Predicting Human Translation Difficulty Using Automatic Word Alignment

Zheng Wei Lim, Trevor Cohn,* **Charles Kemp** and **Ekaterina Vylomova**
The University of Melbourne
z.lim4@student.unimelb.edu.au, {t.cohn,c.kemp,vylomovae}@unimelb.edu.au

Abstract

Translation difficulty arises when translators are required to resolve translation ambiguity from multiple possible translations. Translation difficulty can be measured by recording the diversity of responses provided by human translators and the time taken to provide these responses, but these behavioral measures are costly and do not scale. In this work, we use word alignments computed over large scale bilingual corpora to develop predictors of lexical translation difficulty. We evaluate our approach using behavioural data from translations provided both in and out of context, and report results that improve on a previous embedding-based approach (Thompson et al., 2020). Our work can therefore contribute to a deeper understanding of cross-lingual differences and of causes of translation difficulty.

1 Introduction

Words can be hard to translate for many reasons including cultural context and differences in semantic subdivisions across languages (Hershcovich et al., 2022; Chaudhary et al., 2021). For instance, the emotional/moral sense of English *heart* is commonly translated as Malay *hati*, but in a medical setting *heart* should be translated as *jantung* and *hati* refers to a different bodily organ (the liver). Examples such as this are challenging for language learners and translators because they go beyond simple one-to-one correspondences between source and target words.

Translation difficulty has been studied by researchers from multiple disciplines including psycholinguistics (Degani et al., 2016), computational linguistics (Cotterell et al., 2018), machine translation (Koehn and Knowles, 2017) and translation studies (Carl et al., 2016b). Understanding translation difficulty is an important scientific challenge in its own right, but methods for measuring difficulty can also be applied in a number of ways. First, difficulty measures have previously been used to identify word meanings of cultural significance (Toury, 2021; Thompson et al., 2020). Second, difficulty measures can help develop targeted evaluations for Neural Machine Translation (NMT) systems (Bugliarello et al., 2020; Yin et al., 2021). For example, automatic difficulty ratings allow for the generation of translation samples of varying difficulties and facilitate human evaluation of machine translation. A third potential application is to reweight and calibrate NMT performance across data sets, language pairs and domains based on their varying levels of difficulty – an objective crucial to NMT quality estimation tasks (Fomicheva et al., 2020; Behnke et al., 2022). Finally, in second language learning and human translator training, translation difficulty ratings allow instructors to identify potential challenges to language learners, and to curate translation assignments for translation students of different levels of experience (Sun, 2015; Chaudhary et al., 2021).

In this work, we use surprisal and entropy derived from word alignment to estimate word translation difficulty. Different pairs of aligned words are collected as translation alternatives (e.g., *heart-hati* and *heart-jantung*) and used to infer a word's translation distribution and compute our information-theoretic difficulty measures. Among previous studies of translation our approach is closest to the work of Chaudhary et al. (2021), as it leverages large-scale human translation data and extracts word-level translations from aligned sentences. Unlike previous studies, however, we are the first to use word alignments to directly address translation difficulty as a psychological aspect of lexical semantics that is measurable by behavioural data. We evaluate our difficulty estimates against translation norms (Tokowicz et al., 2002; Prior et al., 2007; Allen and Conklin, 2014; Bracken et al., 2017; Lee et al., 2022; Tseng et al., 2014) that measure

*Now at Google DeepMind.

Findings of the Association for Computational Linguistics: ACL 2023, pages 11590–11601

translation difficulty out of context and translation process features (Carl et al., 2016b) that measure translation difficulty in context. We also compare against a previous approach that uses multilingual embeddings to develop a measure of translation difficulty (Thompson et al., 2020).

Relative to embeddings, we suggest that word alignments better capture lexical and morphological distributions, and hence allow for a better measure of translation difficulty.[1] Our measures of translation difficulty are interpretable, and as we show in later sections, help improve the understanding of human language and translation processing.

2 Related Work

Our approach builds on two lines of work from the psycholinguistic literature on translation. One line of work relies on translation norms derived from tasks in which bilingual participants translate single words presented out of context or rate semantic similarity between pairs of words (Tokowicz et al., 2002; Prior et al., 2007; Allen and Conklin, 2014; Bracken et al., 2017; Lee et al., 2022; Tseng et al., 2014). High variation in translation responses to a given word provides evidence of translation ambiguity (Kroll and Tokowicz, 2001; Tokowicz, 2000); whereas perceived degree of cross-lingual semantic overlap informs lexical choice and is predictive of response time (Allen and Conklin, 2013; Van Assche et al., 2009; Dijkstra et al., 2010; Van Hell and De Groot, 1998). A second line of work studies translation in context by measuring reading time and production duration as translators process realistic texts (Carl et al., 2016b). Behavioral approaches like these provide gold-standard measures of translation difficulty but are costly and do not scale.

Within the computational literature, Thompson et al. (2020) and Carl (2021b) derive automatic measures of translation difficulty based on the idea that difficult-to-translate words are hard to align across word embedding spaces. The former use embeddings to compare semantic neighbourhoods of bilingual word pairs and report significant correlations with human semantic similarity judgements. Carl (2021b) learned a cross-lingual embedding projection to estimate word pair similarities, and showed that these estimates predict translation process data. Bugliarello et al. (2020) and Yin et al. (2021) probe translation ambiguity from NMT models using cross-mutual information, which is useful for identifying contextual translations in NMT models. Chaudhary et al. (2021) use word alignment distributions to reveal lexical semantic distinctions across languages. Their work shows that word alignments, with properly extracted descriptions, help language learners disambiguate fine-grained lexical distinctions, but does not directly address the general notion of translation difficulty.

3 Assessing word-level translation difficulty through word alignments

Assume that we have a parallel corpus and a word aligner and are interested in the translation distribution of word w from a source language, L1, to a target language, L2. The most natural approach is a count-based distribution, where counts for words aligned with w are normalized by the frequency of w. From here, $p_{al}(v|w)$, the probability of word w being translated to v, can be computed over aligned word pairs. In addition to alignment counts, most word aligners assign a score for each pair of aligned words. Given two parallel sequences, $x = [x_0, ..., x_m]$ and $y = [y_0, ..., y_n]$, let $x_i \leftrightarrow y_j$ indicate that the ith token from x is paired with the jth token from y, and let $s_{x_i \leftrightarrow y_j}$ denote the alignment score.[2] This allows a weight-based distribution parallel to the count-based method above. In general, we calculate $p_{al}(v|w)$ by:

$$p_{al}(v|w) = \frac{S_{w \leftrightarrow v}}{\sum_{u \in V} S_{w \leftrightarrow u}}. \quad (1)$$

For the weight-based distribution, $S_{w \leftrightarrow v}$ represents the sum of alignment scores of all $w \leftrightarrow v$ pairings. For the count-based distribution, $s_{x_i \leftrightarrow y_j} = 1$, i.e., $S_{w \leftrightarrow v}$ is the number of times w is aligned with v in the entire corpus.[3] The final distribution, $p_{al}(v|w)$, is normalized given the total scores of all possible alignments with w, where V refers to the vocabulary of L2 in the corpus.

The concept of surprisal in psycholinguistics is often associated with cognitive workload, which in translation studies is connected to word translation information (ITra) (Wei, 2022; Carl, 2021a).

[1] Code available at https://github.com/ZhengWeiLim/pred-translation-difficulty.

[2] In Dou and Neubig (2021), $s_{x_i \leftrightarrow y_j}$ is the harmonic mean between $p(y_j|x_i)$ and $p(x_i|y_j)$, the probability of x_i being aligned to y_j over all possible words in y, and vice versa.

[3] $S_{w \leftrightarrow v} = \sum_{x \leftrightarrow y} \sum_{i,j} s_{x_i \leftrightarrow y_j} \mathbb{1}_{\{w,v\}}(x_i, y_j)$

Translation surprisal is defined as:

$$I_{al}(v|w) = -\log p_{al}(v|w). \quad (2)$$

Low surprisal values indicate that v is a stable translation of w, which is expected to require low effort to produce. The translation uncertainty associated with a source word w can be formulated as the entropy (or expected surprisal):

$$H_{al}(w) = -\sum_{u \in V} p_{al}(u|w) \log p_{al}(u|w). \quad (3)$$

Surprisals derived from count-based and weight-based distributions are denoted by I^c_{al} and I^w_{al} respectively. Likewise, H^c_{al} and H^w_{al} will be used as shorthands for their respective entropy values. Word pairs with higher surprisal are expected to be more difficult. Higher translation entropy indicates a greater range of translations for a source word, which is expected to contribute to translation difficulty.

4 Experiments

Dataset and pre-processing We obtain parallel data of English with German (de), Spanish (es), Japanese (ja), Malay (ms), Dutch (nl) and Chinese (zh) from OpenSubtitles (Lison et al., 2018). All sentences are tokenized by the spaCy tokenizer (Honnibal and Montani, 2017), except Malay, for which we use Aksara (Hanifmuti and Alfina, 2020). We choose to preserve word forms in subtitles and evaluation data, because morphological variation, as we see in later sections, partly contributes to translation ambiguity. `awesome-align` is used to infer word alignments from the tokenized parallel sentences.[4] We then calculate surprisal and entropy based on Equations 2 and 3.[5]

Evaluation. We evaluate our methods against context-free translations compiled in existing norms, which include i) the number of unique translations of a word, and cover Spanish, Japanese, Malay, Dutch and Chinese (to and from English); and ii) semantic similarity ratings of paired words between English and Japanese, Dutch and Chinese (Tokowicz et al., 2002; Prior et al., 2007; Allen and Conklin, 2014; Bracken et al., 2017; Lee et al., 2022; Tseng et al., 2014). Measures of translation in context are derived from CRITT TPR-DB, a behavioural data set extracted from translation logs collected using key loggers and eye trackers (Carl et al., 2016b). We focus on three such process features:

- `Dur` specifies the time taken to produce the target token corresponding to a source word.
- `Munit` describes the number of micro units, which are distinct translation activities marked by pauses of a fixed length. Thus, easier translations correspond to lower values of `Munit`.
- `HTra` refers to translation entropy based on manual alignments in TPR-DB.

More details about these three features and about the preprocessing applied are described in Appendix A. We validate against data sets in Japanese (ENJA15, Carl et al., 2016a), German (SG12, Carl et al., 2016b) and Spanish (BML12, Mesa-Lao, 2014), for which information about translation at the token-level is readily available.

Baselines. We compare I^c_{al} and I^w_{al} with Thompson et al.'s (2020) embedding-based approach, which has been framed explicitly as an account of translation difficulty. Following their work we expand the initial NorthEuraLex translations (Dellert et al., 2020) to include all translation pairs in the evaluation data and recompute word-pair semantic alignments using Common Crawl and Wikipedia fastText embeddings (Grave et al., 2018).

The final values are negated to match the sign of I_{al}, and denoted here by S_{emb}.[6] Thompson et al. (2020) do not provide an embedding-based analog of H^c_{al} and H^w_{al} that can be used to estimate

		es	ja	ms	nl	zh
→ **en**	M_{emb}	.300	.341	-	.247	-
	H^c_{al}	.442	.563	.255	.264	-
	H^w_{al}	**.451**	**.570**	**.266**	**.270**	-
en →	M_{emb}	.351	.461	-	**.358**	.284
	H^c_{al}	**.487**	.525	.430	.250	.348
	H^w_{al}	**.487**	**.538**	**.440**	.248	**.351**

Table 1: Pearson correlation (the higher the better) between alignment distribution entropy and number of translations in translation norms. All values shown are significant ($p < .001$). '-' indicates missing values (e.g. the Chinese norms do not include English translations).

[4] without `--train_co` option for consistency optimization.

[5] Other pre-processing steps are described in Appendix B.

[6] In Appendix C, we include alternative results computed from OpenSubtitles embedding and translation pairs with additional top 3 aligned translations of the initial vocabulary.

		ja	nl	zh
→ en	S_{emb}	**-.422**	-.302	-.332
	I^c_{al}	-.200	**-.587**	**-.474**
	I^w_{al}	-.194	**-.587**	-.471
en →	S_{emb}	-.422	-.284	-.332
	I^c_{al}	**-.474**	**-.476**	**-.486**
	I^w_{al}	.-.471	-.474	-.484

Table 2: Pearson correlation with word-pair similarity ratings (the lower the better). All values presented are significant ($p < .01$).

		de	es	ja
HTra↑	M_{emb}	.322	.298	.273
	H^c_{al}	.427	**.512**	**.406**
	H^w_{al}	**.428**	.511	.405
Dur (ms) ↑	S_{emb}	-.363	-.466	/
	I^c_{al}	**-.109**	**-.195**	-.161
	I^w_{al}	-.120	-.205	**-.156**
Munit ↑	S_{emb}	.067	/	/
	I^c_{al}	**.269**	**.269**	**.176**
	I^w_{al}	.263	.260	.170

Table 3: Correlations ($p < .05$) between alignment distribution entropy and behavioural data. Non-significant values are omitted with '/'. ↑ denotes the direction of increasing difficulty.

the translation uncertainty associated with a single source word. We therefore compare H^c_{al} and H^w_{al} with a simple embedding-based measure M_{emb} defined as the highest value of S_{emb} associated with a source word. We limit all comparisons to the same set of vocabulary and translation pairs.[7]

5 Results and Discussion

Context-free translations. Table 1 reports the Pearson correlation of all methods given translations to English (→ en) and translations from English (en →). Both H^c_{al} and H^w_{al} achieve moderately high correlations with Spanish and Japanese norms. H^w_{al} is a weight-based entropy, which captures more nuances in its translation distribution than does the count-based approach, and is, in most languages, the most predictive of a source word's translation difficulty. Table 2 summarizes the correlation of I^c_{al}, I^w_{al} and S_{emb} of word pairs against semantic similarity ratings.[8] The count-based and weight-based entropy measures achieve similar correlations and outperform the embedding-derived measure in 5 out of 6 cases.

Context-dependent translations. Table 3 shows that our corpus-derived entropy measures strongly correlate with entropies based on TPR-DB (HTra), and that I^c_{al} and I^w_{al} are moderately predictive of Munit. However, Dur correlates weakly but negatively with I^c_{al}, I^w_{al} and S_{emb}. This finding is surprising — we previously argued that low-surprisal translations and word pairs with high embedding alignment have a larger degree of semantic overlap, which should have contributed to easier translation and shorter production time. The gap between the embedding and word-alignment approaches for Munit and Dur is also considerably larger than for our previous results. We now offer two partial explanations for these observations.

Lexical and morphological variation. Relative to the embedding-based approach, our word-alignment approach more accurately captures the distribution of lexical choices and morphological variants. Rare and morphologically complex words have long been known to affect NMT modeling difficulty (Belinkov et al., 2017; Cotterell et al., 2018), and have relatively poor representations in both static and contextual embeddings (Bahdanau et al., 2017; Conneau et al., 2017; Schick and Schütze, 2019; Athiwaratkun et al., 2018; Schick and Schütze, 2020; Anastasopoulos and Neubig, 2020). Word embeddings are also typically optimized to minimize the contribution of frequency information (Gong et al., 2018; Mu and Viswanath, 2018; Liang et al., 2021; Spliethöver et al., 2022). Ignoring frequency, however, is problematic for our task because frequency captures information about which translation choices are most typical and natural (Baker, 2018). In our data, *varied* is more commonly translated to Spanish feminine form, *variada*, than the masculine form *variado*. Table 4 suggests that *variado* took longer to produce because it appears less frequently in parallel text together with *varied*, as indicated by its surprisal value. Another example that reflects lexical distribution is *disliked*, where *disgustaba* is a more popular translation than *detestaba*. Here S_{emb} fails to distinguish the difference in usage.

[7] The vocabulary size, evaluation set and the number of translations in comparison, are reported in Appendix B.

[8] Unlike alignment distributions, S_{emb} and similarity judgements (except for Dutch norms) are non-directional, resulting in the same values in both directions.

en	es	Dur	S_{emb}	I^c_{al}	I^w_{al}
disliked	*disgustaba*	5.90	-.555	2.79	2.71
	detestaba	8.53	-.555	4.45	4.37
region	*región*	8.01	-.393	0.21	0.20
	zona	7.36	-.391	2.74	2.73
varied	*variada*	6.70	-.587	1.66	1.64
	variado	7.15	-.628	2.05	2.02
	diversa	7.11	-.600	5.76	5.73
	diverso	6.95	-.586	5.76	5.73

Table 4: Translation examples from English to Spanish, showing influences of form similarity, morphology and lexical distribution on production duration.

Effects of form similarity. Our counterintuitive result for Dur is consistent with previous evidence that cognates are both produced with high probability and associated with relatively long production times.[9] Heilmann and Llorca-Bofí (2021) show that the cognate status of a source word increases translation duration (particularly cognate-to-cognate translation), due to hesitation and self-monitoring. Additional evidence that form overlap influences translations is provided by Prior et al. (2011) and Schwartz and Kroll (2006), who found that context helps facilitate non-cognate alternatives to compete in lexical selection. Consistent with these results, we found significant negative correlations between I^c_{al} and I^w_{al} with cognate rating in Spanish norms (Prior et al., 2007) and Mean Form Sim Rating in Dutch norms (Tokowicz et al., 2002), which shows that cognates are indeed more probable translations.[10] For Japanese, we conducted a t-test on surprisals and found a significant difference ($p < .001$) between borrowings and non-borrowings (Allen and Conklin, 2014).[11] Table 4 shows *región* as a more common but slower translation of *region*. Unlike *variada* and *diversa*, the surprisal differential between *variado* and *diverso* is not enough to overcome its cognate effect.

6 Conclusion

We developed predictors of translation difficulty based on word alignment distributions and tested them using translation norms and translation processing data. Compared to the embedding-based approach, our measures derived from word alignment do not depend on lexical databases and more accurately capture lexical choice distributions and morphological variation. Our results show improved estimates of translation difficulty, but suggest that a comprehensive account of human translation difficulty must also consider additional factors such as form similarity.

7 Limitations

Although form similarity is demonstrably responsible for slower translation processing, we are unable to ascertain if it is the primary reason. The work also reveals one shortcoming of alignment distributions — the measure tends to be biased towards translations with similar forms and does not always make accurate predictions about cognates. To address this limitation, future work can evaluate more elaborate models of translation that incorporate variables (e.g., form overlap, syntactic complexity, and morphological complexity) identified as relevant by previous empirical work in psycholinguistics.

Ethics Statement

We obtained all data from cited sources. Our experimental procedures and analysis do not involve human participants and are in compliance with ACL Code of Ethics.[12]

Acknowledgements

This work was supported by ARC FT190100200.

References

David Allen and Kathy Conklin. 2014. Cross-linguistic similarity norms for Japanese–English translation equivalents. *Behavior Research Methods*, 46(2):540–563.

David B Allen and Kathy Conklin. 2013. Cross-linguistic similarity and task demands in Japanese-English bilingual processing. *PloS one*, 8(8):e72631.

Fabio Alves and Daniel Couto Vale. 2017. On drafting and revision in translation: A corpus linguistics oriented analysis of translation process data. *Annotation, exploitation and evaluation of parallel corpora*, pages 89–110.

[9] Here, we use *cognate* liberally to include loanwords and words with high form (orthographical and/or phonological) similarity.

[10] On average, the correlations ($p < .001$) for I^c_{al} are -.213 (es) and -.214 (nl), whereas I^w_{al} are -.214 (es and nl).

[11] S_{emb} is largely uninfluenced by formal similarity.

[12] https://www.aclweb.org/portal/content/acl-code-ethics

Antonios Anastasopoulos and Graham Neubig. 2020. Should all cross-lingual embeddings speak English? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8658–8679, Online. Association for Computational Linguistics.

Ben Athiwaratkun, Andrew Wilson, and Anima Anandkumar. 2018. Probabilistic FastText for multi-sense word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–11, Melbourne, Australia. Association for Computational Linguistics.

Dzmitry Bahdanau, Tom Bosc, Stanisław Jastrzębski, Edward Grefenstette, Pascal Vincent, and Yoshua Bengio. 2017. Learning to compute word embeddings on the fly. *arXiv preprint arXiv:1706.00286*.

Mona Baker. 2018. *In other words: A coursebook on translation*. Routledge.

Hanna Behnke, Marina Fomicheva, and Lucia Specia. 2022. Bias mitigation in machine translation quality estimation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1475–1487.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do Neural Machine Translation Models Learn about Morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872.

Jennifer Bracken, Tamar Degani, Chelsea Eddington, and Natasha Tokowicz. 2017. Translation semantic variability: How semantic relatedness affects learning of translation-ambiguous words. *Bilingualism: Language and Cognition*, 20(4):783–794.

Emanuele Bugliarello, Sabrina J Mielke, Antonios Anastasopoulos, Ryan Cotterell, and Naoaki Okazaki. 2020. It's Easier to Translate out of English than into it: Measuring Neural Translation Difficulty by Cross-Mutual Information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Michael Carl. 2021a. Information and entropy measures of rendered literal translation. In *Explorations in Empirical Translation Process Research*, pages 113–140. Springer.

Michael Carl. 2021b. Translation norms, translation behavior, and continuous vector space models. In *Explorations in Empirical Translation Process Research*, pages 357–388. Springer.

Michael Carl, Akiko Aizawa, and Masaru Yamada. 2016a. English-to-Japanese translation vs. dictation vs. post-editing: comparing translation modes in a multilingual setting. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4024–4031.

Michael Carl, Moritz Schaeffer, and Srinivas Bangalore. 2016b. The CRITT translation process research database. In *New directions in empirical translation process research*, pages 13–54. Springer.

Michael Carl and Moritz Jonas Schaeffer. 2017. Why translation is difficult: A corpus-based study of non-literality in post-editing and from-scratch translation. *HERMES-Journal of Language and Communication in Business*, (56):43–57.

Aditi Chaudhary, Kayo Yin, Antonios Anastasopoulos, and Graham Neubig. 2021. When is Wall a Pared and when a Muro?: Extracting Rules Governing Lexical Selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6911–6929.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Ryan Cotterell, Sabrina J Mielke, Jason Eisner, and Brian Roark. 2018. Are All Languages Equally Hard to Language-Model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 536–541.

Tamar Degani, Anat Prior, Chelsea M Eddington, Ana B Arêas da Luz Fontes, and Natasha Tokowicz. 2016. Determinants of translation ambiguity: A within and cross-language comparison. *Linguistic approaches to bilingualism*, 6(3):290–307.

Johannes Dellert, Thora Daneyko, Alla Münch, Alina Ladygina, Armin Buch, Natalie Clarius, Ilja Grigorjew, Mohamed Balabel, Hizniye Isabella Boga, Zalina Baysarova, et al. 2020. Northeuralex: A wide-coverage lexical database of northern eurasia. *Language resources and evaluation*, 54(1):273–301.

Ton Dijkstra, Koji Miwa, Bianca Brummelhuis, Maya Sappelli, and Harald Baayen. 2010. How cross-language similarity and task demands affect cognate recognition. *Journal of Memory and language*, 62(3):284–301.

Zi-Yi Dou and Graham Neubig. 2021. Word Alignment by Fine-tuning Embeddings on Parallel Corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Manuel Gimenes and Boris New. 2016. Worldlex: Twitter and blog word frequencies for 66 languages. *Behavior research methods*, 48(3):963–972.

Chengyue Gong, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2018. Frage: Frequency-agnostic word representation. *Advances in neural information processing systems*, 31.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Muhammad Yudistira Hanifmuti and Ika Alfina. 2020. Aksara: An Indonesian morphological analyzer that conforms to the UD v2 annotation guidelines. In *2020 International Conference on Asian Language Processing (IALP)*, pages 86–91. IEEE.

Arndt Heilmann and Carme Llorca-Bofí. 2021. Analyzing the Effects of Lexical Cognates on Translation Properties: A Multivariate Product and Process Based Approach. In *Explorations in Empirical Translation Process Research*, pages 203–229. Springer.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, et al. 2022. Challenges and Strategies in Cross-Cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Judith F Kroll and Natasha Tokowicz. 2001. The development of conceptual representation for words in a second language. *One mind, two languages: Bilingual language processing*, 2:49–71.

Soon Tat Lee, Walter JB van Heuven, Jessica M Price, and Christine Xiang Ru Leong. 2022. Translation norms for Malay and English words: The effects of word class, semantic variability, lexical characteristics, and language proficiency on translation. *Behavior Research Methods*, pages 1–17.

Yuxin Liang, Rui Cao, Jie Zheng, Jie Ren, and Ling Gao. 2021. Learning to remove: Towards isotropic pre-trained bert embedding. In *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part V 30*, pages 448–459. Springer.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: extracting large parallel corpora from movie and TV subtitles. In *10th conference on International Language Resources and Evaluation (LREC'16)*, pages 923–929. European Language Resources Association.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Bartolomé Mesa-Lao. 2014. Gaze behaviour on source texts: An exploratory study comparing translation and post-editing. In *Post-editing of machine translation: Processes and applications*, pages 219–245. Cambridge Scholars Publishing.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective post-processing for word representations. In *6th International Conference on Learning Representations, ICLR 2018*.

Anat Prior, Brian MacWhinney, and Judith F Kroll. 2007. Translation norms for English and Spanish: The role of lexical variables, word class, and L2 proficiency in negotiating translation ambiguity. *Behavior Research Methods*, 39(4):1029–1038.

Anat Prior, Shuly Wintner, Brian MacWhinney, and Alon Lavie. 2011. Translation ambiguity in and out of context. *Applied Psycholinguistics*, 32(1):93–111.

Moritz Schaeffer and Michael Carl. 2017. Language processing and translation. *Empirical modelling of translation and interpreting*, 7:117–154.

Moritz Schaeffer, Barbara Dragsted, Kristian Tangsgaard Hvelplund, Laura Winther Balling, and Michael Carl. 2016. Word translation entropy: Evidence of early target language activation during reading for translation. In *New directions in empirical translation process research*, pages 183–210. Springer.

Timo Schick and Hinrich Schütze. 2019. Attentive mimicking: Better word embeddings by attending to informative contexts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 489–494, Minneapolis, Minnesota. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2020. BERTRAM: Improved word embeddings have big impact on contextualized model performance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3996–4007, Online. Association for Computational Linguistics.

Ana I Schwartz and Judith F Kroll. 2006. Bilingual lexical activation in sentence context. *Journal of memory and language*, 55(2):197–212.

Maximilian Spliethöver, Maximilian Keiff, and Henning Wachsmuth. 2022. No word embedding

model is perfect: Evaluating the representation accuracy for social bias in the media. *arXiv preprint arXiv:2211.03634*.

Sanjun Sun. 2015. Measuring translation difficulty: Theoretical and methodological considerations. *Across languages and cultures*, 16(1):29–54.

Bill Thompson, Seán G Roberts, and Gary Lupyan. 2020. Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, 4(10):1029–1038.

Natasha Tokowicz. 2000. *Meaning representation within and across languages*. Ph.D. thesis, The Pennsylvania State University.

Natasha Tokowicz, Judith F Kroll, Annette De Groot, and Janet G Van Hell. 2002. Number-of-translation norms for Dutch—English translation pairs: A new tool for examining language production. *Behavior Research Methods, Instruments, & Computers*, 34(3):435–451.

Gideon Toury. 2021. The nature and role of norms in translation. In *The translation studies reader*, pages 197–210. Routledge.

Alison M Tseng, Li-Yun Chang, and Natasha Tokowicz. 2014. Translation ambiguity between English and Mandarin Chinese: The roles of proficiency and word characteristics. *The development of translation competence: Theories and methodologies from psycholinguistics and cognitive science*, pages 107–165.

Eva Van Assche, Wouter Duyck, Robert J Hartsuiker, and Kevin Diependaele. 2009. Does bilingualism change native-language reading? Cognate effects in a sentence context. *Psychological science*, 20(8):923–927.

Janet G Van Hell and Annette MB De Groot. 1998. Conceptual representation in bilingual memory: Effects of concreteness and cognate status in word association. *Bilingualism: Language and cognition*, 1(3):193–211.

Jeroen Van Paridon and Bill Thompson. 2021. subs2vec: Word embeddings from subtitles in 55 languages. *Behavior research methods*, 53(2):629–655.

Yuxiang Wei. 2022. Entropy as a measurement of cognitive load in translation. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Workshop 1: Empirical Translation Process Research)*, pages 75–86.

Kayo Yin, Patrick Fernandes, André FT Martins, and Graham Neubig. 2021. When Does Translation Require Context? A Data-driven, Multilingual Exploration. *arXiv e-prints*, pages arXiv–2109.

A Translation behavioural data

We evaluate translation difficulty in context using CRITT TPR-DB, which includes logs for translations of the multiLing corpus (six English source texts) into various languages (Carl et al., 2016b).[13] Here we briefly describe all features relevant to translation difficulty.

HTra is similar to H^c_{al} in that these methods quantify the degree of uncertainty in a lexical distribution. Where H^c_{al} measures the entropy of word alignments, HTra does the same for source and target tokens in multiLing translations (Schaeffer et al., 2016). Words with high HTra have less obvious translation choices, which means that the lexical decisions of the translator require more cognitive effort. This measure has been shown to affect total target production duration, First Fixation Duration and Source Token Reading Time (Carl and Schaeffer, 2017; Schaeffer and Carl, 2017; Schaeffer et al., 2016).

Munit refers to the number of micro translation units, which are units of translation activity separated by pauses of a given length, as monitored by a key logger or an eye tracker (Alves and Vale, 2017). This records the number of activities involved in the translation process, where the translator might read, plan, revise, edit or reconsider a previously translated token.

Dur refers to the production duration of a target token given a source token, i.e., the time taken from the first keystroke to last keystroke in producing the relevant token.

Following Heilmann and Llorca-Bofí (2021) and Carl (2021b), we remove all values of Dur smaller than 20ms and log scale all remaining values. Across participants and translation sessions, HTra is averaged by source words, whereas Munit and Dur are averaged by translation pairs.

B Experiment and data specification

The pre-processing steps before word alignment include white space cleaning and removal of any sentence pairs containing non-ASCII-decodable characters. After word alignment, we exclude entropy values of words that have been aligned fewer than 20 times, or have frequency lower than 50 in Worldlex (Gimenes and New, 2016).[14]

[13] https://sites.google.com/site/centretranslationinnovation/tpr-db

[14] http://www.lexique.org/?page_id=250

Measure		de	es	ja	ms	nl	zh
I^c_{al}	→ en	7.5M	11.8M	1.3M	0.9M	9.0M	5.6M
	en →	7.5M	11.8M	1.3M	0.9M	9.0M	5.6M
I^w_{al}	→ en	7.5M	11.8M	1.3M	0.9M	9.0M	5.6M
	en →	7.5M	11.8M	1.3M	0.9M	9.0M	5.6M
H^c_{al}	→ en	41.0K	44.5K	13.5K	10.7K	33.1K	24.7K
	en →	34.6K	38.6K	15.6K	13.8K	37.0K	30.7K
H^w_{al}	→ en	41.0K	44.5K	13.5K	10.7K	33.1K	24.7K
	en →	34.6K	38.6K	15.6K	13.8K	37.0K	30.7K
M_{emb}	→ en	1,973	2,779	2,134	-	1,586	1,834
	en →	1,209	1,883	1,131	-	1,388	1,241
S_{emb}	↔ en	3,011	4,972	4,209	-	1,911	2,004
NoTrans	→ en	-	762	193	1,004	550	-
	en →	-	670	193	844	562	544
Semsim	↔ en	-	-	193	-	1,003	1,282
HTra	en →	415	416	415	-	-	-
Munit	en →	4,419	4,897	12.0K	-	-	-
Dur	en →	4,087	4,240	6,085	-	-	-

Table 5: Vocabulary size and number of paired words for each measure and evaluation data set. NoTrans and Semsim refer to number of translations and human semantic similarity ratings of translation norms respectively. Measures marked with ↔ en are non-directional (except for Dutch semantic similarity ratings, which include ratings in both directions of the same translation pairs).

Table		de	es	ja	ms	nl	zh
1	→ en	-	751	162	713	534	-
	en →	-	670	187	738	559	540
2	→ en	-	-	184	-	988	1,175
	en →	-	-	184	-	988	1,175
3	HTra	366	376	246	-	-	-
	Dur	1,330	1,584	809	-	-	-
	Munit	1,400	1,697	1,334	-	-	-

Table 6: The number of comparisons across measures for each result table in the main text.

Table 5 reports the vocabulary size and number of paired words in each measure and evaluation data set. During evaluation, we also limit our comparisons across methods to the same set of vocabulary and translation pairs. The number of comparisons for all result tables in the main text is summarized in Table 6.

C Additional results for Thompson's embedding-based approach

We found the embedding-based method of Thompson et al. (2020) to be highly sensitive to the quality of the input translation pairs — performance degrades with additional word alignment data. Here, we provide results for two alternative measures. M^{+}_{emb} and S^{+}_{emb} are comparable to M_{emb} and S_{emb} in the main text, but incorporate the top 3 word alignments for each word in the initial vocabulary. Another set of measures are M^{s}_{emb} and S^{s}_{emb}, which are based on the same translation pairs as M_{emb}/S_{emb}, but are computed with OpenSubtitles embeddings (`subs2vec`) (Van Paridon and Thompson, 2021).[15]

Tables 7a and 7b show the results against context-free translations, which correspond to Tables 1 and 2 in the main text. For context-dependent translations, the correlations with translation process features are reported in Table 8. Note that some values are missing from the tables, because `subs2vec` embeddings are not available in Japanese and Chinese.

D Terms for use

For all relevant data, models and code used in the work, we list licenses permitting research use:

- `awesome-align` under BSD 3-Clause License
- spaCy tokenizer and `subs2vec` under MIT License
- Aksara tokenizer under GNU Affero General Public License
- CRITT TPR-DB under CC BY-NC-SA License
- fastText embeddings CC BY-SA 3.0 License
- NorthEuraLex translations under CC BY-SA 4.0 License

We use the code of Thompson et al. (2020) from `https://osf.io/tngba/`, which can be freely used for academic research.[16] Lison and Tiedemann (2016) explicitly made OpenSubtitles corpora "freely available to the research community", whereas translation norms have been created to facilitate multilingual research (Tokowicz et al., 2002; Prior et al., 2007; Lee et al., 2022). The code repository for this project, as referenced from footnote 1, is available under MIT License.

		es	ja	nl	zh
→ **en**	M^{+}_{emb}	.215	/	/	-
	M^{s}_{emb}	.351	-	.212	-
en →	M^{+}_{emb}	.317	.263	.190	.151
	M^{s}_{emb}	.394	-	.335	-

(a) Number of translations

		ja	nl	zh
→ **en**	S^{+}_{emb}	-.316	-.325	-.360
	S^{s}_{emb}	-	-.310	-
en →	S^{+}_{emb}	-.316	-.295	-.360
	S^{s}_{emb}	-	-.302	-

(b) Semantic similarity ratings

Table 7: Alternative results of M_{emb} and S_{emb} using additional word alignments, M^{+}_{emb} / S^{+}_{emb}, and `subs2vec` embeddings, M^{s}_{emb} /S^{s}_{emb}, against context-free translation norms ($p < .001$). The sub-tables correspond to Tables 1 and 2 in the main text.

		de	es	ja
HTra↑	M^{+}_{emb}	.332	.314	.254
	M^{s}_{emb}	.276	.296	-
Dur (ms) ↑	S^{+}_{emb}	-.339	-.401	/
	S^{s}_{emb}	-.352	-.497	-
Munit ↑	S^{+}_{emb}	.110	.075	/
	S^{s}_{emb}	.064	/	-

Table 8: Alternative results ($p < .05$) corresponding to Table 3 in main text.

[15] `https://github.com/jvparidon/subs2vec`

[16] `https://www.nature.com/nature-portfolio/editorial-policies/self-archiving-and-license-to-publish#terms-for-use`

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
6

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3

☑ B1. Did you cite the creators of artifacts you used?
3

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix D

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
1,3, Appendix D

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix B

C ☒ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.