Rapidly Piloting Real-time Linguistic Assistance for Simultaneous Interpreters with Untrained Bilingual Surrogates

Alvin Grissom II[1] **Jo Shoemaker**[2] **Ben Goldman**[1] **Ruikang Shi**[1]
Craig Stewart[3] **C. Anton Rytting**[2] **Leah Findlater**[4] **Jordan Boyd-Graber**[3]
[1]Haverford College [2]University of Maryland
[3]Carnegie Mellon University [4]University of Washington

Abstract

Simultaneous interpretation is a cognitively taxing task, and even seasoned professionals benefit from real-time assistance. However, both recruiting professional interpreters and evaluating new assistance techniques are difficult. We present a novel, realistic simultaneous interpretation task that mimics the cognitive load of interpretation with crowdworker surrogates. Our task tests different real-time assistance methods in a Wizard-of-Oz experiment with a large pool of proxy users and compares against professional interpreters. Both professional and proxy participants respond similarly to changes in interpreting conditions, including improvement with two assistance interventions—translation of specific terms and of numbers—compared to a no-assistance control.

Keywords: simultaneous interpretation, human-computer interaction, translation assistance, evaluation

Simultaneous interpretation (SI) is the translation of utterances from one language into another in *real time*, in contrast to consecutive translation, where the translator waits until the end of each sentence. Real-time assistance presents interpreters with information to facilitate better interpretations. We make three contributions: (1) a survey of professional interpreters to gauge their preferred kinds of assistance, (2) evaluation of assistance methods with expert translators, and (3) a fast evaluation scheme for real-time SI assistance that only requires bilingual participants.

SI combines the challenges of translation—where the literal meaning, subtext, tone, and register of an utterance must be rendered into another language—with the time pressure of keeping up with the interpretee. The cognitive load of this is immense. Even with advance preparation, professionals struggle with technical terms (Pignataro, 2012), dates and numbers (Mazza, 2001), proper nouns, and colloquialisms. To alleviate this, interpreters work in pairs, with one person actively interpreting while the other takes notes, looks up references, or suggesting translations (Gile, 2009).

Just as computer assistance developed by the NLP community can assist traditional translators (Green et al., 2013), it could also facilitate SI. Computer assistance could also improve communication in low-resource settings where nonprofessionals interpret *ad hoc* (Valero-Garcés, 2015).

Some interpreters already use computer assistance. Existing tools help in preparing for SI (Costa et al., 2014; Fantinuoli, 2016; Rütten, 2017), but such systems still rely on predictions of what will be said, which in many cases is extremely error-prone (Grissom II et al., 2016; Li et al., 2020). In contrast, real-time assistance recognizes and displays translations for difficult terms as they are uttered by the interpretee, reducing the preparation time before a session.

Vetting real-time assistance is hampered by the logistical and financial (AIIC, 2011) difficulties of finding and hiring from a small pool of SI professionals. In contrast, while crowdsourcing platforms provide near-instant access to multilingual crowd workers, crowd workers lack SI training. We address this disparity by comparing whether researchers' early design phases for SI, vetted with crowd workers (**proxies**), can be used as temporary surrogates for the former. We describe a novel SI task, empirical metric, and interface suitable for proxy participants (Section 1) and describe the procedure and results of an experiment in which professional interpreters and proxy participants perform SI with different assistance (Section 2). We evaluate with standard automatic MT metrics. The two groups respond to the task and assistance in qualitatively similar ways, suggesting that our evaluation is useful for piloting real-time SI assistance with proxies.

We surveyed SI professionals and students (N=20). Interpreters are most interested in receiving help with items difficult to retrieve faithfully from memory, either due to high information density or lexical infrequency. The top three items that respondents said they want displayed are "dates and numbers mentioned by the speaker", "names of people and places mentioned by the speaker", and "translations of individual key terms".[1] But our re-

[1]Other options were "biographical information about the speaker", "full text documents relevant to the speech", "short excerpts from documents mentioned by the speaker", "a visual record of speaker utterances", "automatic written translation of the speech," "previous translations/interpretations of speakers/cited documents", "translations of idiomatic expressions", and a write-in option.

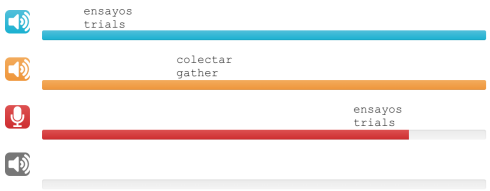


Figure 1: Display for SI task for proxy participants. As source audio plays, progress bars fill from top to bottom. Blue progress bars indicate that users need only listen, orange progress bars warn that the user should prepare to translate, and red progress bars indicate that the user should be translating the presented audio. In the HELP conditions, the source and suggested translation for terms are above the progress bar, aligned with the occurrence of the word in the audio.

spondents were acutely aware that providing too much information could have the opposite effect and be distracting. When asked what they thought disadvantages or challenges of using an assistive interface might be, the most frequent concern was being overloaded with input and distracted from the incoming words. To minimize this risk, assistance should focus on the high-priority items: numerals and dates, named entities, and key terms.

1. An Experimental SI Task

Interpreters are in-demand, rare professionals. Broadening the participants who can vet SI interfaces would allow researchers to rapidly pilot changes in the early design process and perform trials for statistically reliable results (Travis, 2016), enabling more rapid progress, much as BLEU (Papineni et al., 2002) facilitated MT progress. Untrained bilingual participants are not perfect stand-ins for interpreters, but they might suffice as **proxies** (Boyd-Graber et al., 2006) to guide the design.

Professional simultaneous interpreters are trained to cope with the challenges of their work, and they show advantages in *cognitive control*—the ability to multitask or rapidly switch between tasks—compared to other multilingual professionals such as translators (Becker et al., 2016). However, we hypothesize that bilingual proxies will respond similarly to interpreters when working in conditions that address their comparative shortcomings. Our task aims to present proxy participants with the same challenges faced by professionals while providing scaffolding to prevent them from being overwhelmed.

The display for our proxy SI task (Figure 1) alleviates some of SI's cognitive overhead while guiding participants to listen to or interpret sentences of a speech heard over headphones.

SI challenges fall into four major areas (Gile, 2009): listening while speaking (multitasking), rapid translation (timing), speech topic area understanding (content), and maintaining stamina (stamina).

Multitasking and timing are key to SI, and while we expect proxy participants to struggle with these, completely removing them would invalidate our results. We make two slight concessions to mitigate multitasking: first, progress bars indicate how much of the sentence remains; second, we add a three-second delay before the participant translates, giving them time to finish processing. Content is our focus here: our experimental conditions manipulate whether translations for terms appear above the sentence progress bar just after they are uttered. Stamina is the most directly related to practice, so it can reasonably be mitigated for proxy participants. In our task, participants are only directed to interpret intermittent nonconsecutive sentences.[2] This technique was used in Lasecki et al. (2012) to help untrained participants in a monolingual real-time captioning task.

2. Evaluating Assistance

We compare the quality of interpretations produced across four assistance conditions with our SI task:

- **NO HELP**: baseline condition where no terms or translations are displayed to the participant
- **EXPERT HELP**: baseline condition where the system displays tokens that a human interpreter deemed difficult
- **NUMBER HELP**: system displays translations for the names of numeric values[3]
- **TERM HELP**: system displays the most difficult non-stopword tokens according to a metric described in A, closely related to word translation entropy (Schaeffer et al., 2015)

Each HELP condition displays roughly the same number of terms per speech.

2.1. Finding Speeches

The Spanish speeches come from audio files for five Creative Commons-licensed YouTube videos (Mayo Clinic, CDC, and FDA) that were either re-recorded or human-captioned in Spanish and English to provide gold standard translations. English speeches are four excerpted recordings of quarterly earnings calls from four companies (Alphabet, Blizzard, Costco, and United Healthcare)

[2] At least two sentences play to provide initial context before asking participants to translate. At least one sentence comes between translated sentences.

[3] The idea for NUMBER HELP came from a term paper written by Nathan Anderson at Brigham Young University, which showed that professionals benefit from such assistance. Desmet et al. (2018) also show this.

transcribed and professionally translated into Spanish. The translations to display for each speech and HELP condition are decided based on the transcriptions for each speech and its gold translation. All speeches are about five minutes.

2.2. Procedure for Proxies

Proxy participants (N=64) were recruited from Upwork and Amazon Mechanical Turk (AMT) for either an English-to-Spanish or Spanish-to-English interpretation task. They received 12 USD for their participation in a 45-minute session. At the beginning of the task, they tested whether their computer's microphone worked and took a short screening quiz for English and Spanish fluency.[4] Before the task begins, participants click through a tutorial introducing the interface and the task instructions, including a short video modeling how to follow the visual cues from the interface. They then try the task on the 90-second practice video. Afterwards, participants are recorded in each of three conditions, in random order, on three randomly selected speeches from the appropriate language. For the Spanish-English tasks, the three conditions were NO, EXPERT, and TERM HELP. For the English-Spanish tasks, the three conditions were NO, TERM, and NUMBER HELP.[5]

2.3. Procedure for Professionals

Our simplified SI procedure is only useful for assistance testing if proxies in our task behave like professional interpreters doing full-blown SI. We compare proxy participants' translation quality to that of professionals who each interpret three full five-minute speeches rather than selected sentences, without any pauses between sentences.

Professional participants (N=13) were recruited from ProZ.com, a freelancing website for language professionals, for either a Spanish-English or English-Spanish interpretation. They received 25 USD for their participation in a 20-minute session. After testing microphone functionality, these participants are introduced to the interface through a short tutorial. Then they record each of three conditions, in random order, on three randomly selected speeches.

[4]This quiz played audio clips of a sentence in the speech language and then users selected a target-language translation from four choices. Only participants who chose three out of four correct responses were allowed to continue.

[5]NUMBER HELP replaces EXPERT HELP for the English-to-Spanish experiments because the Spanish-to-English experiments came first and indicated that EXPERT HELP was not useful. We switch to English as the source language because earnings calls—which are number-heavy—are conducted in English.

2.4. Patterns in Translation Quality

Each participant's translation is transcribed by a professional transcription service. Transcribed sentences are scored against the corresponding gold-standard translation with the METEOR metric (Banerjee and Lavie, 2005), which is preferred over BLEU for sentence level quality assessment (Stanojević and Sima'an, 2014). The more recall-oriented METEOR is also preferable to the precision-based BLEU for SI, for which meaning preservation is favored over literalism.

We want to know whether proxies doing simplified SI and professionals doing real SI respond similarly to changing conditions, including our HELP conditions. Thus, we fit a linear mixed-effects model to examine factors' effects on translation quality. We focus on factors likely to affect translation quality and consider their interaction with whether a participant is a professional or not. The dependent variable is sentence METEOR score, with fixed effects of participant type (`Pro` in Table 1), condition (with NO HELP as the control), source sentence word count (`WordCount`), total term translation difficulty[6] according to our difficulty metric (`Difficulty`), and number of previously attempted sentences (`SentenceNo`). We consider interactions between `Pro` and all other factors. We account for participant and source speech as random effects.[7]

3. Results and Analysis

Similar Patterns in Proxies and Professionals

Proxy METEOR scores varied more (0.035–0.380) than professional METEOR scores (0.140–0.277). Table 1 shows mixed-effects coefficients.

On 1583 sentences from 64 proxy participants, sentences in the TERM HELP and NUMERIC HELP conditions have higher METEOR scores than in the other two conditions (Figure 2). The same is true of the 1233 sentences from 13 professionals.

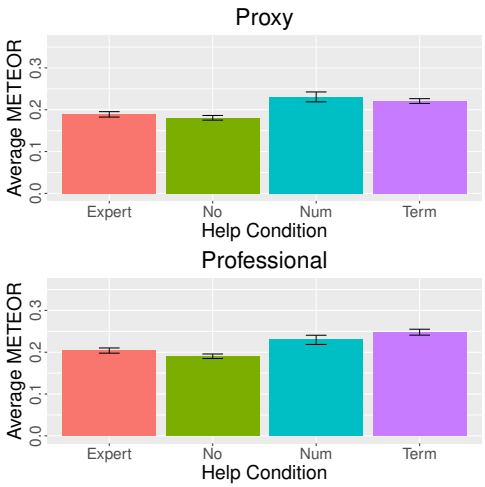
Comparing Professionals and Proxies

While we also include smoothed BLEU[8] to show that both professionals and proxies obtain similar score patterns across more than one metric, we focus on METEOR scores because they are more robust to synonymy and emphasize unigram recall, both of which are more appropriate for SI. SI demands multitasking in short, punctuated bursts:

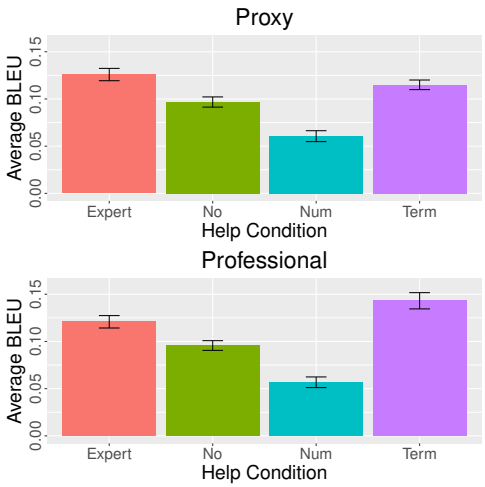
[6]Sum of translation difficulties for each term.

[7]A model including the source language and its interactions with all other factors does not significantly change the main result but increases model complexity. Including source speech as a random effect implicitly accounts for factors that may relate to speech language.

[8]We use the NIST geometric sequence smoothing method investigated by Chen and Cherry (2014).



(a) Average METEOR



(b) Average sentence-level BLEU

Figure 2: Average sentence-level METEOR/BLEU scores across our four conditions with std. error. Both proxies and professional interpreters respond similarly to help, although proxies get more help from NUM assistance.

Fixed Effect	β	p
Intercept	$\mathbf{0.220}$	$\mathbf{0.000*}$
Pro	0.003	0.890
EXPERT HELP	0.009	0.252
TERM HELP	$\mathbf{0.020}$	$\mathbf{0.003*}$
NUMBER HELP	$\mathbf{0.029}$	$\mathbf{0.009*}$
WordCount	$\mathbf{0.001}$	$\mathbf{0.049*}$
Difficulty	$\mathbf{-0.011}$	$\mathbf{0.000*}$
SentenceNo	-0.000	0.360
Pro&EXPERT HELP	0.002	0.188
Pro&TERM HELP	0.001	0.381
Pro&NUMBER HELP	0.019	0.262
Pro&WordCount	-0.001	0.180
Pro&Difficulty	$\mathbf{0.008}$	$\mathbf{0.018*}$
Pro&SentenceNo	0.000	0.830

Table 1: Summary of linear mixed-effects model. TERM and NUMBER HELP conditions, as well as increasing WordCount, have significant positive effects on translation quality, while Difficulty negatively impacts quality. Coefficients assigned to the effect β in a linear regression predicting METEOR score with statistically significant effects ($p < 0.05$) are bolded with an asterisk.

proxies' METEOR scores are comparable to professionals'. In some cases, proxies have higher overall METEOR scores than professionals. This is unsurprising, since it is much easier to interpret nonconsecutive sentences with pauses between.

Proxies and professionals have similar responses to the HELP conditions. METEOR is significantly higher in the TERM HELP and NUMBER HELP conditions, but the EXPERT HELP condition fails to significantly improve scores compared to the NO HELP baseline.

Overall, there is a significant positive, albeit small, effect of WordCount, so both groups better translate longer sentences. Proxies do not suffer significant translation quality losses due to fatigue by the end of long sentences.

Regarding SentenceNo's lack of interaction, both groups could be expected to do worse as the task goes on because of fatigue, but proxies who may never have interpreted before might be expected to improve with exposure and practice.

Our model shows one clear difference between the groups: as Difficulty of a sentence increases, translation quality suffers. However, the significant positive interaction effect between the Difficulty and Pro factors is almost large enough to counteract this trend for the professional group. In other words, professionals are better than proxies at handling sentences with more difficult terminology. Thus our difficulty metric does not accurately predict translation difficulty for professionals, though it is strongly correlated with proxies' translation difficulty.

Why TERM HELP beats EXPERT HELP

The model predicts that TERM HELP and NUMBER HELP lead to higher METEOR versus the NO HELP condition. This is consistent with the higher mean METEOR scores for these conditions (Figure 2). It is curious that these interventions lead to more reliable improvement than EXPERT HELP. One reason may be that a single annotator's opinions on translation difficulty may not generalize to people with different experiences and domain knowledge. We also suspect that the higher rate of infrequent cognates translated in the EXPERT HELP condition make it a less effective strategy.

4. Related Work

4.1. Predicting Lexical Difficulty for Interpreters and Bilinguals

Stewart et al. (2018) designed a real-time measure of interpreter difficulty based on features of the interpreter's productions compared to the source speech. Their features were an extension of the QuEst++ (Specia et al., 2015) sentence-level feature set that included more specific indicators of SI difficulty, such as use of near-cognates, use of nonspecific words, and disfluencies. In contrast, our measure is designed to predict difficulty for an interpreter independent of current performance. An ideal assistance system would use both kinds of feedback to tailor its output to each interpreter.

The similarity of vector space representations can predict the strength and timing of psycholinguistic phenomena associated with retrieval, including priming (Ettinger and Linzen, 2016; Jones et al., 2006; Lapesa and Evert, 2013; Mandera et al., 2017) and N400 response in a highly predictive context (Ettinger et al., 2016; Parviz et al., 2011). All of these results are for a monolingual setting, however, and we are unaware of any previous work investigating psychological correlates in multilingual embedding spaces. We chose the more transparent topic model-based vector space representations (Appendix A) for our experiments in hopes that a simpler and lower-resource method could maintain predictive power. Our performance metric of translation score is very far downstream from the psycholinguistic measurements investigated in previous work. Thus it is possible that our assistance paradigm improves translation quality for different reasons than we expect and doesn't actually ease word recognition and retrieval as would be reflected by such measurements.

4.2. Adapting a Task for Nonprofessionals

In human-computer interaction research focusing on highly specific user groups, it is very common to conduct case studies and qualitative assessments insteaf of trying to gather statistically robust results. There is precedent, however, for shortening the working interval of a normally continuous task so that untrained professionals can keep up. Lasecki et al. (2012) used this technique for crowdsourcing real-time captions for an ongoing speech. This task is similar to SI insofar as it requires listening while producing, though the sentence production is written rather than spoken and the task monolingual.

5. Limitations and Conclusions

Despite some differences, our evaluation setup provides consistent results between expert and proxy participants. This can allow more user-focused development of aids for simultaneous translation, enabling human-in-the-loop testing more quickly and with lower cost. Additionally, the translation difficulty metric detailed in Appendix A predicts downstream translation difficulty for nonprofessionals when used with gold-standard translations. One important caveat, however, is that our experiments compare professionals and proxies solely on the basis of automated metrics. Future work can explore the finer-grained differences between professional and nonprofessional interpretation assistance affects them.

We explore two novel ideas: a term-translation difficulty metric based on vector difference and a paradigm for testing interpretation interventions with untrained bilingual participants. We demonstrate that untrained bilingual participants can perform an adapted interpretation task and produce translations whose quality differs systematically with changes in stimuli. Other researchers with real-time interpreter assistance proposals can use this paradigm to pilot their designs online with a large population before seeking out professional case study participants.

Our experimental results support that vector distance between a source term and its target translation predicts translation difficulty. This finding may be relevant outside of interpretation assistance, for instance in interpreter training or in prioritizing vocabulary practice for second language learners. Future work cam address whether this difficulty score is still predictive when founded on automatic instead of gold-standard speech recognition and translation and when based on vector representations other than topic model-derived ones.

5.1. Interpretation for Non-professionals

Much more testing and comparison is required before we can conclude that non-professional behavior on our task mirrors relevant aspects of professional interpreter behavior. However, we have demonstrated as a proof of concept that untrained bilingual participants can perform an adapted interpretation task and produce translations whose quality differs systematically with changes in stimuli. Other researchers with ideas for real-time interpreter assistance may find this task paradigm useful for piloting their designs online with a large population before seeking out professional case study participants.

Ethical Considerations

Our study is approved by an institutional review board: professional interpreters are paid $50 per hour and proxies are paid $12 per hour.

The goal of the research is to lower the mental load and mental stress of professional interpreters. Our hope is that this would make the profession more accessible to potential interpreters. One theme that came up in discussions with professional interpreters was whether this technology could replace skilled interpreters. Given the current state of the research, this is highly unlikely. Indeed, given the difficulty of the task and the high stakes, it would be extremely ill-advised to attempt to replace interpreters with technology.

6. Bibliographical References

AIIC. 2011. How much will it cost. *International Association of Conference Interpreters*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Maxi Becker, Torsten Schubert, Tilo Strobach, Jürgen Gallinat, and Simone Kühn. 2016. Simultaneous interpreters vs. professional multilingual controls: Group differences in cognitive control as well as brain structure and function. *Neuroimage*, 134:250–260.

Y. Bengio, A. Courville, and P. Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Jordan Boyd-Graber, Sonya S. Nikolova, Karyn A. Moffatt, Kenrick C. Kin, Joshua Y. Lee, Lester W. Mackey, Marilyn M. Tremaine, and Maria M. Klawe. 2006. Participatory design with proxies: Developing a desktop-PDA system to support people with aphasia. In *SIGCHI*.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 362–367.

Hernani Costa, Gloria Corpas Pastor, and Isabel Durán Muñoz. 2014. Technology-assisted interpreting. *MultiLingual*, 143(25):3.

Bart Desmet, Mieke Vandierendonck, and Bart Defrancq. 2018. Simultaneous interpretation of numbers and the impact of technological support. In *Interpreting and technology*, pages 13–27. Language Science Press.

Allyson Ettinger, Naomi Feldman, Philip Resnik, and Colin Phillips. 2016. Modeling N400 amplitude using vector space models of word representation. In *CogSci*.

Allyson Ettinger and Tal Linzen. 2016. Evaluating vector space models using human semantic priming results. In *Workshop on Evaluating Vector-Space Representations for NLP*, pages 72–77.

Claudio Fantinuoli. 2016. Interpretbank. redefining computer-assisted interpreting tools. In *Proceedings of the Translating and the Computer 38 Conference in London*, pages 42–52.

Daniel Gile. 2009. *Basic concepts and models for interpreter and translator training*, volume 8. John Benjamins Publishing.

Spence Green, Jeffrey Heer, and Christopher D. Manning. 2013. The efficacy of human post-editing for language translation. In *SIGCHI*, page 439–448.

Alvin Grissom II, Naho Orita, and Jordan Boyd-Graber. 2016. Incremental prediction of sentence-final verbs: Humans versus machines. In *Conference on Computational Natural Language Learning*, pages 95–104, Berlin, Germany. Association for Computational Linguistics.

Michael N Jones, Walter Kintsch, and Douglas JK Mewhort. 2006. High-dimensional semantic space accounts of priming. *Journal of memory and language*, 55(4):534–552.

Sugandha Kaur and Bidisha Som. 2018. Context effects in bilingual language processing. In *Psycholinguistics and Cognition in Language Processing*, pages 140–168. IGI Global.

Gabriella Lapesa and Stefan Evert. 2013. Evaluating neighbor rank and distance measures as predictors of semantic priming. In *Workshop on Cognitive Modeling and Computational Linguistics*.

Walter Lasecki, Christopher Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey Bigham. 2012. Real-time captioning by groups of non-experts. In *UIST*.

Wenyan Li, Alvin Grissom II, and Jordan Boyd-Graber. 2020. An attentive recurrent model for

incremental prediction of sentence-final verbs. *Findings of EMNLP*.

Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2017. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92:57–78.

Cristina Mazza. 2001. Numbers in simultaneous interpretation. Technical report, EUT-Edizioni Università di Trieste.

Andrew Kachites McCallum. 2002. MALLET: A machine learning for language toolkit.

David Mimno, Hanna Wallach, Jason Naradowsky, David Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Conference of Empirical Methods in Natural Language Processing*.

Behrang Mohit and Rebecca Hwa. 2007. Localization of difficult-to-translate phrases. In *StatMT*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Mehdi Parviz, Mark Johnson, Blake Johnson, and Jon Brock. 2011. Using language models and latent semantic analysis to characterise the n400m neural response. In *Australasian Language Technology Association Workshop*.

Clara Pignataro. 2012. Terminology and interpreting in lsp conferences: A computer-aided vs. empirical-based approach. Technical report.

Anja Rütten. 2017. Terminology management tools for conference interpreters–current tools and how they address the specific needs of interpreters. *The SCATE Prototype: A Smart Computer-Aided Translation Environment*, page 98.

Moritz Schaeffer, Barbara Dragsted, Kristian Tangsgaard Hvelplund, Laura Winther Balling, and Michael Carl. 2015. Word translation entropy: Evidence of early target language activation during reading for translation.

Lucia Specia, Gustavo Paetzold, and Carolina Scarton. 2015. Multi-level translation quality prediction with QuEst++. *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 115–120.

Miloš Stanojević and Khalil Sima'an. 2014. Fitting sentence level translation evaluation with many dense features. In *Conference of Empirical Methods in Natural Language Processing*, pages 202–206.

Craig Stewart, Nikolai Vogler, Junjie Hu, Jordan Boyd-Graber, and Graham Neubig. 2018. Automatic estimation of simultaneous interpreter performance. In *Proceedings of the Association for Computational Linguistics*, Melbourne, Australia.

Mark Steyvers and Thomas Griffiths. 2007. Probabilistic topic models. *Handbook of Latent Semantic Analysis*.

D Travis. 2016. Usability testing with hard-to-find participants. *Userfocus. co. uk*.

Carmen Valero-Garcés. 2015. Cross-fertilization of training and research in a master's program in public service interpreting and translation: Some challenges and results. In *Handbook of Research on Teaching Methods in Language Translation and Interpretation*, pages 397–415. IGI Global.

A. Term Help Method

To find difficult terms, our scoring algorithm finds words with sense-dependent translations. Our metric's focus on distribution over possible senses is closely related to the "probability of translation ambiguity" (Mohit and Hwa, 2007) to predict terms difficult for MT systems. It's also similar to word translation entropy—the number of alternate translations possible for a given word—which has been correlated with cognitive difficulty in eye tracking studies (Schaeffer et al., 2015).

Infrequent or specific terms are not always the hardest to translate; nor are common words necessarily easy. For example, *nasopharynx* is a medicine-specific, infrequent English term, but its Spanish cognate *nasofaringe* is easy for an interpreter to retrieve because of cognate facilitation (Kaur and Som, 2018). Conversely, consider the English word *thread*, which has at least three distinct senses and appropriate Spanish context-dependent translations (*hilo* for sewing *thread*, *tema* for online discussion *thread*, and *subproceso* for computational *thread*). Technical language abounds in such cases of commonplace terms taking on specialized meanings.

Representation learning (Bengio et al., 2013), particularly distributed semantic representations, can capture words' real world usage. Vector representations of meaning can predict the strength and timing of psycholinguistic phenomena associated with retrieval, including priming (Ettinger and Linzen, 2016; Jones et al., 2006; Lapesa and Evert, 2013; Mandera et al., 2017) and N400 response in a highly predictive context (Ettinger et al., 2016; Parviz et al., 2011).

We use on multilingual topic models (Mimno et al., 2009) to uncover coarse senses (Steyvers and Griffiths, 2007). If a source term has a different distribution of topics from its target translation, the term and translation should have more distant representations. Topic models are attractive in this context for multiple reasons: they are easy to train for small corpora, can capture document-level themes (as opposed to local context), and are intuitive for humans to understand. The last feature is particularly important, as topic distributions can help an interpreter organize or display terms.[9]

Each term w has a distribution over topics $\vec{v}_w \equiv p(z \mid w) = p(w \mid z)p(z)/p(w)$. We hypothesize the translation difficulty between two terms $\delta_{s,t}$ as

$$\delta_{s,t} = \text{JDist}(\vec{v}_s, \vec{v}_t) + \frac{1}{\log(n_s + n_t)} \quad (1)$$

where n_x is the number of occurrences of type x in the training corpus and JDist is Jaccard Distance.[10] The second term in the equation boosts infrequent terms, which may be hard to translate simply because they haven't been previously encountered.

We train a 400-topic model for 1,000 iterations on Wikipedia pages that have English and Spanish versions using Mallet (McCallum, 2002), These values were chosen by manual inspection of topic quality. providing representations of 5.9M English and 2.5M Spanish lemmatized words and phrases. Before training, multiword phrases with Wiktionary entries are grouped together as single words. Our translations were compiled into a dictionary by scraping entries in English and Spanish Wiktionary, and we ended up with multiple translations for over 55k Spanish words and phrases.

We attempted to use our topic-space representations to automatically select the best translation for each word in our experimental stimuli—Spanish-language medical PSAs pulled from YouTube—but found that the bag-of-words context information considered by the topic model was not sufficient for accurate translation selection. Our translation selection method was to select the translation for a term whose topic profile best matched the topic profile of the speech as a whole. Given a current context of topic c from our topic model and a set of translations T_s for source term s, we determined the best translation for s based on:

$$\underset{T_s}{\text{argmax}}\ v^c_{t_i} p(t_i) \quad (2)$$

where $p(t_i)$ is the probability of t_i in the corpus, and $v^c_{t_i}$ is the probability of assigning topic c to t_i according to our topic model—in other words, the value of the vector v_{t_i} at the position corresponding to topic c. The context c was selected as the argmax of the elementwise product of v_s and v_d, the topic distribution of the transcript of the video as inferred from our topic model. Unsurprisingly, the resulting translations are not very accurate: the most appropriate translation for a term was selected only 46% of the time across all five experimental stimuli. Since our primary concern is the validity of piloting interventions intended for professional interpreters with nonprofessionals, we abstract away from selecting term translations in our experiment and instead opt for gold-standard translations. A real assistance system would require a high-fidelity MT selection scheme and would need to account for its uncertainty when determining translations to display.

[9] We envision offline vocabulary organization as another feature of a potential assistive interface.

[10] For vectors x and y, Jaccard Distance is defined as $1 - \sum_i \frac{\min(x_i, y_i)}{\max(x_i, y_i)}$

Term	Translation	Score	Log Inverse Frequency
nefrotoxina	nephrotoxin	1.434	0.434
leucotoxina	leukotoxin	1.319	0.319
urticaria	hives	1.093	0.118
lote	lot	0.427	0.097
artritis	arthritis	0.427	0.118
reemplazo	replacement	0.425	0.088
estenosis	stenosis	0.132	0.132
fenotipo	phenotype	0.131	0.116
biopsia	biopsy	0.120	0.120

Table 2: Difficulty scores for the highest-, median, and lowest-scoring term-translation pairs in our experimental stimuli. 'Log Inverse Frequency' corresponds to the second term in our difficulty score given in Equation 1. In bold are terms that our informant marked as worth displaying. In italics are terms that our model selected as worth displaying.

A.1. Assessing Validity of the Model's Difficulty Ratings

For an initial assessment of validity of this difficultly metric, an informant with Spanish–English interpreting experience labeled terms in our Spanish experimental stimuli that they considered difficult. Table 2 characterizes some of the differences between the informant's judgments and our model's.

The lowest-scoring terms, as expected, are cognates that are relatively frequent in both English and Spanish. The two highest-scoring terms are also cognates, but appear fewer than four times each in Spanish Wikipedia and did not receive accurate distributions over contexts in our topic model. The third highest, *urticaria* (which can mean either "skin rash" or "anxiety"), is far more frequent but is translated in our source material into *hives*, which has a different set of senses (primarily "insect hive", secondarily "skin rash").