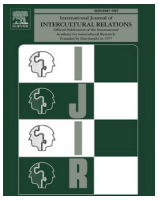
International Journal of Intercultural Relations 97 (2023) 101888

Contents lists available at ScienceDirect

International Journal of Intercultural Relations

journal homepage: www.elsevier.com/locate/ijintrel



Utilizing AI questionnaire translations in cross-cultural and intercultural research: Insights and recommendations

Jonas R. Kunst [a,*], Kinga Bierwiaczonek [a,b]

[a] Department of Psychology, University of Oslo, Norway
[b] Instituto Universitário de Lisboa (ISCTE-IUL), Centro de Investigação e Intervenção Social, Portugal

ARTICLE INFO

Keywords:
AI
Questionnaire
Survey
Translation
Machine Learning
GPT

ABSTRACT

In this research paper, we investigated the viability of AI-supported translations of survey materials in intercultural and cross-cultural research, comparing the quality of machine translations to traditional human translations. Focusing on the HEXACO personality inventory, we translated the original English inventory using Google Translate and GPT-3.5 into 33 languages for which validated human translations exist. The statistical similarity between human- and machine-generated translations varied considerably between the target languages. It was highest for target languages from the same language family as the source language, arguably because this relatedness allowed for more direct machine translations. Consistent with this reasoning, the genetic similarity between languages largely explained the differences observed. GPT's temperature setting determining how stringently or freely a text is translated had little influence on the similarity estimates, but very high levels tended to produce somewhat lower statistical similarity. To validate the quality of the machine translations, a group of social scientists rated the translation in a language for which the human and machine translations statistically converged strongly. Although the human translation was rated as being of higher quality than four out of five machine translations, these differences were relatively small. Crucially, the social scientists did not rate the human translation as significantly better than the GPT 3.5 translation with the lowest temperature setting. Based on these insights, we propose a framework outlining four recommendations for utilizing AI-supported translation in cross-cultural and intercultural research, involving AI to varying degrees in the forward-back translation process.

Social scientists seek to uncover the intricacies that underlie human behavior, interactions, and societal structures. A vital aspect of this endeavor is the comparative analysis of specific phenomena across different cultures. By juxtaposing diverse cultural backgrounds and settings, researchers often strive to determine their findings' universality or cultural specificity (Van De Vijver & Poortinga, 1982), enriching our understanding of the human experience. Crucial to this comparative approach is the development of valid and accurate cross-cultural translations of survey materials, as any discrepancies or inaccuracies in translations could lead to flawed conclusions and misleading comparisons.

Traditionally, forward-back translation (Brislin, 1970) by researchers or specialized professional translators has been considered the gold standard in ensuring the accuracy of cross-cultural translations. This method involves translating survey materials from the source language to the target language by one individual or group and then translating them back to the source language by a different

* Correspondence to: Forskningsveien 3A, 0373 Oslo, Norway.
E-mail address: j.r.kunst@psykologi.uio.no (J.R. Kunst).

https://doi.org/10.1016/j.ijintrel.2023.101888
Received 5 April 2023; Received in revised form 4 September 2023; Accepted 6 September 2023
Available online 11 September 2023

individual or group unaware of the source text. Discrepancies are detected by comparing the source and back-translated texts, and adjustments are made to the translation accordingly. While commonly used, this method has its limitations. First, it is time-intensive and can involve high costs if professional translators are involved. Moreover, the quality of translations depends considerably on the motivation, attention to detail, training, and linguistic skills of the individuals involved in the process (Ozolins et al., 2020).

With recent advancements in artificial intelligence (AI), there is an apparent potential in harnessing AI's power to revolutionize the cross-cultural translation process. Pioneering technologies such as Google Translate and the Generative Pre-trained Transformer (GPT) series have made high-quality translations readily accessible to a vast audience, sparking a debate on the potential of AI-supported translations as a more time- and resource-efficient alternative to traditional methods of survey translation (Groves & Mundt, 2015; Lear et al., 2016). However, the pressing question is whether AI-assisted translation of research materials currently reaches the quality of validated human translations. Does its quality differ between languages? What do insights into these questions tell us about where

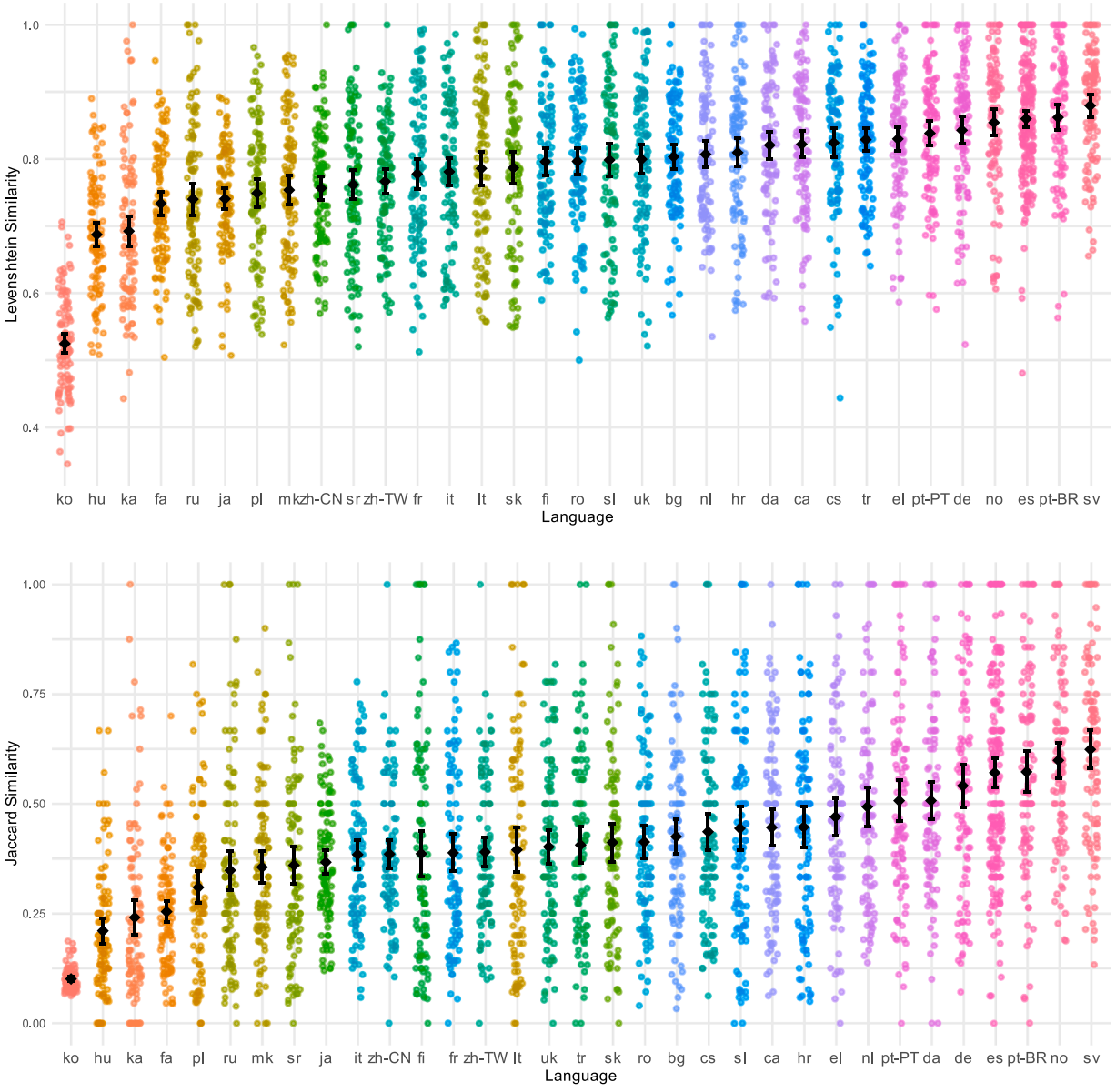


Fig. 1. Similarity Between Google Translations and Human Translation. Note. Black point estimates represent mean values and error bars represent 95% confidence intervals. Colored points represent data points. bg = Bulgarian, ca = Catalan, zh-CN = simplified Chinese, zh-TW = traditional Chinese, hr = Croatian, cs = Czech, da = Danish, nl = Dutch, fi = Finnish, fr = French, ka = Georgian, de = German, el = Greek, hu = Hungarian, it = Italian, ja = Japanese, ko = Korean, lt = Lithuanian, mk = Macedonian, no = Norwegian, fa = Persian, pl = Polish, pt-BR = Brazilian Portuguese, pt-PT = European Portuguese, ro = Romanian, ru = Russian, sr = Serbian, sk = Slovak, sl = Slovenian, es = Spanish, sv = Swedish, tr = Turkish, uk = Ukrainian.

scientific translation methods in the social sciences are headed? Considering AI's transformative potential, the present research aimed to provide empirical insights into the viability of machine-supported translations for survey materials in intercultural and cross-cultural research, making a set of general recommendations.

Our empirical investigation focused on translations of the HEXACO personality inventory (Lee & Ashton, 2004) as it is one of the most cross-culturally used psychometric instruments with validated translations in various languages. In the first step, we used Google Translate and GPT-3.5 to translate the original 100-item version of the inventory from English into the 33 languages for which validated human translations are available (Lee & Ashton, 2023). For the GPT translations, we then systematically altered the temperature parameter. This parameter controls the randomness of the model's output, with higher values increasing diversity and creativity at the cost of coherence, while lower values promote more focused and deterministic responses. To the extent that similarity between the original, validated human translation and the machine-generated translations can be considered a proxy indicator of high-quality translations, we then used two standard estimates of similarity between texts (i.e., the Jaccard and the Levenshtein indices of similarity; Jaccard, 1912; Levenshtein, 1966) to calculate similarity scores between the human translations and machine translations. Further, we tested whether scores on these similarity indices may be explained by the "genetic similarity" between languages (Beaufils & Tomin, 2023), arguably because more similar languages allow for a more direct machine translation. We also tested whether the percentage of online content in a given language (W3 Techs, 2023) may explain the performance of the machine translations, given that more content may have allowed for better training of the model. Finally, we validate the previous results by letting a group of social scientists rate the quality of the human and machine translations.

Methods

The original English version and 33 official translations of the 100-item HEXACO inventory (Lee & Ashton, 2023) were downloaded and organized in one file (all materials, data, and code are openly available at https://osf.io/mxrfa/?view_only=c375de8d658040ec9b8aa597137a42b0). The 33 different languages from various parts of the world can be seen on the x-axis of Fig. 1 (please note that Mandarin Chinese was represented separately in its traditional and simplified versions and that two official Spanish translations were available and, hence, analyzed combined).

Next, we translated the original English inventory version into each of these 33 target languages using Google Translate and GPT. Google Translate is a widely used translation service developed by Google, which supports over 100 languages. Initially based on statistical machine translation (SMT), Google Translate has evolved to incorporate neural machine translation (NMT; Bahdanau et al., 2014) to improve translation quality. Its NMT system leverages a deep learning architecture called the sequence-to-sequence (seq2seq; Google, 2023) model. This model consists of an encoder-decoder framework, where the encoder processes the input text and generates a fixed-size vector representation, which the decoder then uses to generate the translated text. In addition, Google Translate is trained on vast amounts of parallel text data collected from various multilingual sources such as websites, books, and other translated documents. This data helps the NMT system learn and generalize patterns across different languages, enabling it to produce more accurate translations.

GPT is a language model developed by OpenAI. Based on the Transformer architecture (Radford et al., 2018; Vaswani et al., 2017), GPT has been trained on vast amounts of data and is capable of generating human-like text. Whereas it was not specifically designed for translation, GPT can be fine-tuned for translation tasks. It utilizes an attention-based mechanism, which aims at improving performance by flexibly utilizing the most relevant parts of the input text. This attention mechanism helps GPT understand and capture the context of a given text, making it more effective in producing accurate translations. In addition to its attention mechanism, GPT benefits from a vast pre-training process that exposes it to diverse language structures and idiomatic expressions. This extensive training allows GPT to understand language intricacies deeply, making it adept at handling complex translation challenges. Moreover, GPT's self-supervised learning approach enables it to refine its translation capabilities based on the input-output pairs provided during fine-tuning. This process allows GPT to continuously learn and adapt, leading to more accurate and reliable translations.

All text was translated once using Google Translate on March 30, 2023, and five times using the GPT 3.5 turbo model (i.e., the most recent model widely available via API at the time of the study) with different temperature parameters (i.e., 0 [the lowest possible value], 0.25, 0.50, 0.75, 1.00 [the highest possible value]) on March 30 and 31, 2023. For GPT, we used the following instructions suggested by the chatbot itself: "You are a helpful assistant that translates English text to different languages. Translate the following English text to [target language]." Every item was translated through separate API calls.

The resulting translated text was then analyzed using two standard indices of linguistic similarity and with the human ratings of a sample of specialists (described below). We also enriched the data with two potential sources of this similarity: the genetic similarity between the source and target languages (Greenberg, 2005) and the percentage of online content in the target languages as a proxy of the size of the language corpora available for model training. Before the statistical analyses, the non- or semi-alphabetic input texts (i.e., traditional Chinese, Simplified Chinese, Japanese, and Korean) were tokenized using the jiebaR (Wenfeng & Yanyi, 2019) and stringi (Gagolewski, 2022) packages to match the required input format for the analyses. We also cleaned the data before analyzing them. For instance, in some cases, the GPT chatbot added the explanation "Translated into [language]" to the translation. In others, it added moral disclaimers, such as "Please keep in mind that using counterfeit money is illegal and not recommended. This translation is for informational purposes only" to the translation of the item "I'd be tempted to use counterfeit money, if I were sure I could get away with it." Whenever two translations were provided, we kept the first of them.

Levenshtein estimate of similarity

The Levenshtein estimate (1966) quantifies the similarity between two strings (sequences of characters) by calculating the minimum number of single-character edits required to transform one string into the other. These edits can be insertions, deletions, or substitutions of individual characters. The Levenshtein distance is based on the principle of determining the most efficient way to transform one string into another, where the smallest number of edits measures the efficiency.

The Levenshtein distance between two strings A and B can be computed using a dynamic programming algorithm with the following recursive formula:

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j) if \quad \min(i,j) = 0, \\ \min \begin{cases} lev_{a,b}(i-1,j) + 1 \\ lev_{a,b} \quad (i,j-1) + 1 \\ lev_{a,b}(i-1,j-1) + 1_{(a_i \quad \neq b_j)} \end{cases} \quad otherwise. \end{cases}$$

where lev(a, b) is the Levenshtein distance between the first i characters of string a and the first j characters of string b; a_i and b_j represent the i-th and j -th characters of strings a and b, respectively. To normalize the Levenshtein distance, it is divided by the maximum possible distance (i.e., the length of the longer string). In the present paper, we calculated this metric using the stringdist R library (van der Loo, 2014).

Jaccard estimation of similarity

The Jaccard (1912) similarity coefficient is a statistic used to quantify the similarity between two sets. In text analysis, these sets can represent two documents' words, phrases, or other textual elements. The Jaccard similarity coefficient is based on the principle of comparing the intersection and the union of the two sets in question. Thereby, the Jaccard similarity coefficient formula is:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

J(A, B) represents the Jaccard similarity coefficient between sets A and B. Next, $|A \cap B|$ is the size of the intersection of sets A and B (i.e., the number of elements common to both sets), $|A \cup B|$ is the size of the union of sets A and B (i.e., the total number of unique elements in both sets). By dividing the number of words in common by the total number of unique items, we get a value that helps us understand how similar the two texts are. The higher the value, the more similar the texts. The resulting Jaccard similarity coefficient ranges from 0 to 1. A value of 0 indicates that the two sets have no common elements, implying that the texts are entirely dissimilar. A value of 1, on the other hand, indicates that the two sets are identical.

Genetic similarities between languages

The genetic similarity between languages (Greenberg, 2005) can be calculated using a computerized model based on linguistics and mathematics, similar to methodologies used in biology. The process involves encoding lexical material in a format that a computer can process. A set of rules is determined to identify cognates based on relationships and sound change. The scoring system ranges from 0 to 100, with 100 being the highest possible score (same language). The model also calculates the statistical context for all results, addressing the issue of chance exposure in language evolution analysis. Over 160 languages can currently be compared using this method, with the system performing millions of comparison tasks within seconds. For the present study, we obtained scores from Beaufils and Tomin (2023), who adapted the MEGA7 molecular evolutionary genetics model (Kumar et al., 2016) for their analysis.

Percentage of websites corpus in languages

We obtained the estimated percentage of all websites in the different target languages from the W3 Techs (2023) Web Technology Surveys. W3 Techs determines the percentage of web content in a specific language by analyzing web pages like a search engine. It fetches web pages and uses specific patterns to identify the languages used. Websites are visited approximately once a month, and deeper crawling is performed on samples. The reports are updated daily, with more extensive market reports generated monthly.

Human ratings

Finally, we selected one of the languages that converged the most with the human translations (Norwegian) to be rated by humans. A repeated-measurements G*POWER (Faul et al., 2009) analysis indicated that 10 participants would provide 90% power to detect a small effect size ($f = 0.25$) provided 50 measurements in five groups (five translations; human, Google, three GPT versions) and an $r = .80$ between measures. This power analysis and procedure were pre-registered at https://osf.io/23xw9/?view_only=eb6f06420f274403b428ce5a06508536.

We recruited 11 experts (6 identifying as women, 5 as men), all social scientists with an average of 7.82 years of experience. On a scale from 0 (*not experienced at all*) to 6 (*very experienced*), participants rated their experience in survey translation at 4.73 on average

(*SD* = 1.74). Based on the European classification system for language competency, all participants had C2 competency in Norwegian, equal to being a native language speaker. In terms of English competency, six had C1 and five C2. Participants read the following instruction:

"In research, the accurate translation of scales is important to allow for valid comparisons between countries. Next, we will present you with items from the HEXACO personality inventory and different translations into Norwegian. We would like you always to judge the quality of each translation. When evaluating the quality, please pay attention to the following:

1. Accuracy: Ensure that the translated content maintains the original meaning and intent of the source questionnaire without distortion, omission, or embellishment. This is crucial in obtaining consistent and comparable responses across languages.
2. Cultural Relevance: Check whether the translation has been adapted to the target audience's cultural context. This may involve adjusting idioms, metaphors, and examples to avoid confusion, misunderstanding, or offense.

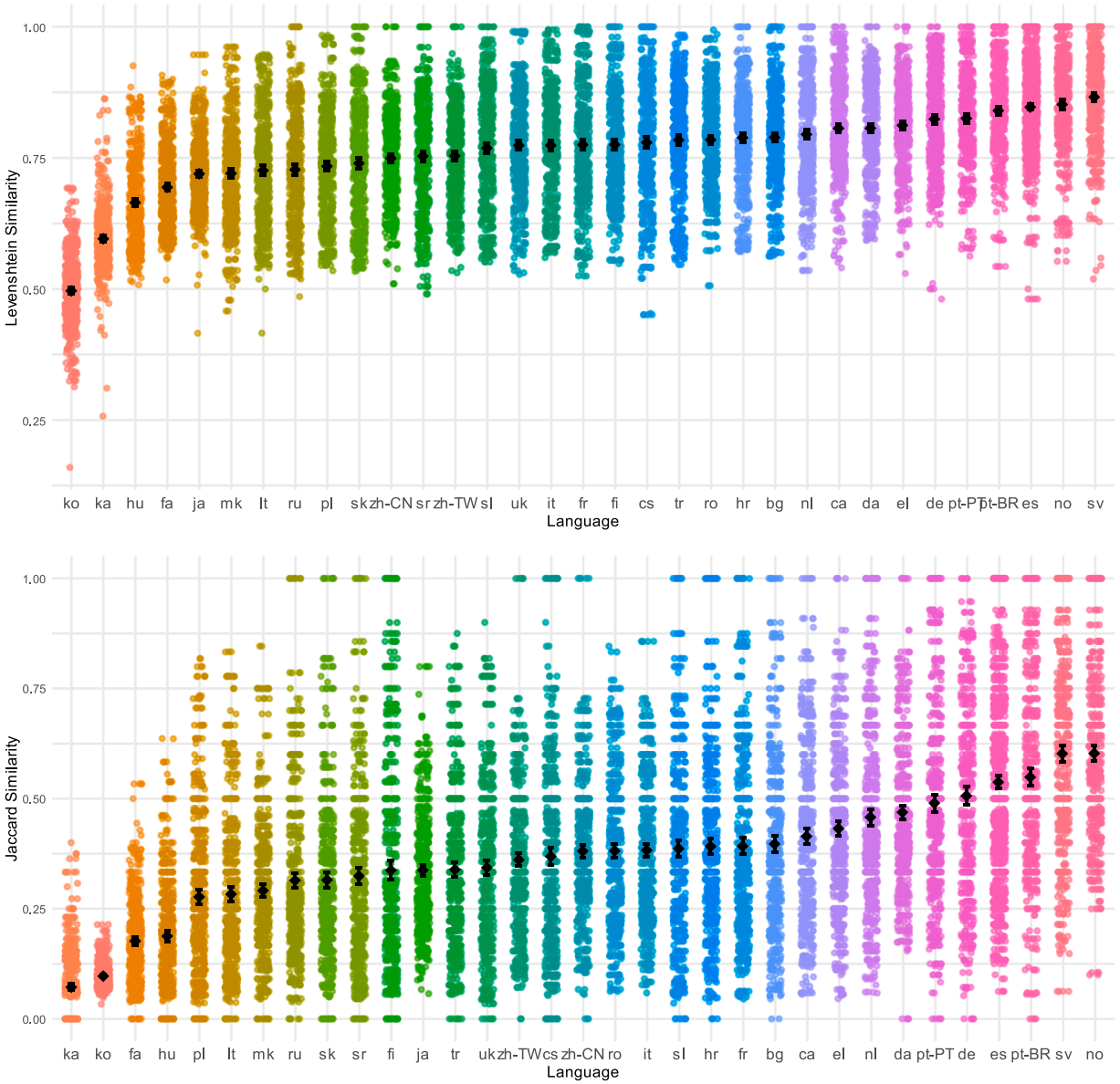


Fig. 2. Similarity Between GPT 3.5 Translations and Human Translation Across All Temperature Settings. Note. Black point estimates represent mean values and error bars represent 95% confidence intervals. Colored points represent data points. bg = Bulgarian, ca = Catalan, zh-CN = simplified Chinese, zh-TW = traditional Chinese, hr = Croatian, cs = Czech, da = Danish, nl = Dutch, fi = Finnish, fr = French, ka = Georgian, de = German, el = Greek, hu = Hungarian, it = Italian, ja = Japanese, ko = Korean, lt = Lithuanian, mk = Macedonian, no = Norwegian, fa = Persian, pl = Polish, pt-BR = Brazilian Portuguese, pt-PT = European Portuguese, ro = Romanian, ru = Russian, sr = Serbian, sk = Slovak, sl = Slovenian, es = Spanish, sv = Swedish, tr = Turkish, uk = Ukrainian.

3. Clarity and Readability: The translated statements should be clear, concise, and easy to understand. The language used should be appropriate for the intended respondents, avoiding jargon, overly technical terms, or ambiguous phrasing.

Please note that the translations may be the same or similar. It is still important that you rate each of them."

The human raters then always saw one of the original English HEXACO items on top of the screen and were asked to rate the quality of five translations presented under it: (1) the human translation, (2) the Google translation, and (3−5) three GTP 3.5 translations (all five translations presented in random order). For the GPT translations, we selected the version with 0, 0.5, and 1 temperatures to allow for the evaluation of the full range of temperature parameters while limiting participant fatigue. For each translated item, the participants rated the quality from 1 (*very low quality*) to 7 (*very high quality*). In total, each participant rated the translations of 50 items randomly selected from the HEXACO. They were blind to which type of translation was presented to them. This procedure resulted in a sample size of 550 ratings for each of the five translations (2750 data points in total). As pre-registered, we applied Holm correction in data analyses to prevent the inflation of Type 1 errors due to multiple comparisons.

Results

Differences in statistical similarity between machine learning approaches

The average on the Levenshtein similarity index was high across the different translation types (see Fig. 1). However, in multi-level models using the lme4 (Bates, 2010; Douglas et al., 2015), lmerTest (Kuznetsova et al., 2016), and emmeans (Lenth et al., 2019) packages in R controlling for the target language, the Google translation showed a significantly larger similarity with the human translation ($M = 0.785$, 95% CI [0.761, 0.810]) than the GPT translations at each temperature level (0.00: $M = 0.766$, 95% CI [0.741, 0.791]; 0.25: $M = 0.765$, 95% CI [0.741, 0.790]; 0.50: $M = 0.765$, 95% CI [0.740, 0.789]; 0.75: $M = 0.763$, 95% CI [0.739, 0.788]; 1.00: $M = 0.760$, 95% CI [0.735, 0.784]; all $ps < .001$). Nevertheless, it is essential to note that these differences were minor, $|ds_{rm}| < 0.15$ (as calculated with the EMAtools R package, v. 0.1.4; Kleiman, 2021). The GPT temperature groups did not differ significantly from each other. However, the highest (i.e., 1.00) temperature group trended lower on the Levenshtein similarity index than the lowest (i.e., 0.00) temperature group, $p = .075$.

The same pattern was observed for the Jaccard similarity index. The Google translation showed a significantly larger similarity with the human translation ($M = 0.411$, 95% CI [0.369, 0.453]) than the GPT translations at each temperature level (0.00: $M = 0.375$, 95% CI [0.333, 0.417]; 0.25: $M = 0.373$, 95% CI [0.331, 0.415]; 0.50: $M = 0.372$, 95% CI [0.330, 0.414]; 0.75: $M = 0.369$, 95% CI [0.327, 0.411]; 1.00: $M = 0.362$, 95% CI [0.320, 0.404]; all $ps < .001$). Still, these differences were again small, $|ds_{rm}| < 0.15$. The GPT temperature groups did not differ significantly from each other, but the highest temperature group again trended lower than the lowest temperature group, $p = .067$.

Treating the temperature variable as a continuous variable, a negative effect was observed of temperature on both the Levenshtein, $B = -0.006$, $SE = 0.002$, $t(16970) = -2.84$, $p = .005$, $d_{rm} = -0.04$, and Jacard similarity indices, $B = -0.012$, $SE = 0.004$, $t(16970) = -2.88$, $p = .004$, $d_{rm} = -0.04$, in corresponding multi-level models. Thus, the similarity decreased as the temperate parameter increased, but these effects were minimal.

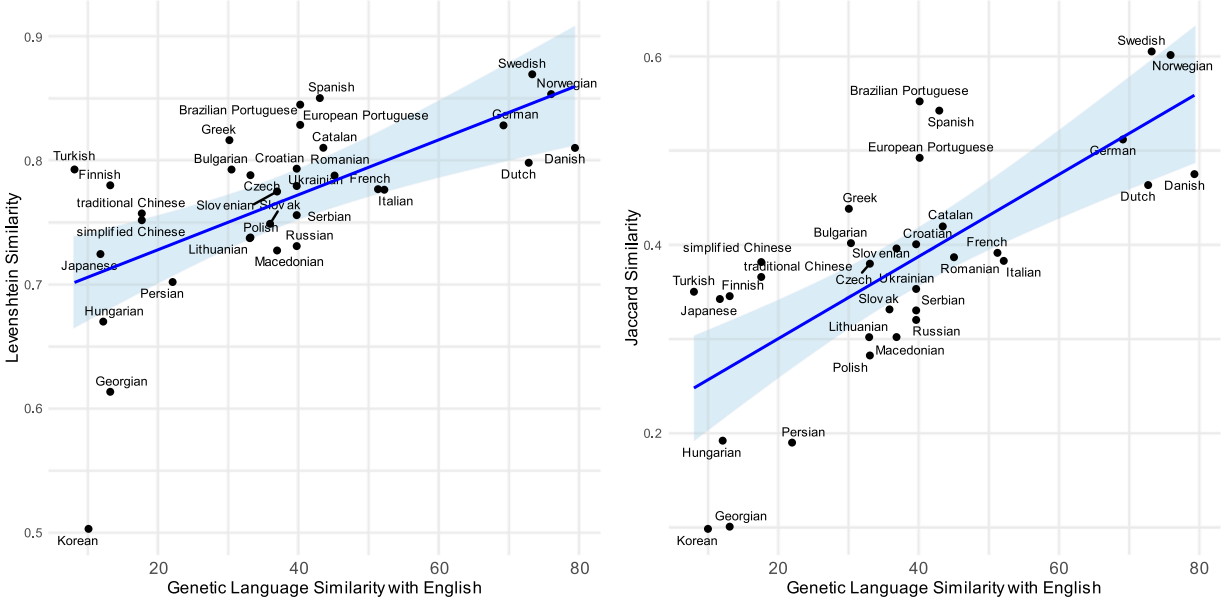


Fig. 3. The Role of Genetic Similarity for Scores on the Similarity Indices at the Language Level. Note. Ribbons represent 95% confidence intervals.

Differences in Statistical Similarity Between Languages

Next, we investigated differences in similarity scores between languages. As displayed in Figs. 1 and 2 (created with the ggplot2 and tidyverse R packages; Wickham et al., 2019; Wickham et al., 2016), these differences were substantial. For example, machine-based translations into European languages such as Swedish, Norwegian, Spanish, Portuguese, and German were the most similar to the human translations, whereas the similarity of translations into languages such as Korean, Georgian, or Hungarian was the lowest.

To understand this variation better, we tested the influence of the language-level variables of genetic similarity (rescaled to a 0–1 scale to avoid inflated variances) and language corpus size in multi-level models (controlling for nesting in languages and in types of translations, i.e., Google and the five GPT different temperature variants) with the Jaccard and Levenshtein similarities as dependent variables. The genetic similarity between English (the source language of the HEXACO) and the target language strongly and positively predicted the Levenshtein, $B = 0.157$, $SE = 0.037$, $t(30) = 4.29$, $p < .001$, $d_{rm} = 1.56$, and Jaccard similarities, $B = 0.309$, $SE = 0.055$,

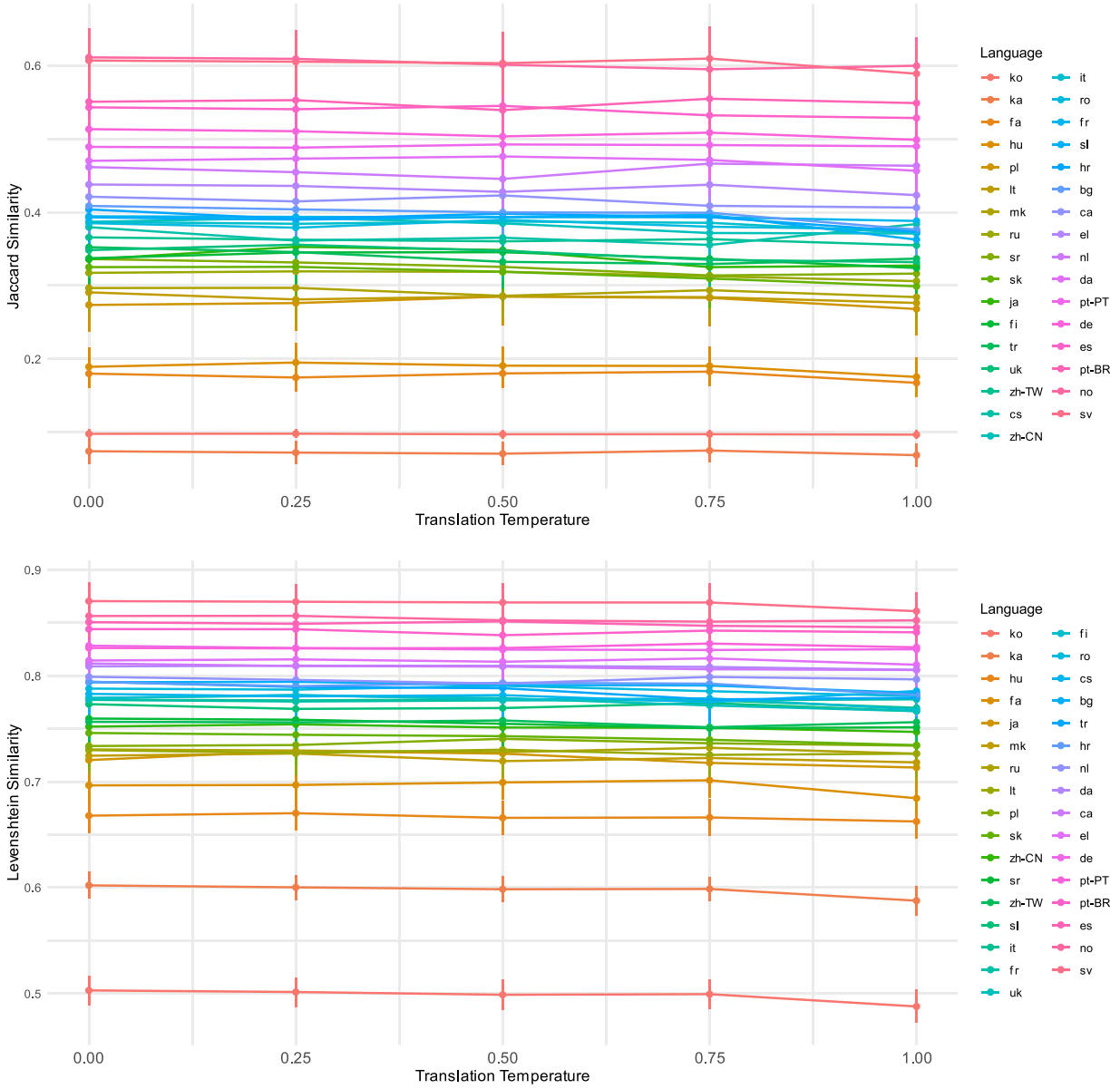


Fig. 4. The Influence of GPT 3.5 Temperature Parameter on Similarity. Note. Point estimates represent mean values and error bars 95% confidence intervals. bg = Bulgarian, ca = Catalan, zh-CN = simplified Chinese, zh-TW = traditional Chinese, hr = Croatian, cs = Czech, da = Danish, nl = Dutch, fi = Finnish, fr = French, ka = Georgian, de = German, el = Greek, hu = Hungarian, it = Italian, ja = Japanese, ko = Korean, lt = Lithuanian, mk = Macedonian, no = Norwegian, fa = Persian, pl = Polish, pt-BR = Brazilian Portuguese, pt-PT = European Portuguese, ro = Romanian, ru = Russian, sr = Serbian, sk = Slovak, sl = Slovenian, es = Spanish, sv = Swedish, tr = Turkish, uk = Ukrainian.

$t(30) = 5.66$, $p < .001$, $d_{rm} = 2.06$. For none of the similarity indices, the size of the language corpus had an effect, $ps > .225$. To visualize these differences, we aggregated the data to the language level. Here, the genetic similarity was strongly and positively correlated with the Levenshtein, $r(31) = .616$, $p < .001$, and Jaccard similarity sores, $r(31) = .714$, $p < .001$, see Fig. 3. The language corpus variable was unrelated to the similarity scores at the aggregate level, $ps > .290$, mirroring the multi-level results.

We also tested whether the temperature settings for the GPT translations would influence the similarity scores. However, as visualized in Fig. 4, the similarity scores were generally very similar in each language. Yet, a slight reduction may be observed in some languages at the highest temperature level.

Human Quality Ratings

Finally, we tested whether the sample of social scientists would rate the quality of human and machine translations differently. In a multi-level model controlling for the clustering of responses within participants and items, the human translation was rated as being of higher quality than the Google translation, $B = 0.644$, $SE = 0.191$, $t(44.8) = 3.37$, $p = .014$, $d_{rm} = 0.57$, and the GPT 3.5 translations with medium (0.50) temperature, $B = 0.569$, $SE = 0.175$, $t(30.4) = 3.26$, $p = .022$, $d_{rm} = 0.51$, and high (1.00) temperature, $B = 0.591$, $SE = 0.171$, $t(36.2) = 3.46$, $p = .014$, $d_{rm} = 0.53$, see Fig. 5. However, the human translation did not significantly differ from the GPT 3.5 translation with the lowest (0.00) temperature, $B = 0.431$, $SE = 0.157$, $t(27.0) = 2.75$, $p = .073$, $d_{rm} = 0.38$, although this small difference trended in favor of the human translation. The GPT translations did not differ significantly, $ps > .790$.

Discussion

Social scientists interested in comparative research across cultures and contexts regularly face the challenge of accurately translating surveys into a set of target languages. As this process is time and resource-intensive, utilizing AI in the translation process may be a viable option. However, the use of AI for this purpose has also been met with resistance due to its potential linguistic inaccuracy (Groves & Mundt, 2015; Lear et al., 2016). Alleviating some of these concerns, our results demonstrate that AI-generated translations can already today achieve a quality approaching that of human translations. Moreover, as the existing language models rapidly evolve and new versions are released frequently, the small quality gap observed in the present study will likely narrow, if not disappear, in the future. However, at least for now, researchers should be aware of several factors that influence the quality of the translations.

The statistical convergence between human and machine translations seems to depend strongly on the genetic similarities between the source language (in this case, English) and the target language. In the present study, machine translations from English into other European languages showed the highest statistical convergence with human translations. Arguably, such genetic similarity in languages' vocabulary, grammatical structure, and cultural content allows for more direct machine translations, leading to higher convergence with human translations. Thus, from a practical perspective, more involvement of humans in the translation process will likely be necessary for target languages with less similarity to the source language. However, as language models evolve, the importance of language similarity will likely decrease.

The human quality ratings suggested that human translations still have an advantage over machine translations, although this advantage was relatively small, depending on the translation method. Here, the temperature settings of the GPT 3.5 model seemed to

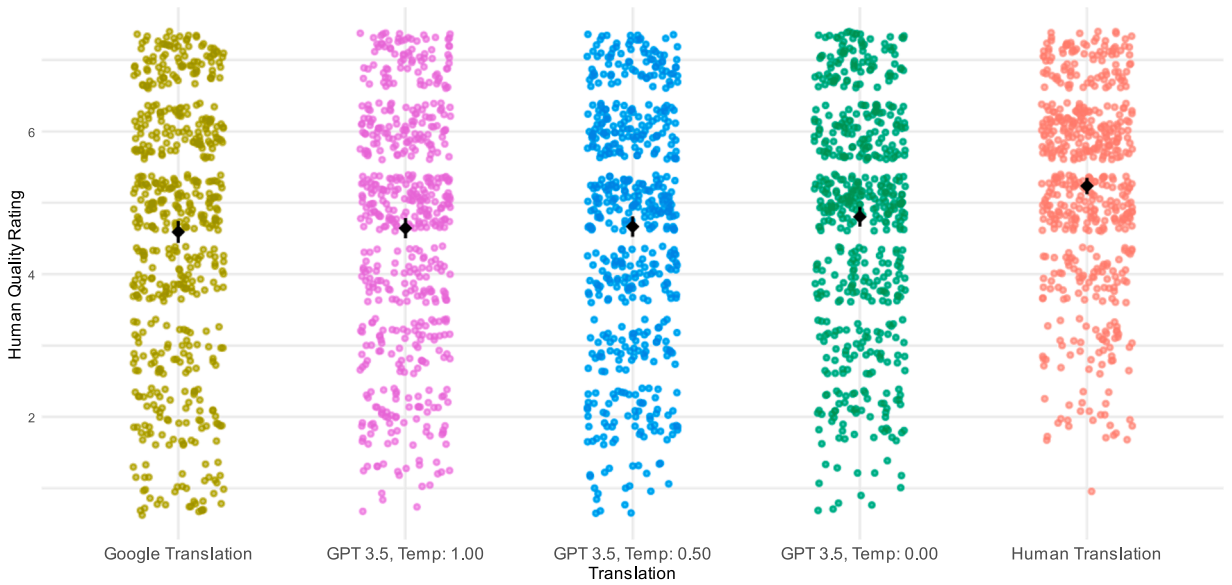


Fig. 5. The Human Quality Ratings of the Different Translations. Note. Black point estimates represent mean values and error bars 95% confidence intervals. Colored points represent data points.

matter. When the model was instructed to translate text from English into the target language stringently with little degrees of freedom (i.e., as reflected in the temperature setting of 0), the difference between quality ratings of the human and machine translations did not reach significance. This finding suggests that low temperature settings can be advantageous in machine-supported survey translation. Still, given our focus on one language for these analyses, more research is needed to elucidate how temperature settings influence translation quality in different target languages that vary in their similarity to the source language. We aimed to give a glimpse into the *potential* of machine translation in the social sciences, and we, therefore, selected a language with high statistical convergence. However, we acknowledge the need to validate translations into other languages with human raters.

Whereas the present research provided novel insights into the viability of AI in the translation of cross-cultural research materials, it should be seen as a first step considering several limitations. We compared the machine-generated translations to the validated human translations of the HEXACO (Lee & Ashton, 2004), assuming that the latter can be considered high quality. However, translating between languages from different language families can make translations that are both direct (i.e., closely following the structure, wording, or literal meaning of the source text) and natural-sounding (i.e., conveying language fluently and idiomatically in the target language) difficult. In such cases, different translations may reproduce the original text's meaning with similar accuracy, even if they show varying degrees of similarity when compared with the human translation. We acknowledge that while the HEXACO has been

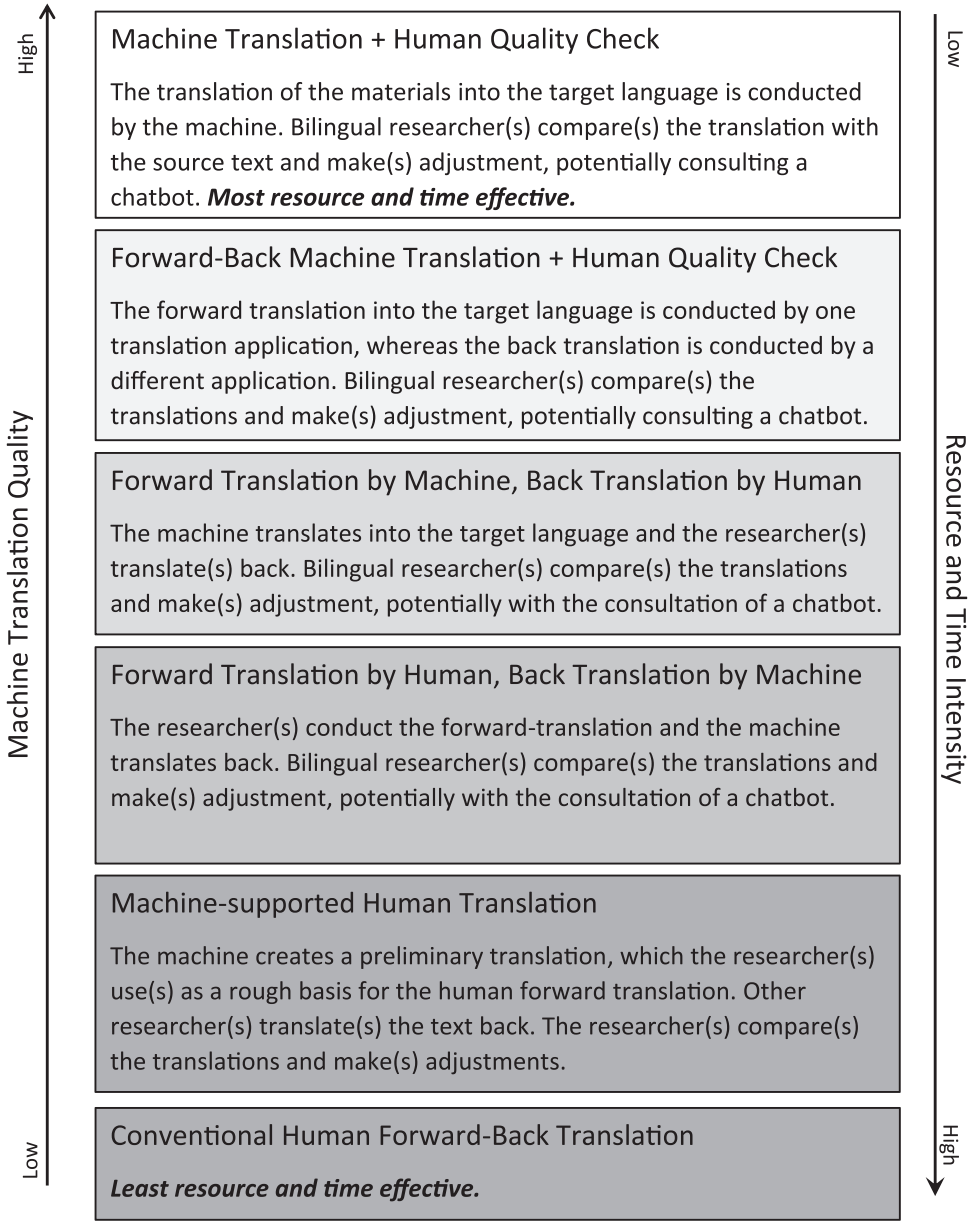


Fig. 6. Conceptual Framework for Machine-Supported Translation.

extensively researched across various cultures, it is imperative for future studies to validate our findings using other instruments that are at the core of intercultural and cross-cultural research. Suitable instruments include the social value scales (Schwartz, 1992), the tightness-looseness scale (Gelfand et al., 2011), and the relational mobility scale (Thomson et al., 2018). These tools have validated translations in numerous target languages and often achieve high levels of measurement invariance, making them suitable candidates for such investigative efforts.

Our study is not well-positioned to quantify how well the language models by Google or OpenAI intrinsically translate text into different languages. We found the genetic similarity between the source and target languages to predict the convergence of human and machine translations strongly. Future research is needed to establish whether model- and other language-specific factors can account for the variance unexplained by the genetic similarity of the included languages.

Recommendations

Based on the empirical insights and considering accelerating technological developments that will improve the accuracy of machine translations, we conclude with recommendations for researchers who want to utilize AI technology when translating materials into one or several languages for cross-cultural or intercultural purposes. We propose a framework that identifies translation methods based on the machine translation quality (see Fig. 6). Specifically, we suggest four alternatives to standard human forward-back translation:

1. **Machine Translation + Human Quality Check:** This method involves using machine translation to unidirectionally convert the survey materials into the target language, followed by bilingual researchers who assess the translation's accuracy and make necessary adjustments by comparing the source to the target versions. The researchers may also consult a chatbot for assistance during the process. This approach relies on the machine generating high-quality translations and is the most resource and time-effective among the options. Currently, it may be best suited for translations between source and target languages with high genetic similarity.
2. **Forward-Back Machine Translation + Human Quality Check:** In this approach, the forward translation into the target language is performed by one machine translation application, while a different application (based on a different language model) handles the back-translation. Bilingual researchers then compare and adjust the translations, potentially consulting a chatbot.
3. **Forward-Translation by Machine, Back-Translation by Human:** The machine handles the forward translation into the target language, and humans perform the back translation. Bilingual researchers compare the translations and make adjustments, potentially consulting a chatbot.
4. **Forward-Translation by Human, Back-Translation by Machine:** In this approach, humans conduct the forward translation, and the machine performs the back translation. Bilingual researchers then compare and adjust the translations, potentially consulting a chatbot. This method can be adopted when the translation quality from the machine does not reach the level needed for the previous options.

In cases where these four methods are not viable, possibly due to only medium-quality machine translations and/or low genetic similarity between the languages, we still recommend that researchers use AI during the translation process where possible to reduce the time intensity of the task. Many language models can produce at least acceptable translations into most modern languages. Thus, researchers may utilize the support of AI even if its translation quality does not reach standalone quality. For instance, the machine can generate a preliminary translation, which researchers use as a rough basis for their forward translation. Other researchers handle the back translation, potentially also supported by AI. The researchers then compare and adjust the translations.

It is important to note that the proposed framework needs to be empirically validated. Our analyses suggest that the most time- and resource-effective options might, at the moment, be most viable for target languages genetically similar to the source language. To put it another way, the best approach might not be dependent on particular languages, but rather on how closely the source and target languages resemble each other. Nonetheless, additional studies are required to confirm this notion, as certain linguistic features intrinsic to specific languages and language families might pose greater challenges for AI to emulate than others. Moreover, translations into isolated languages (i.e., languages with little to no genetic similarity with other languages) might pose challenges that need to be investigated. Therefore, we encourage researchers to empirically compare the translation results of the different methods in different languages to identify their benefits and boundary conditions. Future studies could help pinpoint specific thresholds in machine translation accuracy. This information could guide decisions about which proposed option is best suited for a particular scenario. However, as the translation accuracy of machines will only improve in the future, we are confident that an increasing number of researchers will gravitate toward the least resource- and time-intensive translation options in our framework that involve machine translations to a large extent.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used GPT 4.0 in order to optimize the writing process, with a main focus of refining the manuscript's language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv Preprint arXiv, 1409*, *0473*.

Bates, D.M. (2010). *lme4: Mixed-effects modeling with R*. http://lme4.0.r-forge.r-project.org/lMMwR/lrgprt.pdf.

Beaufils, V., & Tomin, J. (2023). eLinguistics. net. Quantifying the genetic proximity between languages. URL: http://www.elinguistics.net.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1*(3), 185–216. https://doi.org/10.1177/135910457000100301

Douglas, Bates, Martin, Maechler, Ben, Bolker, & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Gagolewski, M. (2022). stringi: Fast and portable character string processing in R. *Journal of Statistical Software, 103*(2), 1–59. https://doi.org/10.18637/jss.v103.i02

Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., Duan, L., Almaliach, A., Ang, S., Arnadottir, J., Aycan, Z., Boehnke, K., Boski, P., Cabecinhas, R., Chan, D., Chhokar, J., D'Amato, A., Ferrer, M., Fischlmayr, I. C., Fischer, R., Fülöp, M., Georgas, J., Kashima, E. S., Kashima, Y., Kim, K., Lempereur, A., Marquez, P., Othman, R., Overlaet, B., Panagiotopoulou, P., Peltzer, K., Perez-Florizno, L. R., Ponomarenko, L., Realo, A., Schei, V., Schmitt, M., Smith, P. B., Soomro, N., Szabo, E., Taveesin, N., Toyama, M., Van de Vliert, E., Vohra, N., Ward, C., & Yamaguchi, S. (2011). Differences between tight and loose cultures: A 33-nation study. *Science, 332*(6033), 1100–1104. https://doi.org/10.1126/science.1197754

Google. (2023). seq2seq. In https://google.github.io/seq2seq/.

Greenberg, J. (2005). *Genetic linguistics: essays on theory and method*. Oxford: OUP.

Groves, M., & Mundt, K. (2015). Friend or foe? Google Translate in language for academic purposes. *English for Specific Purposes, 37*, 112–121. https://doi.org/10.1016/j.esp.2014.09.001

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist, 11*(2), 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

Kleiman, E. (2021). Data Management Tools for Real-Time Monitoring/Ecological Momentary Assessment Data. In https://cran.r-project.org/web/packages/EMAtools/EMAtools.pdf.

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution, 33*(7), 1870–1874.

Kuznetsova, A., Brockhoff, P.B., & Christensen, R.H.B. (2016). lmerTest: Tests in Linear Mixed Effects Models. In (Version R package version 2.0–33) https://CRAN.R-project.org/package=lmerTest.

Lear, A., Oke, L., Forsythe, C., & Richards, A. (2016). "Why Can't I Just Use Google Translate?" A Study on the Effectiveness of Online Translation Tools in Translation of Coas. *Value in Health, 19*(7), A387. https://doi.org/10.1016/j.jval.2016.09.232

Lee, K., & Ashton, M. C. (2004). Psychometric Properties of the HEXACO Personality Inventory. *Multivariate Behavioral Research, 39*(2), 329–358. https://doi.org/10.1207/s15327906mbr3902_8

Lee, K., & Ashton, M.C. (2023). HEXACO-PI-R Materials for Researchers. http://hexaco.org/hexaco-inventory [Record #3174 is using a reference type undefined in this output style.].

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*(8), 707–710.

Ozolins, U., Hale, S., Cheng, X., Hyatt, A., & Schofield, P. (2020). Translation and back-translation methodology in health research – a critique. *Expert Review of Pharmacoeconomics & Outcomes Research, 20*(1), 69–77. https://doi.org/10.1080/14737167.2020.1734453

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by generative pre-Training*.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 25, pp. 1–65). Academic Press.

Thomson, R., Yuki, M., Talhelm, T., Schug, J., Kito, M., Ayanian, A. H., Becker, J. C., Becker, M., Chiu, C.-Y., Choi, H.-S., Ferreira, C. M., Fülöp, M., Gul, P., Houghton-Illera, A. M., Joasoo, M., Jong, J., Kavanagh, C. M., Khutkyy, D., Manzi, C., Marcinkowska, U. M., Milfont, T. L., Neto, F., Oertzen, T. v, Pliskin, R., Martin, A. S., Singh, P., & Visserman, M. L. (2018). Relational mobility predicts social behaviors in 39 countries and is tied to historical farming and threat. *Proceedings of the National Academy of Sciences, 115*(29), 7521–7526. https://doi.org/10.1073/pnas.1713191115

van der Loo, M. (2014). The stringdist package for approximate string matching. *The R Journal, 6*, 111–122. https://CRAN.R-project.org/package=stringdist.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.

W3 Techs. (2023). Usage statistics of content languages for websites. https://w3techs.com/technologies/overview/content_language.

Wenfeng, Q., & Yanyi, W. (2019). jiebaR: Chinese Text Segmentation. In (Version 0.11) https://CRAN.R-project.org/package=jiebaR.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., Grolemund, G., Hayes, A., Henry, L., & Hester, J. (2019). Welcome to the Tidyverse. *Journal of Open Source Software, 4*(43), 1686.

Wickham, H., Chang, W., & Wickham, M. H. (2016). Package 'ggplot2'. Create elegant data visualisations using the grammar of graphics. *Version, 2*(1), 1–189.