like auto-completion (Green et al., 2014b,a), or intelligibility assessments (Coppers et al., 2018; Vandeghinste et al., 2016, 2019).

While TM is still often valued higher than MT (Moorkens and O'Brien, 2017), a recent study by Vela et al. (2019) shows that professional translators who were given a choice between translation from scratch, TM, and MT, chose MT in 80% of the cases, highlighting the importance of PE of MT. Regarding the time savings achieved through PE, Zampieri and Vela (2014) find 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, and Läubli et al. (2013) find that PE also increases productivity in realistic environments. Furthermore, it has been shown that PE not only leads to reduced time but also reduces errors (Green et al., 2013).

Furthermore, PE changes the interaction pattern (Carl et al., 2010), leading to a significantly reduced amount of mouse and keyboard events (Green et al., 2013). Therefore, we believe that other modalities or combinations thereof might be more useful for PE.

2.2 Multi-Modal Approaches

Dictating translations dates back to the time when secretaries transcribed dictaphone content on a typewriter (Theologitis, 1998); however, the use of automatic speech recognition also has a long history for translation (Dymetman et al., 1994; Brousseau et al., 1995). A more recent approach, called SEECAT (Martinez et al., 2014), investigates the use of automatic speech recognition (ASR) in PE and argues that its combination with typing could boost productivity. A survey regarding speech usage with PE trainees (Mesa-Lao, 2014) finds that they have a positive attitude towards speech input and would consider adopting it, but only as a complement to other modalities. In a small-scale study, Zapata et al. (2017) found that ASR for PE was faster than ASR for translation from scratch. Due to these benefits, commercial CAT tools like memoQ and MateCat are also beginning to integrate ASR.

The CASMACAT tool (Alabau et al., 2013) allows the user to input text by writing with e-pens in a special area. A vision paper (Alabau and Casacuberta, 2012) proposes to instead use e-pens for PE sentences with few errors in place and showcases symbols that could be used for this. Studies on mobile PE via touch and speech (O'Brien et al., 2014; Torres-Hostench et al., 2017) show that participants especially liked reordering words through touch drag and drop, and preferred voice when translating from scratch, but used the iPhone keyboard for small changes. Zapata (2016) also explores the use of voice- and touch-enabled devices; however, the study did not focus on PE, and used Microsoft Word instead of a proper CAT environment.

Teixeira et al. (2019) explore a combination of touch and speech for translation from scratch, translation using TM, and translation using MT. In their studies, touch input received poor feedback since (a) their tile view (where each word is a tile that can be dragged around) made reading more complicated, and (b) touch insertions were rather complex to achieve within their implementation. 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.

The results of an elicitation study by Herbig et al. (2019a) indicate that pen, touch, and speech interaction should be combined with mouse and keyboard to improve PE of MT. In contrast, other modalities like eye tracking or gestures were seen as less promising.

In summary, previous research suggests that professional translators should switch to PE to increase productivity and reduce errors; however, translators themselves are not always eager to do so. It has been argued that the PE process might be better supported by using different modalities in addition to the common mouse and keyboard approaches, and an elicitation study suggests concrete modalities that should be well suited for various editing tasks. A few of these modalities have already been explored in practice, showing promising results. However, the elicited combination of pen, touch, and speech, together with mouse and keyboard, has not yet been implemented and evaluated.

3 The MMPE Prototype

We present the MMPE prototype (see Figure 1) which combines these modalities for PE of MT. A more detailed description of the prototype can be found in Herbig et al. (2020), and a video demonstration is available at https://youtu.be/H2YM2R8Wfd8.

3.1 Apparatus & Overall Layout

On the software side, we decided to use Angular for the frontend, and node.js for the backend. As requested in Herbig et al. (2019a), we use a large tiltable touch & pen screen for the study (see Figure 1b): the Wacom Cintiq Pro 32 inch display with the Flex Arm that allows the screen to be tilted and moved flat on the table, or to be moved up to work in a standing position. We further use the Sennheiser PC 8 Headset for speech input. The goal of this hardware setup was to limit induced bias as much as possible, in order to get results on the modalities and not on a flawed apparatus.

We implemented a horizontal source-target layout (see Figure 1a), where each segment's status (unedited, edited, confirmed) is visualized between source and target. On the far right, support tools are offered as requested in Herbig et al. (2019a): (1) the unedited MT output, to which the users can revert their editing using a button, and (2) a corpus combined with a dictionary.

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 (Herbig et al. (2019a); see Figure 1a). 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 and keyboard input (right). Both 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 be interpreted as both handwriting (middle) and reordering (right). 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).

3.2 Left Target View: Handwriting

For handwriting recognition (see Figure 1c), we use the MyScript Interactive Ink SDK. Apart from merely recognizing the written input, it offers gestures[2] like strike-through or scribble for deletions. For inserting words, one can directly write into an empty space, or create such a space first by breaking the line (draw a long line from top to bottom), and then handwriting the word. All changes are immediately interpreted, i.e., striking through a word deletes it immediately, instead of showing it in a struck-through visualization. The editor further shows the recognized text immediately at the very top of the drawing view in a small gray font, where alternatives for the current recognition are offered. Apart from using the pen, the user can also use his/her finger or the mouse on the left-hand editing view for handwriting.

3.3 Right Target View: Touch Reordering, Mouse & Keyboard

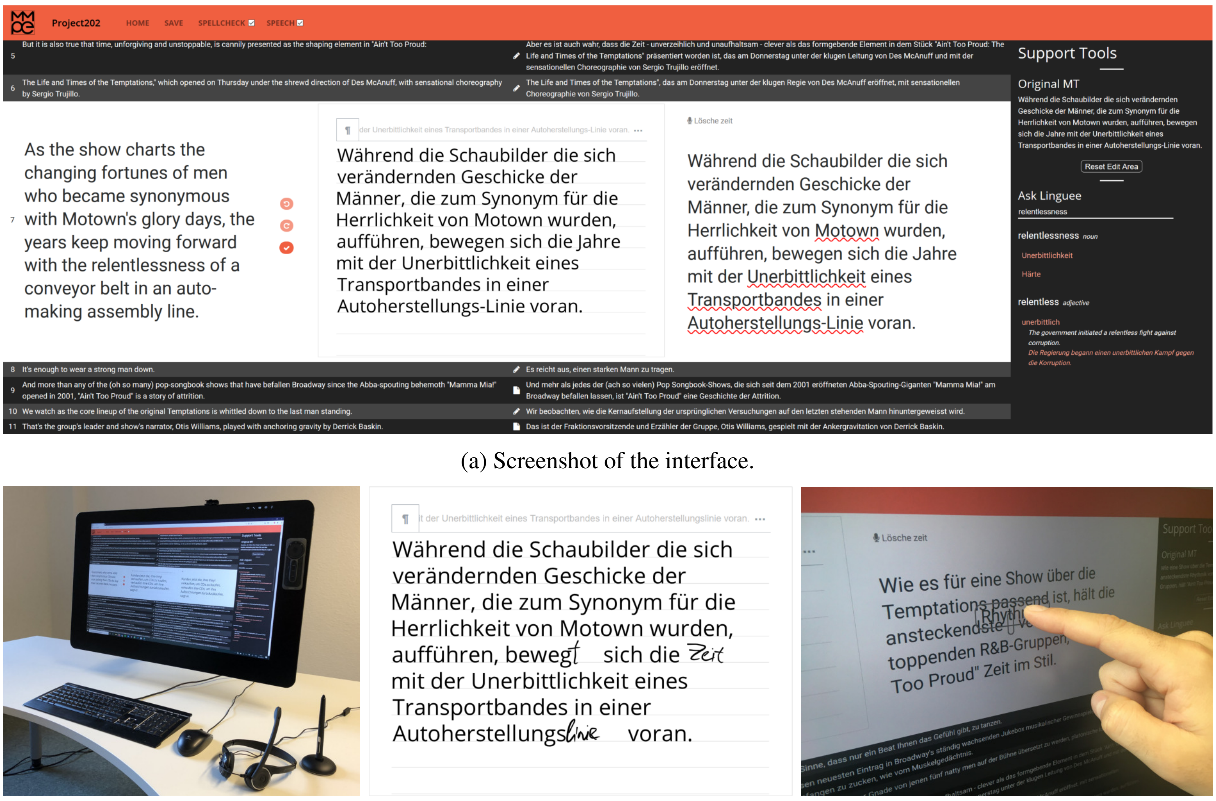
On the right-hand editing view, the user can delete words by simply double-tapping them with pen/finger touch, or reorder them through a simple drag and drop procedure (see Figure 1d), which visualizes the picked-up word as well as the current drop position, and automatically fixes spaces between words and punctuation marks. This reordering functionality is strongly related to Teixeira et al. (2019); however, only the currently dragged word is temporarily visualized as a tile to offer better readability. Naturally, the user can also edit using mouse and keyboard, where all common navigation inputs work as expected from other software.

3.4 Speech Input

For speech recognition, we stream the audio recorded by the headset to IBM Watson 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.

As commands, the user has the option to "insert", "delete", "replace", and "reorder" words or subphrases. To specify the position, if it is ambiguous, one can define anchors as in "after"/"before"/"between", or define the occurrence

[2] see https://developer.myscript.com/docs/concepts/editing-gestures/, accessed 16/04/2020



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.

Lösche zeit

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.

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.

Support Tools

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.

Reset Edit Area

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.

(a) Screenshot of the interface.

¶ 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.

Lösche zeit
Wie es für eine Show über die Temptations passend ist, hält die ansteckende Rhythm ... toppenden R&B-Gruppen ... Too Proud" Zeit im Stil.

(b) Apparatus.

(c) Handwriting on left target view.

(d) Touch reordering on right target view.

Figure 1: Overview of the MMPE prototype.

of the entity ("first"/"second"/"last"). A full example is "insert A after second B", where A and B can be words or subphrases. Character-level commands are not supported, so instead of e.g. deleting a suffix, one should replace the word.

3.5 Multi-Modal Combinations

Last, the user can use a multi-modal combination, i.e., pen/touch/mouse combined with speech. For this, the cursor first needs to be positioned on or next to a word, or the word needs to be long-pressed with pen/touch, resulting in a pickup visualization. Afterwards, the user can then use a simplified 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.

3.6 Logging

In a log file, we store all concrete keypresses, touched pixel coordinates, etc. Much more importantly, we directly log all UI interactions (like *segmentChange*), as well as all text manipulations (like *replaceWord*) together with the concrete changes (e.g. with the *oldWord*, *newWord*, and complete *segmentText*).

4 Evaluation Method

The prototype was evaluated by professional translators[3]. 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, see Section 3.5), while "operations" refers to Insertions, Deletions, Replacements, and Reorderings. The experiment consisted of the following phases and took approximately 2 hours per participant:

4.1 Introduction & Independent PE

First, 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.

After that, participants were given 10–15 minutes to explore the prototype on their own. We

[3] 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/

specifically told them that we are more interested in them exploring the presented features than in receiving high-quality translations. This phase had two main purposes: (1) to let the participants become familiar with the interface (e.g., how best to hold the pen) and to resolve questions early on; (2) to see how participants intuitively work with the prototype. Two experimenters carefully observed the participants and took notes on interesting behavior and questions asked.

4.2 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 an MT system, 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 using each modality, 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. For this, 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 easily show the required change to perform. The reason why we provided the correction to apply was to ensure a consistent editing behavior across all participants, thereby making subjective ratings and 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 Herbig et al. (2019a), participants rated each modality for that operation on three 7-point Likert scales ranging from "strongly disagree" to "strongly agree", namely as to 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.

In the end, after completing all 80 segments, we performed a final unstructured interview to capture high-level feedback on the interface as well as things we missed in our implementation.

4.3 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 like a TM or consistency checker are currently missing. Since our main question was whether the newly implemented features have potential for PE of MT or not, we focus on qualitative feedback, ratings, and timing information, which is more relevant to this research question.

5 Evaluation Results and Discussion

In this section, we present and discuss the study's main findings.

5.1 Participants

Overall, 11 (f=10, m=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 (avg=41.6, σ=9.3)[4], with 3 to 30 years of professional experience (avg=13.3, σ=7.4) and a total of 27 language pairs (avg=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. For most participants, the self-rated CAT knowledge was good (6 times) or very good (4 times, 1 neutral). 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 (avg=11.5, σ=5.1), where participants had used between 1 and 10 distinct CAT tools (avg=4.9, σ=2.7).

5.2 Subjective Ratings

Figure 2 shows the subjective ratings provided for each modality and operation on the three scales

[4] The small number of participants and their age distribution (with 10 participants of age 30 to 48, and only one aged 64) did not us allow to analyze the effect of age on the results.

"Goodness", "Ease of use", and "Good alternative to mouse & keyboard" after having tested each feature (see Section 4.2). 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 P and T were ranked even slightly higher than MK for reorderings. Speech and multi-modal were considered worse here.

5.3 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 1$^{\text{st}}$ and 0.5 times the 2$^{\text{nd}}$ position to both MM and S, while P got 3$^{\text{rd}}$, MK 4$^{\text{th}}$, and T the 5$^{\text{th}}$ position. To get an overall ordering across participants, we then multiplied the total amount of times a modality was rated 1$^{\text{st}}$/2$^{\text{nd}}$/3$^{\text{rd}}$/4$^{\text{th}}$/5$^{\text{th}}$ by 1/2/3/4/5 (similar to Zenner and Krüger (2017)). Consequently, a lower score indicates that this modality is better suited for the operation. The scores for each modality and operation are:

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

5.4 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 3 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.

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; MM was slowest by far. 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 $r > 0.83$).

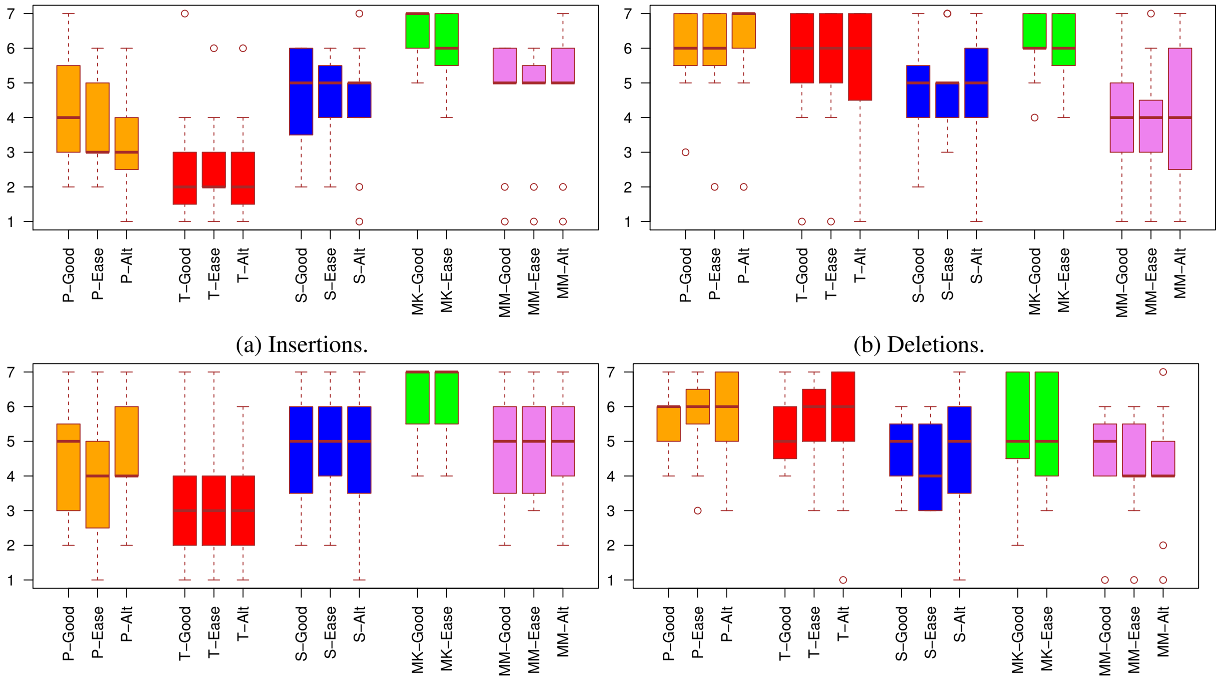
Replacements with MK were the fastest, followed by P, T, S, and MM. MK was significantly faster than all other modalities, and P significantly faster than S and MM (all p < 0.05, all $r > 0.83$), while no significant differences exist between the other three.

5.5 Qualitative Analysis

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

5.5.1 Pen & Touch

Especially for short insertions and replacements, handwriting was seen as a suitable input mode; for more extended changes, one should instead fall back on typing or dictation. Both touch/pen deletion mechanisms (strike-through and double-tap) and touch/pen reordering were highlighted as very useful or even "perfect" as they "nicely resemble a standard correction task". Most participants seemed to prefer the pen to finger handwriting for insertions and replacements due to its precision, although it was considered less direct.



(a) Insertions.

(b) Deletions.

(c) Replacements.

(d) Reorderings.

Figure 2: Subjective ratings.

A major concern was thinking about and creating sufficient space to handwrite into. A suggested improvement was to make the available space configurable to one's own handwriting. Furthermore, placing the palm of the hand on the screen should not be interpreted as input. Six participants also noted that the text jumps around when reordering a word from the end of a line, as the picked-up word is removed from the text, resulting in all remaining words being moved to the front, which could be prevented by adapting the text only on drop.

5.5.2 Speech & Multi-Modal Combinations

Perceptions regarding speech recognition were somewhat mixed, with some thinking it worked "super" while two participants found it exhausting to formulate commands while mentally working with text. Furthermore, speech was considered impractical for translators working in shared offices. 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". Speech deletion was considered to "work fine" and to be simpler than insertion as there is usually no need to specify the position. However, it would be unsatisfactory to have to read 10 words to delete them.

The main advantage of the multi-modal approach was that "one has to speak/think less". However, it was also argued 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 that "if there are no ambiguities, speech is better, but if there are, multi-modal is cool".

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 a few times a word was wrongly recognized (e.g., when replacing an 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 (Dymetman et al., 1994) or training the ASR towards the user.

5.5.3 Mouse & Keyboard

Due to 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: "why do I have to do all of this, why is it not as simple as the pen".

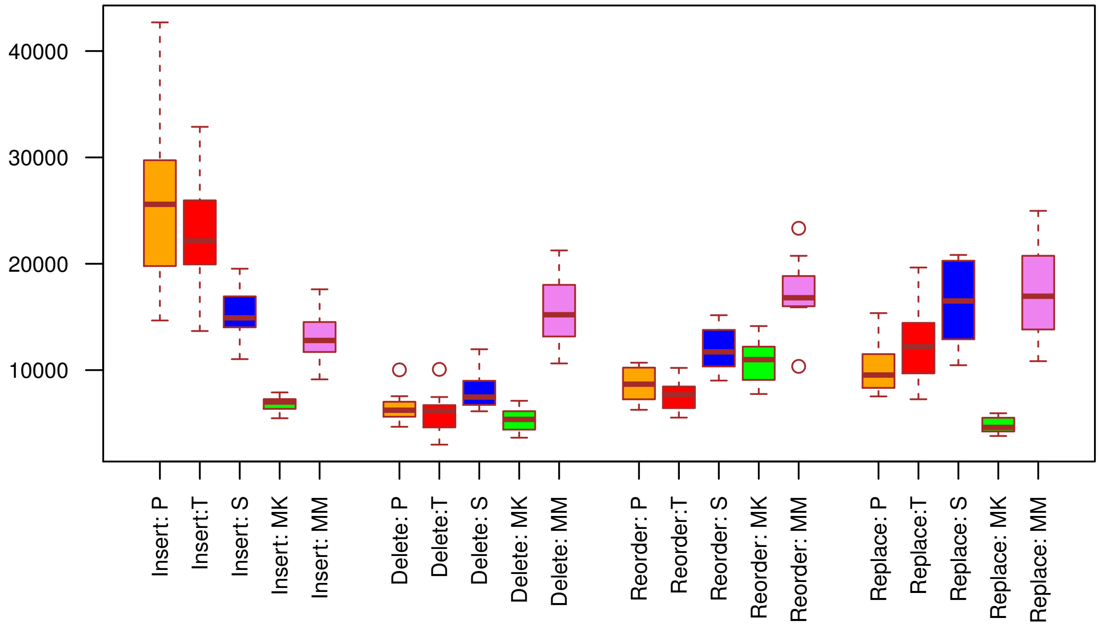


Figure 3: Editing durations (in ms) per operation and modality.

5.5.4 General Feedback

In general, we received lots of positive feedback in the final discussion about the prototype, where participants made statements such as “I am going to buy this once you are ready” or expressed “respect for the prototype”. Multiple participants reported that it would be nice to have multiple options to vary between the modalities. It was frequently suggested to combine the two editing views, e.g. by having a switch to enable/disable the drawing mode. Participants also commented positively on the large typeface for the current segment (“you really see what you are working on”). Suggestions for further improvements included adaptation possibilities for the size of the editing fields and a switch between vertical and horizontal source-target layout.

5.6 Discussion

This section discusses the main takeaways regarding each modality.

5.6.1 Pen

According to ordering scores, subjective ratings, and comments, we see that the pen is among the best modalities for deletions and reordering. However, other modalities are superior for insertions and replacements, where it was 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, even though it was rated lower. The main concern for handwriting was the need to think about space and to create space before actually writing.

5.6.2 Touch

Results for touch were similar, 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, and even for replacements, where it was clearly rated as the worst modality, it actually turned out to be (non-significantly) faster than S and MM.

5.6.3 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 2nd (after MK) and in particular much better than P and T. Timing analysis agrees for insertions, being 2nd after MK. For replacements, however, S and MM were the slowest even though the ratings put them before P and T. An explanation of why MM was slower than S for deletion is that our implementation did not support MM deletions of multiple words in a single command. Still, we would have expected a comparable speed of MM and S for reordering. Insertions are the only operation where the multi-modal approach 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 received a lot of positive feedback for insertions and replacements, where they would be more interesting the more text was to be added. The main advantage of the MM approach, as argued by the participants, was that one has to speak less, albeit at the cost of doing two things at once.

5.6.4 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, stating that it only works well because of "years of expertise".

5.6.5 General

Interestingly, our findings are not entirely in line with translators' intuitions reported in our previous elicitation study (Herbig et al., 2019a): while touch worked much better than expected, handwriting of whole subphrases did not work as well as 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.

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. Furthermore, several promising ideas for improving and extending the prototype have been proposed.

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.

6 Conclusion

While more and more professional translators are switching to the use of PE to increase productivity and reduce errors, current CAT interfaces still heavily focus on traditional mouse and keyboard input, even though the literature suggests that other modalities could support PE operations well. This paper therefore presents MMPE, a CAT prototype combining pen, touch, speech, and multi-modal interaction together with common mouse and keyboard input possibilities, and explores the use of these modalities by professional translators. The study 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 seen as suitable interaction modes; however, mouse & keyboard were still favored and faster here.

As a next step, we will integrate the participants' valuable feedback to improve the prototype. While the presented study provided interesting first insights regarding participants' use of and preferences for the implemented modalities, it did not allow us to see how they would use the modalities over a longer time period in day-to-day work, which we also want to investigate in the future.

Furthermore, participants in Herbig et al. (2019a) were positive regarding the idea of a user interface that adapts to measured cognitive load, especially if it automatically provides additional resources like TM matches or MT proposals. An exploration of multi-modal measuring approaches (Herbig et al., 2019b) shows the feasibility of this, so we will try to combine explicit multi-modal input, as done in this work, with implicit multi-modal sensor input to better model and support the user during PE.

Acknowledgments

This research was funded in part by the German Research Foundation (DFG) under grant number GE 2819/2-1 (project MMPE). We thank AMPLEXOR (`https://www.amplexor.com`) for their excellent support in providing access to professional human translators for our experiments.

References

Vicent Alabau, Ragnar Bonk, Christian Buck, Michael Carl, Francisco Casacuberta, Mercedes García-Martínez, Jesús González, Philipp Koehn, Luis Leiva, Bartolomé Mesa-Lao, et al. 2013. CASMACAT: An open source workbench for advanced computer aided translation. *The Prague Bulletin of Mathematical Linguistics*, 100:101–112.

Vicent Alabau and Francisco Casacuberta. 2012. Study of electronic pen commands for interactive-predictive machine translation. In *Proceedings of the International Workshop on Expertise in Translation and Post-Editing – Research and Application*, pages 17–18.

Nora Aranberri, Gorka Labaka, A Diaz de Ilarraza, and Kepa Sarasola. 2014. Comparison of post-editing productivity between professional translators and lay users. In *Proceeding of AMTA Third Workshop on Post-editing Technology and Practice*, pages 20–33.

Jan van den Bergh, Eva Geurts, Donald Degraen, Mieke Haesen, Iulianna van der Lek-Ciudin, Karin Coninx, et al. 2015. Recommendations for translation environments to improve translators' workflows. In *Proceedings of the 37th Conference Translating and the Computer*, pages 106–119. Tradulex.

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 *Proceedings of Eurospeech Fourth European Conference on Speech Communication and Technology*, pages 193–196.

Michael Carl, Martin Jensen, and Kay Kristian. 2010. Long distance revisions in drafting and post-editing. *CICLing Special Issue on Natural Language Processing and its Applications*, pages 193–204.

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 *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1–13. ACM.

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 Foster, Pierre Isabelle, Yves Normandin, and Pierre Plamondon. 1994. Towards an automatic dictation system for translators: The TransTalk project. In *Proceedings of the ICSLP International Conference on Spoken Language Processing*.

Marcello Federico, Nicola Bertoldi, Mauro Cettolo, Matteo Negri, Marco Turchi, Marco Trombetti, Alessandro Cattelan, Antonio Farina, Domenico Lupinetti, Andrea Martines, et al. 2014. The MateCat tool. In *Proceedings of the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 129–132.

Federico Gaspari, Antonio Toral, Sudip Kumar Naskar, Declan Groves, and Andy Way. 2014. Perception vs reality: Measuring machine translation post-editing productivity. In *Third Workshop on Post-Editing Technology and Practice*, page 60.

Spence Green, Jason Chuang, Jeffrey Heer, and Christopher D Manning. 2014a. Predictive translation memory: A mixed-initiative system for human language translation. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, pages 177–187. ACM.

Spence Green, Jeffrey Heer, and Christopher D Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 439–448. ACM.

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 *Proceedings of the EMNLP Conference on Empirical Methods in Natural Language Processing*, pages 1225–1236.

Nico Herbig, Santanu Pal, Tim Düwel, Kalliopi Meladaki, Mahsa Monshizadeh, Vladislav Hnatovskiy, Antonio Krüger, and Josef van Genabith. 2020. MMPE: A multi-modal interface using handwriting, touch reordering, and speech commands for post-editing machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics.

Nico Herbig, Santanu Pal, Josef van Genabith, and Antonio Krüger. 2019a. Multi-modal approaches for post-editing machine translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 231. ACM.

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

Maarit Koponen. 2012. Comparing human perceptions of post-editing effort with post-editing operations. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 181–190. Association for Computational Linguistics.

Elina Lagoudaki. 2009. Translation editing environments. In *MT Summit XII: Workshop on Beyond Translation Memories*.

Samuel Läubli, Mark Fishel, Gary Massey, Maureen Ehrensberger-Dow, and Martin Volk. 2013. Assessing post-editing efficiency in a realistic translation environment. In *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice*, pages 83–91.

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 *The 17th Annual Conference of the European Association for Machine Translation*, pages 81–88. European Association for Machine Translation.

Bartolomé Mesa-Lao. 2014. Speech-enabled computer-aided translation: A satisfaction survey with post-editor trainees. In *Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation*, pages 99–103.

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. 2015. Post-editing evaluations: Trade-offs between novice and professional participants. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 75–81.

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.

Sharon O'Brien, Joss Moorkens, and Joris Vreeke. 2014. Kanjingo – a mobile app for post-editing. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*.

Lane Schwartz, Isabel Lacruz, and Tatyana Bystrova. 2015. Effects of word alignment visualization on post-editing quality & speed. *Proceedings of MT Summit XV*, 1:186–199.

Carlos S.C. 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.

Dimitri Theologitis. 1998. Language tools at the EC translation service: The theory and the practice. In *Proceedings of the 20th Conference Translating and the Computer*, pages 12–13.

Antonio Toral, Martijn Wieling, and Andy Way. 2018. Post-editing effort of a novel with statistical and neural machine translation. *Frontiers in Digital Humanities*, 5:9.

Olga Torres-Hostench, Joss Moorkens, Sharon O'Brien, Joris Vreeke, et al. 2017. Testing interaction with a mobile MT post-editing app. *Translation & Interpreting*, 9(2):138.

Vincent Vandeghinste, Tom Vanallemeersch, Liesbeth Augustinus, Bram Bulté, Frank Van Eynde, Joris Pelemans, Lyan Verwimp, Patrick Wambacq, Geert Heyman, Marie-Francine Moens, et al. 2019. Improving the translation environment for professional translators. *Informatics*, 6(2):24.

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, et al. 2016. Scate – Smart Computer-Aided Translation Environment. *Baltic Journal of Modern Computing*, 4(2):382–382.

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 *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 8–15.

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.

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.

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

Julián Zapata. 2016. Translating on the go? Investigating the potential of multimodal mobile devices for interactive translation dictation. *Tradumàtica: Traducció i Tecnologies de la Informació i la Comunicació*, 1(14):66–74.

Julián Zapata, Sheila Castilho, and Joss Moorkens. 2017. Translation dictation vs. post-editing with cloud-based voice recognition: A pilot experiment. *Proceedings of MT Summit XVI*, 2.

Anna Zaretskaya and Míriam Seghiri. 2018. *User Perspective on Translation Tools: Findings of a User Survey*. Ph.D. thesis, University of Malaga.

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. *IEEE Transactions on Visualization and Computer Graphics*, 23(4):1285–1294.