Shaping Trust in Machine Translation Suggestions Through AI-Assisted Context Building

JORDAN S. HUFFAKER, Computer Science & Engineering, University of Michigan, USA
SAI R. GOURAVAJHALA, Lilt, USA

Recent advances in AI technology have led to widespread adoption of machine translation (MT) within computer-assisted translation tools. As a result, human translators (linguists) now frequently collaborate with AI to improve productivity and translation quality. However, MT often fails to provide quality suggestions for complex terms that have overlapping connotations. To evaluate the quality of MT suggestions for such terms, linguists engage in a process we call *context building* where they gather the connotations around terms to understand their utility within the translation context. In this work, we first conduct contextual inquiries to explore the different ways in which linguists engage in the process of context building. Based on the findings, we then propose and develop an AI-assisted context building approach that automatically shows relevant context to linguists for difficult terms. We evaluate this system with professional freelance translators and find that we can nudge linguists towards building context for terms where the MT is least reliable.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: language translation, human-AI collaboration, context building

ACM Reference Format:
Jordan S. Huffaker and Sai R. Gouravajhala. 2022. Shaping Trust in Machine Translation Suggestions Through AI-Assisted Context Building. 1, 1 (April 2022), 12 pages.

1 INTRODUCTION

In recent years, the use of artificial intelligence for language translation has skyrocketed within the localization industry [5]. The incorporation of state-of-the-art neural machine translation (NMT) [3] within popular computer-assisted translation (CAT) tools has made the technology more available than ever for linguists[1] to use. However, machine translation (MT)[2] still cannot produce the quality of translation that is often necessary for many types of content. Instead of relying on MT output alone, many linguists opt to work with the AI by correcting its errors after-the-fact [2] or by translating in tandem with the AI through an interactive workflow [20]. In this paper, we explore how these human-AI translator teams collaborate to provide high quality output translations.

Specifically, we examine the process linguists go through to find and correct errors made by the AI when translating difficult terminology (terms). This process—which we call *context building*—involves gathering the connotations of

[1]We use linguist as a more inclusive term for translator, as believe the term more appropriately captures the broad linguistic expertise that is necessary to fulfill the work of a professional translator.
[2]In this paper, in the context of language translation, we equate AI to mean MT and we use both acronyms interchangeably.

Authors' addresses: Jordan S. Huffaker, jhuffak@umich.edu, Computer Science & Engineering, University of Michigan, Ann Arbor, USA; Sai R. Gouravajhala, sai@lilt.com, Lilt, San Francisco, USA.

terms and AI-suggested translations for their applicability within the translation context. For example, an AI might suggest the French translation *bretelle* for the term *leg braces* when referring to a medical exoskeleton bracing system; however, after researching the term, a linguist would likely discover that the MT suggestion is more in reference to suspenders (the clothing) and is therefore not appropriate in the context.

We show that, when faced with difficult-to-translate words or phrases, linguists' existing attitude towards AI is one of distrust: they tend to spend a great deal of effort building context around all terms with a potential for error, even in situations where errors are likely permissible. In particular, we find that there tends to be a disconnect between the expectations linguists have for the quality of their translations, the reliability of the AI-suggested translation, and the actual quality that is expected of them by their clients. As a result, linguists may spend a great deal of time building context for terms in order to find the best quality of translation when they could have simply gone with a less-perfect AI suggestion that still meets their client's expectations.

We propose the idea of incorporating AI in the context building process in order to shape linguists' effort toward building context for terms where the AI is least reliable. Our approach involves automatically compiling the context surrounding MT suggestions into a *context panel* where linguists can more quickly make assessments about their adequacy within the document's translation context. We evaluate the effectiveness of our approach with a wizard-of-oz style study where we provide linguists with a variety of types of context such as common dictionary results and high-level guidance from other linguists. We then measure changes in their context building behaviors and translation outcomes. We demonstrate that with AI-assisted context building, it is possible to nudge linguists towards building context where it matters most.

In this paper, we make the following contributions:

- We explore how linguists go through the process of *context building* to gather connotations around terms in order to evaluate the utility of MT suggestions.
- We propose the idea of *AI-assisted context building* where AI is incorporated in the context building process by summarizing relevant contexts for AI suggestions so as to surface their utility for human collaborators.
- We implement the AI-assisted context building approach on top of an existing commercial CAT tool.
- We evaluate our approach through a wizard-of-oz style study of professional freelance translators and demonstrate its feasibility for shaping linguists' context building effort.

2 RELATED WORK

Our work comes during a significant shift in translation workflows as a result of an influx of CAT tools that incorporate AI. We build upon the work of prior researchers in AI and HCI in understanding linguist's relationship with AI and how to best support their collaboration. Prior to the use of MT, translation workflows were largely limited to the use of concordances, termbases[3], and translation memory[3]. Linguists have long made use of physical and online resources, as well as CAT software in a highly collaborative environment [15].

The introduction of MT has led to a large disturbance in translation workflows; namely, through the emergence of systems that use a *post-editing workflow* such as popular systems like CASMACAT [1], Matecat [7], MemoQ [10], and SDL Trados [11]. CAT tools that leverage a post-editing workflow pre-translate text with MT and then task linguists with correcting the translation after-the-fact [2], an approach that improves translation efficiency by reducing linguist's cognitive load when getting started with a translation [8]. While post-editing is more effective than unassisted

[3]Databases that track and suggest prior translations of terms and similar source text to aid linguists with maintaining consistency in their projects.

translation in contexts where the MT is accurate, further research has raised concerns about its efficacy in settings where the MT is prone to errors [6, 18].

To answer these challenges, recent CAT tools such as Lilt [9] and Transmart [13] make use of a less restrictive workflow—called the *interactive workflow*—where the MT provides real-time translation suggestions to the linguist in an auto-complete style. The interactive workflow improves upon post-editing by providing linguists with more agency in their translation, leading many to prefer it [17]. However, as Moorkens et al. report, MT suggestions are still treated with some hesitancy, including issues of MT trustworthiness and quality [19]. Coppers et al. developed the Intellingo system to explore how different extensions to an interactive interface help with intelligibility [4] and find that intelligibility features are preferred by linguists when information being presented isn't already in the linguist's readily available knowledge.

While there has been substantial research focused on linguist's attitudes towards AI, including trust and intelligibility, we address a growing desire to see research on linguist's relationship with AI and AI's role within broader sociotechnical translation systems [16, 21]. We provide insight this topic by uncovering how linguists assess the quality of AI suggestions and by designing an approach that assists linguists in their evaluation of AI. In the next section, we will discuss the first steps we take by conducting a contextual inquiry with professional freelance translators.

3 PRESTUDY: CONTEXTUAL INQUIRY WITH PROFESSIONAL FREELANCE TRANSLATORS

As a first step toward understanding how linguists collaborate with AI within their translation workflow, we conduct six semi-structured interviews with professional French freelance translators. We design our interviews in the form of a contextual inquiry [22] because it enables us to better examine the thought-process and work behaviors undertaken by the linguists within their typical work environment. Within the interviews, we provide our linguists with the Lilt[4] translation system and asked them to translate three text snippets (see Appendix A for the exact snippets) that range in both the domain, as well as translation difficulty.

Within all six of our interviews, we find participants spend a substantial portion of their time undergoing a process we call *context building*[5]—where linguists research the connotations around source language and potential translations—in order to evaluate the utility of the MT suggestions. In all three of the text snippets, we observe nearly all context building effort was centered around a few terms. Our participants described these terms as the most difficult parts of the text to translate and they used multiple resources to gather connotations around them including dictionaries, concordances, thesauruses, search engines, and even other MT tools.

Additionally, we observe four modes in which our participants underwent context building including:

- *Searching for "official" translations*: involves looking at authoritative sources for definitions or formal translations of a term. An example might include a linguist going to the official Quebec government dictionary for the translation of the term "waive."
- *Searching for the in-group lingo*: involves looking for a translation that is used within a particular group or community. An example could be a linguist trying to build context for the term "weeder," which has a different translation depending on the scrapbooking or landscaping communities.

[4]The Lilt translation system [9] provides linguists with MT suggestions in an interactive workflow fashion (see Figure 1). All interview participants had extensive experience working with the system prior to their interviews.

[5]We assign a name to a common process linguists undergo. Karamanis et al. describe some of the practices linguists at a language service provider use while completing this process prior to the widespread adoption of MT within translation workflows [15]. Now that MT usage is near industry standard, we examine how MT influences linguists use of context building.

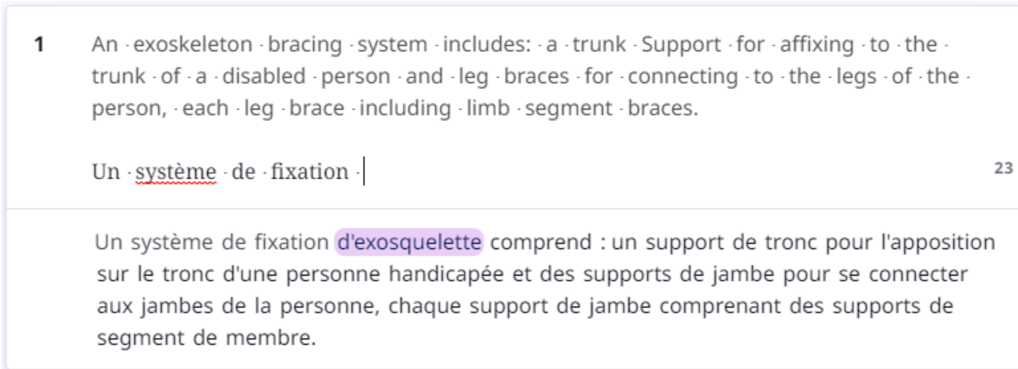

Fig. 1. Our CAT tool enables linguists to interact with machine translation suggestions, which automatically update based on the text typed so far. Linguists can autofill suggestions with a hotkey so that they don't have to move their hands from the keyboard.

- *Searching for credibility signals*: involves looking to see if the translation that the linguist has in mind is being used in similar contexts elsewhere. An example of this would be a linguist searching google images to see if the translation "bretelle" suggested by the MT for the term *leg braces* is appropriate in a medical context.
- *Brainstorming*: involves enumerating potential translations of a term to help find a more creative translation. An example may include going to the concordance *Linguee* to find numerous ways prior linguists have translated the phrase "turn on a dime."

Participants underwent these different modes of context building in a variety of ways—often using more than one mode at once. For example, we see this in the case of linguists searching the MT suggestion "padel" (searching for credibility signals) to understand it's applicability in the competitive padel playing community (searching for in-group lingo). With these findings, we designed a system that can help linguists complete these modes of context building. We will next explain the details of our system and an experiment we developed to evaluate its efficacy.

4 SYSTEM DESIGN

In this section, we discuss the system we designed to assist linguists in the context building process. Based on the results from our initial contextual inquiry interviews, we design an approach we call *AI-assisted context building* where AI is incorporated into the context building process by summarizing the relevant contexts for AI suggestions so that the linguist can understand their utility within the translation context. We implement our approach on top of the Lilt commercial CAT tool [9] that features state-of-the-art neural machine translation [23]. We add modifications to this tool that enable it to automatically summarize relevant contexts for the AI's suggestions within a *context panel*. We will first discuss the relevant features of the existing CAT tool, and then we will explain the modifications we have made.

Our CAT tool uses an interactive translation workflow (Figure 1). As a linguist translates, the system provides them with a next-word suggestion that the linguist can choose to accept (and auto-complete the word) or reject by entering their own translation. The system automatically adapts to the prefix of what a linguist has already entered, so suggestions can only build off of a linguist's existing translation.

We modify this existing CAT tool to incorporate a context panel that includes a summary of the context for several of the most difficult-to-translate terms within the text (Figure 2). For each term, we include a Google search-style preview

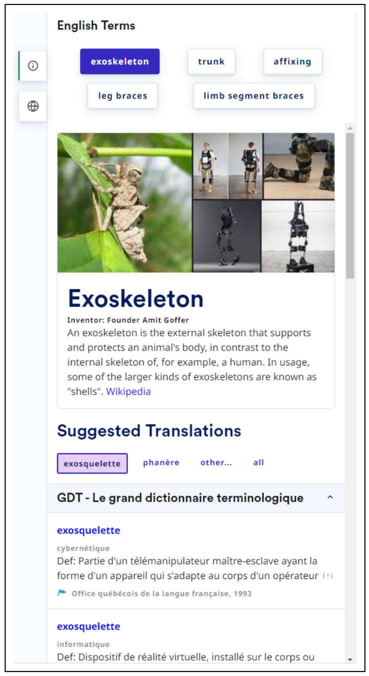

Fig. 2. Context panel: Based on our findings from the contextual inquiries, we develop an AI-assisted context building approach which automatically shows relevant information for terms as the linguist works on a sentence. The context panel comprises information found from a multitude of resources, such as dictionaries, Wikipedia entries, concordance results, and alternative MT suggestions.

of the English term and a list of potential translations compiled from several popular translation resources. For some of the most difficult terms, we included guidance for the linguist, which is a sentence describing what the client might be looking for in the translation. Based on the findings from our prestudy, we selected four popular resources for English-to-French translation whose results we compile in the context panel including the *Grand dictionnaire terminologique* (GDT), *Termium*, *WordReference*, and *Linguee*. Linguists can quickly filter through this context by selecting translations from the provided list, with the AI suggestion highlighted in purple. Now that we have described the components of our system, we will now transition to describing how we evaluated the effectiveness of our approach.

5 STUDY

In this section, we will describe the steps we take to evaluate how our AI-assisted context building approach impacts linguists' context building effort and their translation outcomes. We design our study as a wizard-of-Oz: the system works only for the specific text snippets we use in the study; however, we ground our design decisions and test examples in the findings from our initial contextual inquiry to ensure future implementation of the system is within the realm of feasibility. We will first discuss the nature of our study including our participants, then we will discuss our study's conditions, and we will conclude with our main measures.

5.1 Participants

We recruited professional English-to-French freelance translators from UpWork [12] to participate in an hour-long online study, compensating them $30 for their time. We selected participants that have earned at least $100 on the

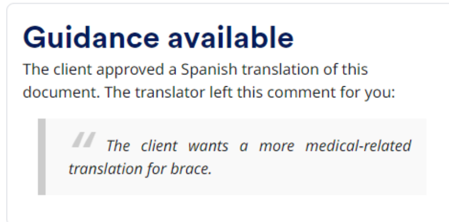

Fig. 3. Guidance: For the most difficult terms, we provide guidance from another experienced linguist to help linguists curate their context building efforts.

UpWork platform doing translation tasks and have a job success rate of at least 70%. In the study, we asked each participant to complete a short prestudy survey, a tutorial of our system, to translate four short text snippets (see Appendix A), and a short post-study survey. Between each translation task, we additionally asked participants to complete a short post-task survey.

To further ensure the expertise of our participants, we treat the first of our text snippets as a gold standard and we remove participants that fail to produce a quality translation on the task. Specifically, we modify the MT so that it would suggest a poor quality translation for the term *credit freeze* and then removed participants that failed to pick the correct translation ‘gel du crédit‘. For the following three snippets, we asked linguists to translate as fast and accurately as possible, and we counter-balanced the order in which translators were given each snippet.

5.2 Conditions

To understand the potential of our AI-assisted context building approach, we created two versions of our approach and we evaluate them against two baselines conditions. We design our experiment to tightly control the ways in which participants are able to engage in context building[6], and then we create conditions to vary the amount of context they are provided for MT suggestions. For each text snippet, we select the 4-5 terms participants spent the most time building context for in our prestudy to include in the context panel. Our conditions are as follows:

- Suggestions only (Baseline): in this condition we provided participants only with AI suggestions, no additional context, and no way to search for additional context.
- Browser only: in this condition we provided participants with a built-in browser that they can use to search for additional context. We tracked both their search terms and the websites they visited while using our browser.
- Browser and context panel: in this condition we provided participants with a built-in browser and our context panel. For each snippet, we compiled the results of four of the most commonly used resources from 4-5 of the most commonly searched for terms in the prestudy. We additionally include a Google search style preview of the English version of each term (see Figure 2).
- Browser and context panel with guidance: in this condition we again provided participants with a built-in browser and our context panel, but this time additionally included guidance from another experienced linguist (see Figure 3). We provided guidance aimed to help linguists curate their context building efforts toward types of

[6]To control the exact context participants are provided, we ask them not to use resources outside of our web page. We give them periodic reminders of this request, including at the moment they click out of our web page. Furthermore, we take the additional step of tracking cases where participants do leave the web page and monitor their behaviors when they return to ensure no external resources influenced their decision making.

translations the client is looking. For each snippet, we provided guidance for one term based on a quote from a participant in our prestudy.

We randomly assigned participants one unique condition per snippet and counter-balanced the order in which participants received conditions, resulting in a repeated measures study design.

5.3 Measures

We measured a number of factors related to time, linguist confidence, and translation quality. They are as follows:

- Task time: our first measure involves the amount of time from the point at which a participant clicked "start task" on a translation task from the time they clicked "submit".
- Browser time: we measure the amount of time participants spent in the built-in browser by adding up all of the times from the point at which a participant made a search, to the point at which they continued typing in their translation. We additionally hand-coded each search a participant made within the browser for the specific resource they were using, term, and translation they were looking for based on the URL they visited and the search terms they used.
- Context panel time: likewise, we measure the amount of time participants spent in the context panel by adding up all of the times from the point at which a participant interacted with the panel through a button click, to the point at which they continued typing in their translation.
- Term confidence: we ask participants to indicate how confident that their translation for each term is the "the best possible translation" for the term on a 5-point Likert scale.
- Term client-confidence: again, we ask participants to indicate for each term how confident they are that they "understand what type of translation" for the term the client is looking for on a 5-point Likert scale.
- Term quality: we hand-coded the translations for each of the terms based on how well they match criteria we determined from the results of the prestudy.

Now that we have described the design of our study, we next will share some of our main findings.

6 RESULTS

In this section, we will discuss the findings of our experiment on how AI-assisted context building impacts linguists' context building behaviors and translation outcomes. We find that our approach impacted linguists in numerous ways including by shaping their context building effort, their confidence in their translations, and the quality of their translations. We will begin with a high-level description of our participants, then we will walk through a quantitative analysis of our findings.

6.1 Participants

Based on the criteria we previously described, we selected a total of 98 professional freelance translators to participate in our study. The majority of these participants have a great deal of experience translating: about 60% with two or more years of experience and about 30% with more than 5 years of experience. Only 10% of our participants had less than six months of experience doing professional freelance translation. Additionally, about half of our participants have received formal education for translation; the other half were self-taught or learned through professional experience.

Our participants have experience doing a wide variety of types of translation. These include doing translation for the types of clients we examine in the study including e-commerce owners, lawyers, and medical content creators, as well as various other types of clients like search engine optimization agencies, book publishers, bloggers, and teachers. About half of our clients indicated that they specialize in a particular type of content; although, their indicated specializations were typically broad enough to include entire classes of content such as healthcare and marketing. Additionally, while all of our participants have experience translating English-to-French, most of our participants have experience translating additional language pairs like English-to-German and English-to-Spanish. Now that we have described the participants of our study, we will transition to an analysis of our participants' context building behaviors.

6.2 Context Building Behaviors

In accordance with the findings from our contextual inquiry, we find that context building behaviors made up a significant portion of the time participants spent translating. In the conditions where participants had access to context building tools, participants spent an average of 17% of their translation time in the browser or searching through our context panel. The distribution of our participants' context building time has a strong skew toward zero and slowly tapers off as the percentage increases.

Again, similar to the findings from our contextual inquiry, we find that participants spent the vast majority of their context building time researching terms. Following the coding process we have previously described, we hand-coded each browser search a participant made and were able to identify the specific term, translation, and resource a participant used to build context for 98% of searches. Of the 2% remaining searches, 1% of them were spent searching for resources and general information about the client, and we were unable to identify the purpose of the final 1%.

More specifically, participants spent most of their context building time building context for the terms we identified in the context panel. Per snippet, the terms 'padel', 'herringbone', and 'turn on a dime' were the most searched for in the marketing snippet, 'waiver' in the code of conduct snippet, and 'leg braces' and 'limb segment braces' for the patent snippet. When participants searched for terms we did not include in the context panel, they generally spread their time across many different terms. The combination of these times is comparable to the time they spent on any one of the context panel terms.

6.2.1 Conditions. In general, we find that making context for terms readily available to participants increased the amount of time they spent building context (Figure 4). We come to this conclusion based on the finding that participants spent the most amount of time doing context building in the condition in which we provide them with a built-in browser and our context panel. Compared to the condition in which participants were provided with only a built-in browser, we see that the increase in context building time is very slight for the code of conduct and marketing snippets, but is much starker for the patent snippet. However, we find that providing participants with guidance in addition to the context panel information can mitigate some of the increase in context building time to a level that is comparable to the browser-only condition for the code of conduct snippet, less for the marketing snippet, and still greater for the patent snippet. As expected, participants completed the least amount of context building in the baseline condition where we restricted their access to the web browser and limited the information available in the context panel to just include translation suggestions for each term.

Digging deeper into how participants allocated their context building time across conditions, we see that participants generally favored using the context panel to research terms instead of using the browser. When they did use the browser, their effort was more allocated toward searching terms and translations that were not include in the context panel. Providing participants with guidance reduced both the time they spend in the browser and in the context panel as

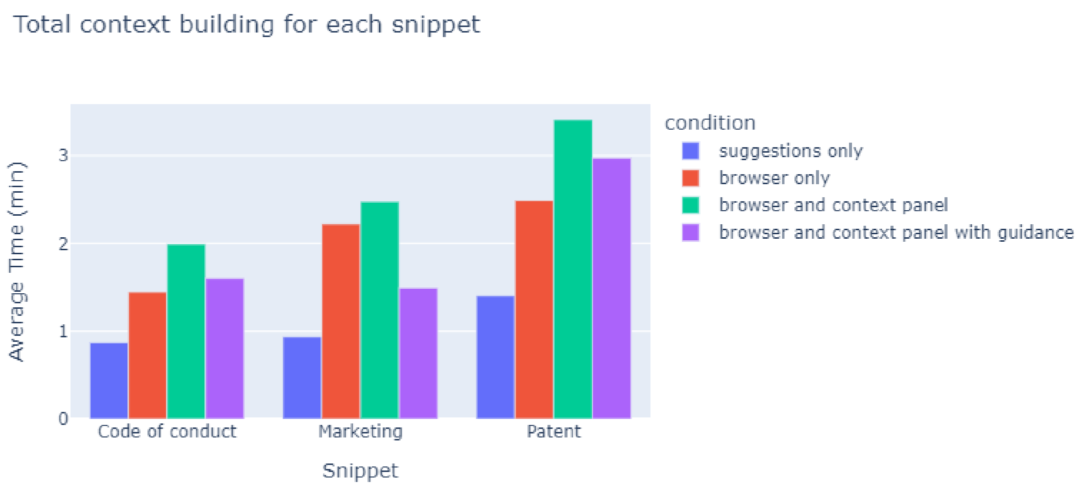

Fig. 4. Providing context for terms can lead to increases in context building time; however, providing guidance reduces context building time for places where reductions in translation quality are more acceptable.

compared to when they have no guidance. Now that we have described how our conditions impacted participants' context building behaviors, we will now explain their impact on translation outcomes.

6.3 Translation Outcomes

Context building impacts key translation outcomes including total translation time, participant confidence, and translation quality. In our initial analysis, we attempted to examine how our conditions predicted these outcomes; however, we have found that these findings are muddled due to a number of factors that cause high variance between our participants. Instead, we examine these outcomes by analyzing the impact of participants' context building behaviors on them.

Our first outcome, translation time, is strongly correlated with the time participants spend building context. As such, we see that translation time generally follows the same trends as indicated in Figure 4, however, the patterns are more muddled due to higher variance.

Likewise, participant confidence is again strongly correlated with the time participants spend building context. As we have previously mentioned, we measure participant confidence in two parts at the term level: the first by their confidence that their translation for a term is the best possible translation for the term, and the second by their confidence that they understand what type of translation the client is looking for. Both of these measures were tied to the amount of time a participant spent building context for a particular term.

Finally, we observe that participants' context building behaviors only impacted term translation quality for the patent snippet. Notably, the amount of time participants spent building context for specific terms in the code of conduct and marketing snippets were not predictive of the quality of translation they pick for the term. For the patent snippet, we observe that the more time they spent building context for the terms 'leg braces' and 'limb segment braces' was correlated with picking high-quality medical translations for these terms. Now that we have explained the quantitative evidence for the impact of our system on translators' context building behaviors and translation outcomes, we will recap how our findings tie back to our contributions and conclude.

7 DISCUSSION AND CONCLUSION

In this paper, we have examined how linguists engage in the process of context building to evaluate the utility of MT suggestions. We first conducted a series of contextual inquiries to better understand the motivations behind how and why linguists engage in context building. Based on the findings from these interviews, we designed an approach that provides assistance to linguists as they engage in context building by summarizing the relevant contexts of MT suggestions. We implemented this approach on top of an existing, commercial CAT tool, and performed a wizard-of-Oz study to evaluate its efficacy. We find that providing linguists with additional context for AI suggestions leads to an increase in linguists' overall context building effort. However, we find that providing linguists with guidance can nudge them to curate their context building effort so that they focus on terms where the MT is least reliable.

Our contributions are therefore: 1) an exploration of how linguists engage in process of context building when seeking to understand the utility of MT suggestions; 2) the idea of incorporating AI in the context building process by summarizing relevant contexts for AI suggestions so as to surface their utility for human collaborators; 3) a system that incorporates our idea of AI-assisted context building for language translation; and, 4) an evaluation of our approach with professional freelance translators where we find that it can be used to shape context building effort.

Our work represents a step toward systems that can shape human trust in AI through by influencing the process of context building. While we cannot say for certain whether our approach shapes human trust in AI suggestions directly, we do find evidence that it can shape the effort human collaborators exert to understand the context of those suggestions, which we find is linked to their acceptance of the suggestion and their confidence that it is correct. We argue that this finding can be used to improve the reliability and efficiency of human-AI teams by shaping context building effort toward places where reliability matters most.

While we have examined one domain in this paper, language translation, we believe that the approach we propose is applicable to any task that involves processing data with social references[7]. For example, our approach might also be applicable in cases where AI assists with creative writing —where it may generate new social references from mash-ups of old ones— or in cases where AI is used to assist with content moderation. We believe that our approach will be important for designing AI systems that can assist humans in delineating these connotations, leading to more trustworthy and reliable human-AI collaborations.

ACKNOWLEDGMENTS

This research was supported by Lilt [8]. We thank the linguists who provided their valuable insights during the initial contextual inquiry and the linguists who helped us evaluate our system by participating in the main study. We thank the members of the research team at Lilt for providing guidance throughout the design our study and the members of the infrastructure team for their assistance in deploying our system online.

[7]As we have coined in our prior work [14], a *social reference* refers to language that invokes connotations by overlapping parts of social and cultural contexts, in an often implicit manner.
[8]https://lilt.com

REFERENCES

[1] Vicent Alabau, Christian Buck, Michael Carl, Francisco Casacuberta, Mercedes García-Martínez, Ulrich Germann, Jesús González-Rubio, Robin L Hill, Philipp Koehn, Luis A Leiva, et al. 2014. Casmacat: A computer-assisted translation workbench. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*. 25–28.

[2] Jeffrey Allen. 2003. Post-editing. *Benjamins Translation Library* 35 (2003), 297–318.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[4] 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*. 1–13.

[5] Donald DePalma. 2021. The Language Sector in Eight Charts. *CSA Research* (2021).

[6] Félix Do Carmo. 2020. 'Time is money'and the value of translation. *Translation Spaces* 9, 1 (2020), 35–57.

[7] 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 *COLING (Demos)*. 129–132.

[8] 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*. 439–448.

[9] https://lilt.com. 2022. The modern way to localize | Lilt Home. *Lilt* (2022).

[10] https://memoq.com. 2022. memoQ | Translation and Localization Management Solutions. *MemoQ* (2022).

[11] https://trados.com. 2022. Trados | Translation Software, CAT Tool & Terminology. *SDL Trados* (2022).

[12] https://upwork.com. 2022. Upwork | The World's Work Marketplace. *UpWork* (2022).

[13] Guoping Huang, Lemao Liu, Xing Wang, Longyue Wang, Huayang Li, Zhaopeng Tu, Chengyan Huang, and Shuming Shi. 2021. Transmart: A practical interactive machine translation system. *arXiv preprint arXiv:2105.13072* (2021).

[14] Jordan S. Huffaker, Jonathan K. Kummerfeld, Walter S. Lasecki, and Mark S. Ackerman. 2020. *Crowdsourced Detection of Emotionally Manipulative Language*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376375

[15] Nikiforos Karamanis, Saturnino Luz, and Gavin Doherty. 2011. Translation practice in the workplace: contextual analysis and implications for machine translation. *Machine Translation* 25, 1 (2011), 35–52.

[16] Dorothy Kenny, Joss Moorkens, and Félix do Carmo. 2020. Fair MT: Towards ethical, sustainable machine translation. *Translation Spaces* 9 (08 2020), 1. https://doi.org/10.1075/ts.00018.int

[17] Rebecca Knowles, Marina Sanchez-Torron, and Philipp Koehn. 2019. A user study of neural interactive translation prediction. *Machine Translation* 33, 1 (2019), 135–154.

[18] Samuel Läubli, Chantal Amrhein, Patrick Düggelin, Beatriz Gonzalez, Alena Zwahlen, and Martin Volk. 2019. Post-editing productivity with neural machine translation: an empirical assessment of speed and quality in the banking and finance domain. *arXiv preprint arXiv:1906.01685* (2019).

[19] Joss Moorkens and Sharon O'Brien. 2017. Assessing user interface needs of post-editors of machine translation. *Human issues in translation technology* (2017), 109–130.

[20] Tanja Munz, Dirk Väth, Paul Kuznecov, Thang Vu, and Daniel Weiskopf. 2021. Visual-interactive neural machine translation. In *Graphics Interface 2021*.

[21] Sebastin Santy, Kalika Bali, Monojit Choudhury, Sandipan Dandapat, Tanuja Ganu, Anurag Shukla, Jahanvi Shah, and Vivek Seshadri. 2021. Language Translation as a Socio-Technical System: Case-Studies of Mixed-Initiative Interactions. In *ACM SIGCAS Conference on Computing and Sustainable Societies*. 156–172.

[22] Dennis Wixon, Karen Holtzblatt, and Stephen Knox. 1990. Contextual Design: An Emergent View of System Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, USA) *(CHI '90)*. Association for Computing Machinery, New York, NY, USA, 329–336. https://doi.org/10.1145/97243.97304

[23] Joern Wuebker, Patrick Simianer, and John DeNero. 2018. Compact personalized models for neural machine translation. *arXiv preprint arXiv:1811.01990* (2018).

A TEXT SNIPPETS

The following are the exact texts we include in the study. The translation interface splits these snippets into two "segments" where each segment consists of one sentence each.

A.1 Gold Standard

We received your Credit Freeze request and a freeze is now in place on your TransUnion credit report. It will stay in place until you request its removal. A credit freeze prevents lenders from checking your credit report in order to open a new account.

A.2 Code of Conduct

To seek a waiver, speak with a manager, who will consider the request in consultation with others, such as Internal Audit, Legal or Human Resources. Waivers of the Code of Conduct require the permission of Intel's Chief Financial Officer, General Counsel, or Chief People Officer.

A.3 Marketing

Competitive padel players seeking a shoe that lends support and flexibility will appreciate the technical design of the GEL-BELA™ 7 style. Its formation includes a lightweight and resilient rubber compound in the outsole with a padel-specific herringbone pattern and pivot points to help you turn on a dime.

A.4 Patent

An exoskeleton bracing system includes: a trunk Support for affixing to the trunk of a disabled person and leg braces for connecting to the legs of the person, each leg brace including limb segment braces. Motorized joints are adapted to provide relative angular movement between the limb segment braces of the leg braces and between the leg braces and the trunk Support.