Multi-Modal Approaches for Post-Editing Machine Translation

Full Paper

Nico Herbig
German Research Center for Artificial Intelligence (DFKI),
Saarland Informatics Campus
nico.herbig@dfki.de

Santanu Pal
Josef van Genabith
Saarland University
{santanu.pal,josef.vangenabith}@uni-saarland.de

Antonio Krüger
German Research Center for Artificial Intelligence (DFKI),
Saarland Informatics Campus
krueger@dfki.de

ABSTRACT

Current advances in machine translation increase the need for translators to switch from traditional translation to post-editing (PE) of machine-translated text, a process that saves time and improves quality. This affects the design of translation interfaces, as the task changes from mainly generating text to correcting errors within otherwise helpful translation proposals. Our results of an elicitation study with professional translators indicate that a combination of pen, touch, and speech could well support common PE tasks, and received high subjective ratings by our participants. Therefore, we argue that future translation environment research should focus more strongly on these modalities in addition to mouse- and keyboard-based approaches. On the other hand, eye tracking and gesture modalities seem less important. An additional interview regarding interface design revealed that most translators would also see value in automatically receiving additional resources when a high cognitive load is detected during PE.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; **Interaction techniques**.

KEYWORDS

Computer-aided translation, multi-modality, post-editing

ACM Reference Format:
Nico Herbig, Santanu Pal, Josef van Genabith, and Antonio Krüger. 2019. Multi-Modal Approaches for Post-Editing Machine Translation: Full Paper. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3290605.3300461

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

1 INTRODUCTION

As machine translation (MT) has been making substantial improvements in recent years[1], more and more professional translators are integrating the technology into their translation workflows [45, 46]. The process of using a pre-translated text as a basis and improving it to the final translation is called post-editing (PE). While older research showed a strong dislike of translators towards PE [17, 39], for which they are sometimes also paid less, the more recent study by Green et al. [12] demonstrated that translators strongly prefer PE and argue that "users might have dated perceptions of MT quality". The type of MT used also impacts PE time, where productivity gains of 36% have been shown with modern neural MT compared to an 18% increase for statistical MT [34].

Apart from MT, which is becoming more and more relevant due to increased performance, translation memories (TM) are already widely accepted tools to increase productivity [31]. 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.

While TMs continue to be useful tools for translating segments that are highly similar to matches in the database, the permanent availability of high-quality MT, or even multiple MT proposals from different engines, has the potential to change the translation process. This requires further investigation in terms of interface design, since the task changes from mostly text production to comparing and adapting MT and TM proposals, or put differently, from control to supervision. We hypothesize that translation environments need to be changed not only on a visualization and tool level, but that stronger changes in terms of using modalities other than mouse and keyboard could facilitate these new operations.

[1] WMT 2018 translation task: http://matrix.statmt.org/

After reviewing the relevant literature, we therefore present the results of an elicitation study conducted with professional translators. For this, 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, (d) how they envision an ideal translation environment setup for PE, and (e) what their thoughts are regarding adapting the translation user interface to measured cognitive load (CL), as was proposed for other interfaces in the HCI domain. 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, and that most participants would see a benefit in automatic adaptations to perceived CL.

2 RELATED WORK

In this section we present related translation environments, analyses of the PE process, existing multi-modal approaches, and last, the concept of elicitation studies.

Translation Environments and Post-Editing

Most professional translators nowadays use so-called CAT (computer-aided translation) tools [36]. These provide features like MT and TM together with quality estimation and concordance functionality [8], alignments between source and MT [30], color coding to show the similarity between input sentences and TM matches [25], interactive MT offering assistance like auto-completion [11, 13], or intelligibility to provide justifications for the suggestions provided by such CAT features [5]. 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 [21]. Adaptively combining MT and TM to get the best translation option was also investigated [47] .

Krings [16] found that PE decreased time 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. Furthermore, it has been shown that PE not only leads to reduced time, but also increases quality [12]. More recent studies also agree regarding the time savings achieved through PE: in [42], the authors find that PE was on average 28% faster for technical translations, [2] shows that PE increases translation throughput for both professionals and lay users, and [18] finds that PE also increases productivity in realistic environments.

Furthermore, PE changes the interaction pattern: while traditional translation consists of gisting, drafting, and revising phases [4], these phases are interleaved in PE [12], leading to a significantly reduced amount of mouse and keyboard events [12]. Therefore, we believe that mouse and keyboard might be less important for PE and that other modalities should be investigated.

Multi-Modal Approaches

The use of speech recognition for dictating translations dates back a long time [3, 7]. A more recent approach, called SEECAT [19], investigates the use of speech recognition in PE and argues that the combination with typing could boost productivity. A survey regarding speech usage with PE trainees [20] finds that these have a positive view on speech input and would consider adopting it; however, it should be used only as a complement.

Casmacat [1] is a CAT tool that offers extensive logging including eye tracking data and allows to input text by writing with e-pens on a special area. Mobile PE on an iPhone via touch and speech input has also been investigated [22, 35]. Participants especially liked reordering words with touch gestures, preferred voice when translating from scratch, but used the iPhone keyboard for small changes. Zapata [43] explores the use of voice and touch enabled devices such as touch screen computers or tablets for translation. However, the study did not focus on PE and participants simply used Microsoft Word instead of a fully-fledged CAT environment.

The presented literature suggests that professional translators should switch to the use of PE to increase productivity and quality, and that the PE process might be better supported by using different modalities (in addition to) the common mouse and keyboard approaches. A few approaches supporting different modalities have been proposed; however, to the best of our knowledge, no systematic evaluation of which PE tasks might be well supported by which modalities was conducted.

Elicitation Studies

To perform such a systematic evaluation, we conduct an elicitation study, which is a specific form of participatory design [29] and a common tool to design natural interfaces. Important aspects include leading participants away from technical thinking [26], making them assume that no recognition issues occur, and considering their behavior as always acceptable [41]. Furthermore, they should only be informed about the essential details of the task so as not to bias them towards existing approaches [40]. Instead, participants should be presented with so-called referents (i.e. common operations) and asked to propose actions to achieve these referents [10]. This approach has been shown to result in an increased immediate usage compared to highly iterated design approaches [40]. We use the formalization for elicitation studies by Vatavu and Wobbrock [37], 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 [9]. The introduced **coagreement rate** defines how much agreement two referents share, and a **significance test** for agreement rates is provided.

Similar to our work, Morris [23] performed a multi-modal (speech + gesture) elicitation study, where users were allowed to suggest more than one interaction per referent. For the analysis, she 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. [24] 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).

3 EVALUATION METHOD

Since professional translators are gradually moving towards PE [44, 45], which changes the interaction patterns with significantly fewer mouse and keyboard activities [12], we investigate which other modalities might be effective for PE based on an elicitation study [37].

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 [41], we gather conceptual complexity ratings on a 5-point scale for all our referents from the expert translators.

Unbiased Elicitation. After this, we conduct a classical elicitation study similar to [37]. First, the general idea of a multimodal 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 [41] 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, 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 [23], 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 [24]. 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.

Cognitive Load Adaptations. PE, and Translation in general, are tasks that can induce a high CL on the translator [16]. In the HCI community, a lot of approaches exist to measure the demand experienced by users [6, 28, 32, 38]. Therefore, the last part of our interview aims to understand (1) if users have interesting ideas on how such measurements could be used within the context of PE, and (2) which proposed user interface adaptations to CL they would find useful. For this, participants are asked to propose ideas themselves, and we discuss possibly interesting adaptations, which are rated on a simple 7-point scale ranging from "very bad" to "very good".

General. 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 they 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.

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. [27] proposes 5 PE operations: correcting word form, correcting word order, adding omission, deleting addition, and correcting lexical choice. Koponen [15] additionally distinguishes between moving single words or groups and the distance of the movement. Temnikova [33] 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 1 that we argue better captures the necessary operations from an interaction perspective.

Table 1: Referents (Ref) used for the elicitation study.

Ref	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*

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

Modalities

The modalities introduced at the beginning of the biased elicitation are: mouse and keyboard (MK), touch (T), a digital pen/stylus (P), mid-air gestures (G), speech (S), eye tracking (E), and combinations thereof (XY) (e.g. TS for touch and speech). We explain each of these modalities to the subjects based on examples drawn from daily life (e.g. touch well known from smartphones, gestures from science fiction movies, etc.) and explain how they can be used (e.g. a pen to draw or for handwriting).

4 EVALUATION RESULTS AND DISCUSSION

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

Participants and Conceptual Complexity

Overall 13 (f=9, m=4) professional translators participated in the experiment, 5 freelance and 8 in-house translators. Their ages ranged from 28 to 62 (avg=40.23, σ=9.11), with 2 to 34 years of professional experience (avg=13.65, σ=9.66) and a total of 39 language pairs (avg=3). For most participants the self-rated CAT knowledge was good (5 times) or very good (5). 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 (avg=9.12, σ=5.23), where participants had used between 1 and 9 distinct CAT tools (avg=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 (avg=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 [41] show that the participants found reordering multiple words the most complex (avg=4.08, σ=0.86), followed by reordering a single word (avg=3.23, σ=0.93), deletion of multiple extra items (avg=3.08, σ=0.76), and replacing an item (avg=2.92, σ=0.64). Adding missing items was rated as (avg=2.69, σ=1.18), correcting the word form as (avg=2.31, σ=1.11), and last, deleting a single extra item (avg=2.08, σ=0.86). 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.

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: [37] found an average agreement rate of .261 (min=.108 with N=12, 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 [37].

The agreement rate and the corresponding cumulative probability with its interpretation per referent is shown in Table 2. 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 [37]. 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 2 shows that all agreement rates are statistically significantly larger than 0.

Table 2: 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 (Ref). The highest AR is shaded; the lowest ARs are marked in bold. S=Speech, T=Touch, P=Pen, E=Eye, M=Mouse, K=Keyboard, Tp=Touchpad, XY=combination of X and Y.

Ref	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)	-

Co-agreement Rates. The co-agreement 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 [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) (cf. Table 2). 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 their 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 [23] (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 [24]. 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.

Biased Elicitation

The second part of the elicitation study was conducted similar to [23] considering the priming and production strategies of [24] to avoid legacy bias. In settings with multiple proposals per participant, agreement rates [37] are not meaningful; instead, we use the metrics max-consensus and consensus-distinct ratio [23].

Participants proposed on average 2 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 4 times while the percentage of touch and speech proposals was even higher. Table 3 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.

The most proposals were given for deleting single items, the least for additions (total) and reordering groups (distinct). Compared to Morris et al. [23] (3.1%) and our unbiased study (18.7%), we see far more multi-modal proposals (avg=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 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

Table 3: Proposals per referent (Ref) in the biased elicitation study. We report the (total and distinct) number of proposals, the percentage of multi-modal proposals (MM%), modalities suggested ≥ 3 times, and the max-consensus (C_m) and consensus-distinct (C_d, threshold = 2) ratios for all and the most frequent modalities (S=Speech, T=Touch, P=Pen, E=Eye, XY=combination of X and Y, e.g. TS for touch and speech). The highest scores are shaded in cyan, the lowest in bold text.

Ref	Number (tot/dist)	MM% tot/dist	Common Proposals	ALL		P		T		S		PS	
				C_m	C_d	C_m	C_d	C_m	C_d	C_m	C_d	C_m	C_d
A	**21**/12	38.1/63.6	P(6), S(4), PS(3)	**46.2**	**0.27**	83.3	0.50	-	-	75.0	**0.50**	100	1.00
RO_s	29/7	**3.5/14.3**	P(10), T(8), E(4), S(3)	76.9	0.71	100	**0.33**	100	1.00	66.7	**0.50**	-	-
RO_g	27/**6**	3.7/16.7	P(11), T(8), S(4)	84.6	0.66	90.9	0.50	87.5	**0.50**	**50.0**	1.00	-	-
RP_s	25/10	48.0/60.0	S(8), PS(5), P(3)	61.5	0.50	66.7	0.50	-	-	100	1.00	80.0	**0.50**
RP_p	22/8	31.8/62.5	P(7), S(5), T(3)	53.9	0.62	100	1.00	66.7	**0.50**	80.0	**0.50**	-	-
D_s	33/14	48.5/71.4	P(7), T(5), S(4), PS(3)	53.9	0.57	85.7	0.50	**60.0**	1.00	100	1.00	100	1.00
D_g	28/14	57.1/71.4	P(6), T(3), PS(3)	**46.2**	0.57	100	1.00	66.7	**0.50**	-	-	**67.7**	**0.50**

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 modality, was proposed. Furthermore, eye tracking was suggested to mark the position the user last looked at within the source and target text, to avoid getting lost when scrolling, similar to the Gazemarks approach [14].

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.

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 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 [5]). 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.

Cognitive Load Adaptations

Here, we present the results of our interview regarding possible CL adaptations.

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 (goodness 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 wasted time. However, it was also noted that it needs to be configurable, and risks showing irrelevant information.

Simplify. We also discussed the opposite idea: to hide elements when a high CL is detected, the hypothesis being that the interfaces are too complex [45] 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: avg=3.39, min=1, max=7, σ=1.94.

Estimate CL. 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 while 3 showed mixed and 1 negative opinions towards this idea. On average the ratings were rather good: avg=4.92, min=3, max=7, σ=1.19.

Breaks. The idea to propose breaks automatically was also discussed: one proposed to display a coffee cup icon, while another argued that this 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 stating that it might help if it is configurable and only suggested but not enforced.

Reordering. 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 often decide to 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 when discussing possible adaptations 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".

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.

5 CONCLUSION

Due to the changed task compared to classical translation, we argue for the need to investigate the PE process more strongly from an interaction, and specifically a modality perspective. For this, we propose a set of common operations necessary for PE and find that translators believe reordering tasks to be the most complex. Based on the agreement shared between replacing single items and parts of items, we 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.

In an initial, completely unbiased elicitation study, participants mostly proposed 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.

Furthermore, most participants 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.

In the future, we will develop and extensively test a PE interface supporting the proposed interactions with different modalities and investigate the feasibility of such CL adaptations in practice. Furthermore, we will investigate if our findings also apply to very related contexts like text review and proofreading, as we expect.

ACKNOWLEDGMENTS

This research was funded in part by the German research foundation (DFG) under grant number GE 2819/2-1 (project MMPE).

REFERENCES

[1] 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 (2013), 101–112.

[2] Nora Aranberri, Gorka Labaka, A Diaz de Ilarraza, and Kepa Sarasola. 2014. Comparison of Post-Editing Productivity between Professional Translators and Lay Users. In *Third Workshop on Post-editing Technology and Practice*. 20–33.

[3] 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 *Fourth European Conference on Speech Communication and Technology*. 193–196.

[4] Michael Carl, Martin Jensen, and Kay Kristian. 2010. Long Distance Revisions in Drafting and Post-editing. *Special Issue: Natural Language Processing and its Applications* (2010), 193–204.

[5] 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 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 1–13.

[6] Vera Demberg and Asad Sayeed. 2016. The Frequency of Rapid Pupil Dilations as a Measure of Linguistic Processing Difficulty. *PloS one* 11,

1 (2016), 1–29.

[7] Marc Dymetman, Julie Brousseau, George Foster, Pierre Isabelle, Yves Normandin, and Pierre Plamondon. 1994. Towards an Automatic Dictation System for Translators: The TransTalk Project. *ArXiv* (1994).

[8] 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 COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*. 129–132.

[9] Leah Findlater, Ben Lee, and Jacob Wobbrock. 2012. Beyond QWERTY: Augmenting Touch Screen Keyboards with Multi-touch Gestures for Non-alphanumeric Input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2679–2682.

[10] Michael D Good, John A Whiteside, Dennis R Wixon, and Sandra J Jones. 1984. Building a User-derived Interface. *Commun. ACM* 27, 10 (1984), 1032–1043.

[11] Spence Green, Jason Chuang, Jeffrey Heer, and Christopher D Manning. 2014. 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*. ACM, 177–187.

[12] 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*. ACM, 439–448.

[13] Spence Green, Sida I Wang, Jason Chuang, Jeffrey Heer, Sebastian Schuster, and Christopher D Manning. 2014. Human Effort and Machine Learnability in Computer Aided Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. 1225–1236.

[14] Dagmar Kern, Paul Marshall, and Albrecht Schmidt. 2010. Gazemarks: Gaze-based Visual Placeholders to Ease Attention Switching. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2093–2102.

[15] Maarit Koponen. 2012. Comparing Human Perceptions of Post-editing Effort with Post-editing Operations. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 181–190.

[16] Hans P Krings. 2001. *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*. Vol. 5. Kent State University Press.

[17] Elina Lagoudaki. 2009. Translation Editing Environments. In *MT Summit XII: Workshop on Beyond Translation Memories*.

[18] 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*. 83–91.

[19] 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*. European Association for Machine Translation, 81–88.

[20] 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*. 99–103.

[21] Joss Moorkens and Sharon O'Brien. 2017. Assessing User Interface Needs of Post-editors of Machine Translation. In *Human Issues in Translation Technology*. Routledge, 127–148.

[22] Joss Moorkens, Sharon O'Brien, and Joris Vreeke. 2014. Kanjingo–A Mobile App for Post-editing. *Tradumàtica: Traducció i Tecnologies de la Informació i la Comunicació* 14, 58–66.

[23] Meredith Ringel Morris. 2012. Web on the Wall: Insights from a Multimodal Interaction Elicitation Study. In *Proceedings of the 2012 ACM International Conference on Interactive Tabletops and Surfaces*. ACM, 95–104.

[24] Meredith Ringel Morris, Andreea Danielescu, Steven Drucker, Danyel Fisher, Bongshin Lee, Jacob O Wobbrock, et al. 2014. Reducing Legacy Bias in Gesture Elicitation Studies. *Interactions* 21, 3 (2014), 40–45.

[25] 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 *Proceedings of the Workshop Natural Language Processing for Translation Memories*. 36–42.

[26] 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*. Springer, 409–420.

[27] 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 *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*. 191–198.

[28] Dennis W Rowe, John Sibert, and Don Irwin. 1998. Heart Rate Variability: Indicator of User State as an Aid to Human-Computer Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 480–487.

[29] Douglas Schuler and Aki Namioka. 1993. *Participatory Design: Principles and Practices*. CRC Press.

[30] Lane Schwartz, Isabel Lacruz, and Tatyana Bystrova. 2015. Effects of Word Alignment Visualization on Post-editing Quality & Speed. *Proceedings of MT Summit XV* 1 (2015), 186–199.

[31] 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 (2016), 1–18.

[32] 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 (2000), 69–85.

[33] Irina Temnikova. 2010. Cognitive Evaluation Approach for a Controlled Language Post–Editing Experiment. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*. 3485–3490.

[34] 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 (2018), 9.

[35] 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 (2017), 138.

[36] 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*. Tradulex, 106–119.

[37] Radu-Daniel Vatavu and Jacob O Wobbrock. 2015. Formalizing Agreement Analysis for Elicitation Studies: New Measures, Significance Test, and Toolkit. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1325–1334.

[38] 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 (2012), 6075–6101.

[39] 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. Dissertation. University of Ottawa.

[40] Jacob O Wobbrock, Htet Htet Aung, Brandon Rothrock, and Brad A Myers. 2005. Maximizing the Guessability of Symbolic Input. In *CHI'05 Extended Abstracts on Human Factors in Computing Systems*. ACM, 1869–1872.

[41] Jacob O Wobbrock, Meredith Ringel Morris, and Andrew D Wilson. 2009. User-defined Gestures for Surface Computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1083–1092.

[42] 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*. 93–98.

[43] 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* 14 (2016), 66–74.

[44] Anna Zaretskaya, Gloria Corpas Pastor, and Miriam Seghiri. 2015. Translators' Requirements for Translation Technologies: Results of a User Survey. In *Proceedings of the Conference New Horizons is Translation and Interpreting Studies*.

[45] Anna Zaretskaya and Míriam Seghiri. 2018. *User Perspective on Translation Tools: Findings of a User Survey*. Ph.D. Dissertation. University of Malaga.

[46] 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 (2016), 91–100.

[47] Jost Zetzsche. 2016. Lilt: Translation Environment Tool of a Different Kind. *Multilingual magazine* (2016), 15–17.