Amsterdam University Press

COMPUTATIONAL COMMUNICATION RESEARCH 1.5 (2025) 1–24
HTTPS://DOI.ORG/10.5117/CCR2025.1.5.CHEW

Evaluating Machine Translation Solutions for Accessible Multi-Language Text Analysis: A Back-translation based Approach

Edward W. Chew
Department of Communication, University of California Davis, Davis, CA

Mahasweta Chakraborti
Department of Communication, University of California Davis, Davis, CA

William D. Weisman
Department of Communication, University of California Davis, Davis, CA

Seth Frey
Department of Communication, University of California Davis, Davis, CA

Abstract

English is the international standard of social research, but scholars are increasingly conscious of their responsibility to meet the need for scholarly insight into communication processes globally. This tension is as true in computational methods as in any other area, with revolutionary advances in the tools for English language texts leaving most other languages far behind. In this paper, we aim to leverage those very advances to demonstrate that multi-language analysis is currently accessible to all computational scholars. We show that English-trained measures computed after translation to English have adequate-to-excellent accuracy compared to source-language measures computed on original texts. We show this for three major analytics—sentiment analysis, topic analysis, and word embeddings—over 16 languages, including Spanish, Chinese, Hindi, and Arabic. We validate this claim by comparing predictions on original language tweets and their back-translations: double translations from their source language to English and back to the source language. Our results suggest that Google Translate, a simple and widely accessible tool, effectively preserves semantic content across languages and methods. Modern machine translation can thus help computational scholars make more inclusive and general claims about human communication.

Keywords: multi-lingual text analysis, computational text analysis, natural language processing, back-translation, topic modeling, word embedding, sentiment analysis

Introduction

Humans communicate in thousands of languages, yet a single language, English, attracts the bulk of Communication research. This not only has the effect of depriving other languages of adequate attention but also deprives English-focused scholars of any sense of where the language stands relative to others. The general use of English-trained tools for English-focused analyses in the social science community is particularly notable given the ubiquity of multi-lingual data and the power of modern computational natural language processing. For example, social media researchers on Twitter typically begin with raw data that is highly multilingual before filtering out all tweets except for English or some other single language. With such practices, scholars miss a tremendous opportunity to test the generalizability of social media-observed big data claims. However, bringing standard text analysis tools to the level of training and refinement that English-trained tools receive is a forbidding prospect that few multi-lingual scholars have the training, resources, and language background to pursue. Building upon a recent special issue in this journal on the subject of computational approaches to multilingual text analysis (Mariken A.C.G Van Der Velden & Baden, 2023), we propose a simple alternative approach that makes texts from over 100 languages accessible to the complete variety of analyses that are typically available to only English-focused scholars. Specifically, we demonstrate that modern machine translation has reached the level of refinement necessary to preserve sentiment, topics, and semantic distance, making multi-language datasets legible to state-of-the-art English-trained tools. More importantly, we provide a method that social scientists can use, without expert manual translations, to validate multilingual analyses based upon large-scale machine translations to English. By providing validation of state-of-the-art machine translation, along with easily adaptable demonstration code, we aim to broaden the horizon of computational research and support Communication scholars in increasing the relevance and generality of their work.

Google Translate is the most popular, accurate, and accessible multilingual neural machine translation service and it has already proven its potential for multilingual social science (De Vries et al., 2018; Hampshire & Salvia, 2010; Lotz & Van Rensburg, 2014). As of the time of this work, Google Translate uses Transformer (Vaswani et al., 2017) based architectures (Conneau et al., 2019; Xue et al., 2020) for generating translations for over 133 languages (Caswell, 2022). In this paper, we demonstrate the efficacy of Google Translate in retaining sentiment valence across translations of

large hand-coded and machine-coded Twitter datasets composed of tweets in 16 global non-English languages from four language families, being of Indo-European, Uralic, Semitic, and Sinitic origin. With our findings that Google Translate preserves semantics across three common text analysis tasks, scholars may be emboldened to utilize Google Translate and other multilingual neural machine translation services to expand the generalizability of their research. Importantly, they can do so with a sense of the accuracy and inaccuracy of their method, leveraging a simple universal approach for validating transitions. As we show, work with non-English languages can benefit from advanced English-trained natural language processing tools, and computational findings usually restricted to the English language can be expanded to broaden scholars' knowledge of global social phenomena.

Academics use Twitter datasets for a wide range of scholarship, including sentiment analysis (e.g., Gautam & Yadav, 2014), algorithmic training (e.g., Braithwaite et al., 2016), and even COVID-19 detection (e.g., Gharavi et al., 2020). English-language corpora have been used to predict election results (Nausheen & Begum, 2018), analyze consumer preferences (Ahmed & Danti, 2016), and explore pro- and anti-childhood vaccine communities' influence on Twitter (Featherstone et al., 2020).

As valuable as this work is, it can only be more valuable extended across languages. Frey et al. (2018) use a corpus of six languages to document the general ripple effects of emotional influence through others and back around to the self. Mocanu et al. (2013) use data on 78 languages to characterize inter-linguistic diversity and intra-linguistic drift across municipalities and regions. Alshaabi et al. (2021) compare the dynamics of social influence on Twitter over 150 languages. In other disciplines, large-scale multi-language comparisons have identified universal patterns in the cross-language naming of colors (Lindsey & Brown, 2009), as well as a universal preference for shortening of dependency length in human sentence production (Futrell et al., 2015).

This work is coming at a time of upheaval in natural language processing and computational methods in general. Large language models have transformed the questions that can be posed and the quality with which we answer them while also providing unprecedented accessibility. However, these changes are motivating work on the frontiers of computational communication science to properly characterize their strengths and weaknesses and understand how they should be integrated into quantitative methodology. Because this work aims to make multi-lingual approaches available

to the largest possible audience, we focus this inquiry on those methods that are familiar, accessible, time-tested, first-principle, and compute efficient. We look eagerly to a consensus on the proper integration of LLMs into computational communication science methodology, so that they can reliably serve as broad an audience as the simpler but more established tools we investigate here. Until then, given the risks and limitations of LLMs, including them does not serve the audience of this study with immediately proven, trusted, understood, and widely available methods. The objective of our work is not to demonstrate or achieve new thresholds of translation for translation-based tasks but to validate the usability of translation at sufficient accuracy across a wide range of functions. Despite the proliferation of state-of-the-art tools, authors such as van der Veen make a compelling case for the enduring place of "cheap-but-good" techniques (Van Der Veen, 2023). While more sophisticated methods are expected to deliver superior performance depending on use cases, simple baseline methods allow us to report practical insights and benchmarks relatable to researchers and social scientists at large.

Our research demonstrates the effectiveness of Google Translate in maintaining sentiments, topic clusters, and semantic distance for tweets in all languages we examine. We validate the approach using "back-translation," a classic validation method for translation in social science (Brislin, 1970; Ervin & Bower, 1952) and computer science (Dostert, 1963) in which scholars compare an original text to a version of that text that has been translated from its original language to another language (in our case English) and then back again to the original (Figure 1). This makes it possible to directly compare the accuracy of English-trained tools on English translations to original-language-trained tools on original-language texts while controlling for semantic drift introduced by the translation process itself. A significant problem with computational approaches to large datasets is the problem of garbage-in-garbage-out, that long analysis pipelines with technical steps are error-prone. This can be addressed with a commitment to validation at every stage of an analysis. With our validation of Google Translate, we develop our argument for back-translation and the importance of including it or some other quantitative validation in multilingual pipelines. With some exceptions (e.g. Maier et al. (2022)), this method remains obscure to computational text scholars.

Back-translation continues to be a popular validation method in computer science's machine translation literature because it provides researchers access to an autoencoder approach to model training and development (Heo

Figure 1: **We compare analytics computed on texts in their original languages to translated English language analytics and texts translated back to the original language.** Differences between the original and translated texts are typically difficult to attribute to semantic differences between the languages and "semantic" imposed by poor translation. Comparing original and back-translated texts enables us to control for the effect of drift and focus on semantics.

& Choi, 2023; Zhang et al., 2022). It is particularly useful in the absence of large professionally translated "ground truth" corpora, a common scenario for low-resource languages and domain-specific corpora. Back-translation has been used creatively for style transfer (Prabhumoye et al., 2018), as a data augmentation technique, and to automatically generate instances of a kind of text (in this case offensive comments) from a small corpus of examples (Dai et al., 2022; Ganganwar & Rajalakshmi, 2022), while iterative back-translation has been used as a noise reduction technique in machine translation research (Hoang et al., 2018; NLLB Team et al., 2022).

Despite its early role and enduring popularity in computer scientific translation research, it continues to be rare in computational social science. For example, it goes entirely unmentioned in a recent review of multilingual text analysis methods (Reber, 2019). And in CCR's own Special Issue on Multilingual Text Analysis (Mariken A.C.G Van Der Velden & Baden, 2023). For example, an overview by Licht and Lind emphasizes several potential problems with machine translation approaches to multilingual text analysis, as well as many potential solutions (Licht & Lind, 2023). The problems they name include the forbidding programming skills required to leverage machine translation packages, the cost of commercial solutions, and the complexity of preprocessing. They also name two more technical problems. One is the issue that "input alignment does not guarantee output alignment": that two texts in different languages that may or may not mean the same thing, when translated into the same language, may not or may appear as meaning the same thing. Another challenge is the evaluation of context sensitivity for the case of domain-specific corpora: that specialist documents translated through general tools may lose their specialist meaning. For solutions they emphasize manual inspection, leverage of domain knowledge,

or quantitative comparison based on documents that are known a priori to be identical or at least comparable. However, this discussion omits back-translation which, when implemented on a commercial platform such as Google Translate, offers scalable solutions and strategies for all but one of these problems (cost), and is preferable to all other proposed solutions.

The approach we offer overcomes the programming challenge because we publish minimal working code for implementing our method. Because Google Translate is designed to leverage naturalistic inputs, it requires almost no preprocessing. And back-translation offers solutions to output alignment and context sensitivity by offering direct same-language comparisons of a text to itself after potentially disaligning or context-mangling acts of translation. It improves on other validation methods by being quantitative, automatic, scalable, and well-controlled, in the sense that it can be performed early in a multilingual data analysis pipeline, unlike other quantitative validation approaches, which tend to work by comparing outputs from a further downstream multi-step analysis pipeline. Recent influential multilingual projects in social science have relied on special "bitext" datasets in which a single artifact has already been professionally transcribed into several languages (Proksch et al., 2019) or manual inspection of comparable outputs for validation of their multilingual translation (Lucas et al., 2015). However, given the relative paucity of datasets uniquely suited for multilingual analysis, there is an acute need for a simple, general, proven approach such as back-translation.

We first test the preservation of sentiments using two large public multilingual Twitter datasets, one with hand-coded sentiments (Mozetič et al., 2016) and another with machine-coded sentiments (Imran et al., 2022). The second portion of our research applies the same two datasets to show that Google Translate preserves topic clusters after back-translation, demonstrating a similar level of semantic conservation for this second common text analysis task. In the third and final portion of our present study, we demonstrate the effectiveness of out-of-the-box machine translation on a third common text analysis approach: neural word embeddings. After back-translation, 10 of the 16 languages examined performed better than chance in maintaining a minimal embedding distance. Our findings provide strong support for the use of modern machine translation to expand academic attention to the languages spoken by most humans.

Methods

Datasets

We utilized two large, multilingual Twitter datasets. First, we examine the Mozetič et al. (2016) dataset, which contains over 1.6 million general Twitter posts hand-labeled as containing "positive", "negative", or "neutral" labels for 15 European languages: Albanian, Bosnian, Bulgarian, Croatian, English, German, Hungarian, Polish, Portuguese, Russian, Serbian, Slovak, Slovenian, Spanish, and Swedish. To expand the scope of our research beyond European languages, we added tweets from the Imran et al. (2022) COVID-19 dataset, a larger (70 million tweet) corpus including tweets in Chinese, Hindi, and Arabic. While these two datasets are comparable (both include sentiment labels), they differ in subject and date, as well as in how they determine sentiment scores. Those in Mozetič et al. (2016) were applied by human language-domain experts, while tweets from the Imran et al. (2022) dataset were determined by algorithms (all trained within-language).

Data cleaning and preprocessing

Before translation and subsequent analysis, we preprocessed all Twitter data to remove Twitter handles, Twitter retweet formatting, URLs, numbers, and empty tweets and converted all content to lowercase. We dropped Serbian from our analysis halfway through the study after discovering that the Mozetič et al. (2016) dataset contains Cyrillic Serbian, but Google Translate only supports Latin-character Serbian. We obviously excluded all English-language tweets from validation by back-translation through English.

To reduce our dataset to a more manageable—and affordable—size (the Google Translate API is paid), we randomly sampled 20,000 tweets from each of the 13 applicable European languages from Mozetič et al. (2016) dataset, and 9,000 tweets from Chinese, Hindi, and Arabic from the Imran et al. (2022) dataset, for a total of 16 languages.

Translation process

Utilizing the Google Translate API, we translate all tweets from their "original language" datasets into English, saving the results as our "English translated" dataset. We then translate all the English translated tweets back to their original language, saving it as our "backtranslated" dataset (Figure 1). Our results are based only partly on three-way comparisons between these

datasets. Where there is no meaningful correspondence between English- and original-language analyses, we use only two-way comparisons between the original and back-translated datasets.

To enable each of the following tasks, we used the open-source software Polyglot's language-specific tokenization feature, based on the Unicode Text Segmentation algorithm (Al-Rfou et al., 2015).

With this manuscript we share the scripts and instructions, to enable researchers to easily extend their single-language corpus research to multiple languages. The code is available at https://osf.io/jx476/.

Sentiment analysis

We conduct our sentiment analysis task with Polyglot's polarity lexicons (Chen & Skiena, 2014). Polyglot allows the generation of sentiment labels in more than 100 languages, with "-1" indicating negative sentiment, "0" indicating neutral sentiment, and "1" indicating positive sentiment for each word in each original-language tweet. Based on the difference between the number of positive sentiment words and negative sentiment words, we generate an overall polarity for each tweet. Polyglot's lexicon-based sentiment analysis relies on a valence dictionary of positive and negative words, computing the sentiment of a text as the simple sum of the valences of its words, normalized back down to the [-1, 1] interval. Our pipeline excluded neutrally labeled tweets: as a result of Polyglot's lexicon-based sentiments, short texts like Twitter posts are overwhelmingly labeled as neutral, which makes it difficult to distinguish the performance of sentiment analyses across translations.

We computed confidence intervals around the accuracy of each language's sentiments with bootstrapping, taking 1000 iterates with a sample size equal to the original datasets. The final sentiment accuracies are the medians of the bootstrapped accuracies.

Topic clustering

While sentiment analysis is a common natural language tool for behavioral analysis, it is a categorical approach that only supports a limited conceptualization of opinion and preferences. Meanwhile, topic analysis is another popular approach with the potential to support even more inclusive abstraction of insights from text. We expand our investigation of Google Translate's ability to preserve the content of translated text through topic analysis.

The scope of translation for corpus-based analytics can be assessed through the consistency of text topics across English translations. Perturbations introduced cumulatively over rounds of translation may influence how the text gets interpreted and categorized by clustering/topic modeling methods. Assignment under the original topic, or reassignment to a disparate topic is determined by the extent of change induced in context and themes with respect to the original text.

We model our topic clustering approach after Yin and Wang (2014) who present the open-source software GSDMM ("Gibbs Sampling algorithm of the Dirichlet Multinomial Mixture). Our choice of method was motivated by several considerations. While LDA, another popular topic modeling approach, produces probability distributions of themes for any text, GSDMM is based on a robust statistical clustering algorithm that can deduce the best topic label for concise texts typical of social media environments e.g., Twitter (Yin & Wang, 2014).

Common computational tools are typically built to support only the English language, from stopwords to supported character sets. We follow the data cleaning steps from Yin and Wang (2014) by removing both emojis and stopwords. We excluded Albanian and Bosnian due to their incompatibility with our data-cleaning dependencies.

Our cluster analysis process was as follows. For each language, we used a total of five iterations of the clustering algorithm. We then classified the back-translated tweets into the clusters generated on the original language tweets using the topic model trained on the original text data. To estimate the success of machine translation at thematic preservation under topic analysis, we evaluate the semantic closeness between the topics of the original text and the back-translated text. We use the BERTopic (Grootendorst, 2022) library to calculate the drift between the topics assigned to a text by GSDMM, before and after back-translation. It generates TF-IDF-based word vector representations for the topic clusters based on the texts they comprise. Robust to stopwords, the vectors generated through this approach weigh words based on their importance within the topic and distinctiveness with respect to other topics. It should be noted that while BERTopic is itself also a popular topic modeling library, it is currently only being used for representing and comparing topic clusters (obtained through GSDMM) through its topic model heatmap feature.

The higher cosine dot product between the vectors corresponds to greater similarity, with values closer to 1 indicating near-identical topics. Note this is a non-linear measure and should be interpreted as a proportion of simi-

larity (like the Pearson correlation). Given limited support for Chinese in BERTopic's default tokenizer, we additionally use the Jieba library to process Chinese texts before TF-IDF.

Like many clustering algorithms, GSDMM requires researchers to impose a desired number of clusters rather than identifying the number of clusters through the same emergent process as cluster assignments. But the ability of back-translation to preserve topic clusters depends on the number of clusters. Therefore, we observe the effectiveness of topic preservation across a range of clusterings by training models on each original language dataset for 2, 5, 10, 15, 20, 50, 100, 150, and 200 clusters.

Unlike our evaluation of sentiment analysis, topical comparison is only applicable across the original language and its back-translation: it is not able to compare either to English. While the framework of sentiment analysis assumes the universally accepted perception of "positive" and "negative" sentiments independent of languages and datasets, topics can be more dynamic as they reflect themes composed of words in a specific text. Within a corpus from a particular language, a set of themes can hold across the original language and back-translated tweets, given their shared lexicons and latent semantics. But lexicons in English and each original language are mostly non-overlapping, and there is ultimately no basis to map English translations to original-language-derived topics.

Word embeddings

Polyglot (Al-Rfou et al., 2015) also supports semantic word embeddings across its languages. We determine semantic preservation under word embeddings by calculating sentence embeddings of the original and back-translated tweets (as the sum of the embeddings of their words) and calculating their cosine distance. Under this formalism, a distance of zero indicates perfect preservation of semantics after translation.

To measure how well machine translation preserves semantics under word embeddings, we compared the embedding distances after back-translation to two baseline distances. For each tweet in a 5,000-tweet sample of each language, we computed its average distance and minimum (non-zero) distance from the other 4,999 tweets. The average values from the 5,000 tweets became the average baseline and the minimum baseline. If the average distance between the original and back-translated tweets is lower than either of the baseline distances, it suggests meaning is preserved despite the semantic drift imposed by the machine translation process. The minimum baseline is the more rigorous of the two.

Results

Our primary finding is that Google Translate is faithful enough to preserve the semantics of multilingual texts under three common text analysis tasks: sentiment analysis, topic analysis, and word embeddings.

Application 1: Preservation of sentiment

We find that the overall accuracy of sentiment scores decreases less than 2% after back-translation, from a median 65.36% accuracy (with a very tight 99% high-confidence interval (HCI) of [65.35, 65.38]) to 64.03% [64.02, 64.05]. While small, this decline was statistically significant, as measured by the separation of the 99% HCI bars. We display this result in Figure 2, below.

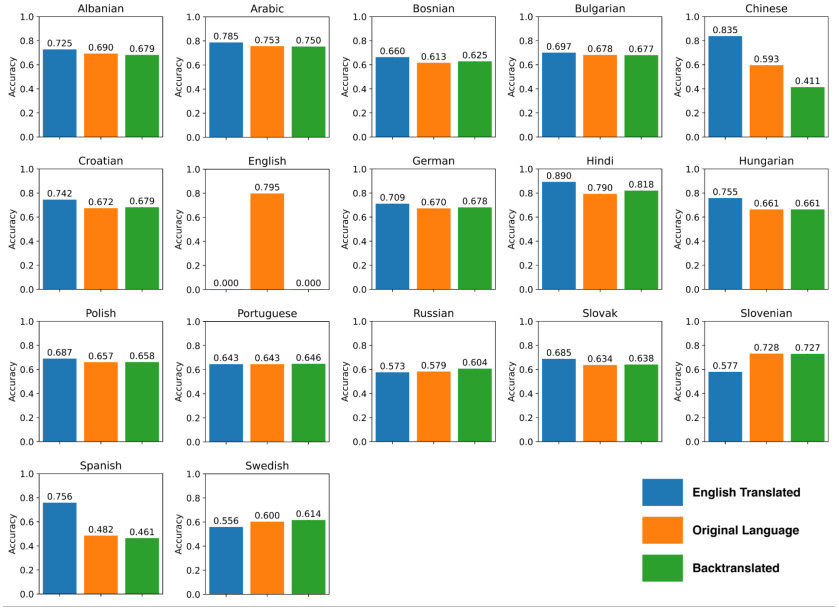

Figure 2: **Sentiment analysis overall retains accuracy after back-translation by machine methods.** Median sentiment detection accuracy increases 4.78% from original language to English translated language datasets, and falls 1.33% from original language datasets to back-translated language datasets. Note that 99% error bars are too narrow to be displayed.

We did have one surprise from this process. We expected that the accuracy of English-trained sentiment on the English-translated tweets would be between or below the accuracy of the original or back-translated tweets, whose "ground truth" sentiments were computed with models trained specif-

ically for those languages. Instead, sentiment accuracy increases by 4.78% following the original languages' translations into English (original language median accuracy: 65.36%, HCI [65.35, 65.38]; English translated median accuracy: 70.14%, HCI [70.12, 70.15]). Sentiment labels extracted from English translations are more accurate than sentiment labels of original language tweets, despite the process of translation in between (Figure 2). We speculate on this result in the Discussion section.

Looking specifically at how different languages performed, we found the expected decrease in accuracy rates between the original language datasets and the back-translated datasets for Albanian, Arabic, Bulgarian, Chinese, Slovenian, and Spanish (Figure 3). Unexpectedly, the remaining language datasets belonging to the languages of Bosnian, Croatian, German, Hindi, Hungarian, Polish, Portuguese, Russian, Slovak, and Swedish experienced an increase in sentiment accuracy from the original language to the back-translated form. Although languages, on average, showed higher accuracy in English translation, the original language datasets of Russian, Slovenian, and Swedish show a drop in sentiment accuracy when translated to English (while the remaining others, Albanian, Arabic, Bosnian, Bulgarian, Chinese, Croatian, German, Hindi, Hungarian, Polish, Portuguese, Slovak, and Spanish all improve).

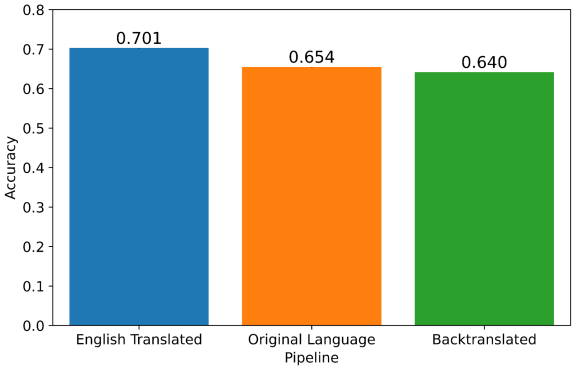

Figure 3: **Comparison of sentiment labeling accuracy across languages, before, during, and after back-translation.** Seventeen language sentiment detection accuracy from original language > English translated > back-translated datasets. Note that 99% error bars are too narrow to be displayed.

Application 2: Preservation of lexical topic assignments

We compare topic assignments after back-translation in order to assess Google Translate's quality at preserving lexical topics (Fig. 4). We plot the pairwise similarities between original and back-translated topics for the 14 languages. To increase the interpretability of topic modeling across back-translation, we provide a baseline reference. For a given number of topic clusters, the baseline depicts the mean similarity between a text's original topic and the next closest topic, across all languages. This represents a reasonable threshold on thematic shift for the likely event where perturbations introduced from back-translation only generally result in the reassignment of every text to the next best topic (with reference to the original).

Our assessments find evidence that Google Translate reasonably preserves thematic representations, outperforming the baseline for all languages in the study. In both tests, we consistently see the worst performance for Chinese, the only Sinitic example in our dataset, even with pre-processing and specialized utilities to ensure comparable modeling and validation across languages. Some observations about Chinese are that its script is logographic rather than alphabetic, source messages had more "pollution" from English words than other languages, and, while messages from all languages came back shorter after back-translation (by 20-25%), Chinese messages were much shorter after back-translation (closer to 40%). We further confirmed that all languages outperform their respective baselines, i.e., for every language, the inter-topic similarity from back-translation was still greater than that between the original topic and its closest neighbor.

Topic modeling can be sensitive to the choice of N. Fewer topics lead to larger umbrella clusters, each composed of several related themes. Minor perturbations from translations and resulting changes in texts' word content (i.e., removal/addition/replacement of words between rounds of translation) generally keep it within the topic as long as central themes and words are mostly intact. However, a high number of clusters results in more granular, mutually distinct topics, as depicted by the baseline's sharp gradient, where for a given N, the baseline represents mean distances between pairwise closest topics. Considerable changes from back-translation can sometimes result in the text no longer associating with the original topic. This is more pronounced for higher values of N, as even minor text variations can cause a text to no longer be associated with a closely fitted topic. Table 2 provides illustrative examples of how a higher N and the extent of variations from back-translation can impact topical consistency. Consequently, the average similarity between original and back-translated topics slightly decreases as

their cosine similarity drops steadily from 0.98 (n = 2) to 0.81 (n = 50), after which it remains pretty stable, reaching 0.80 at = 200 (Figure 4).

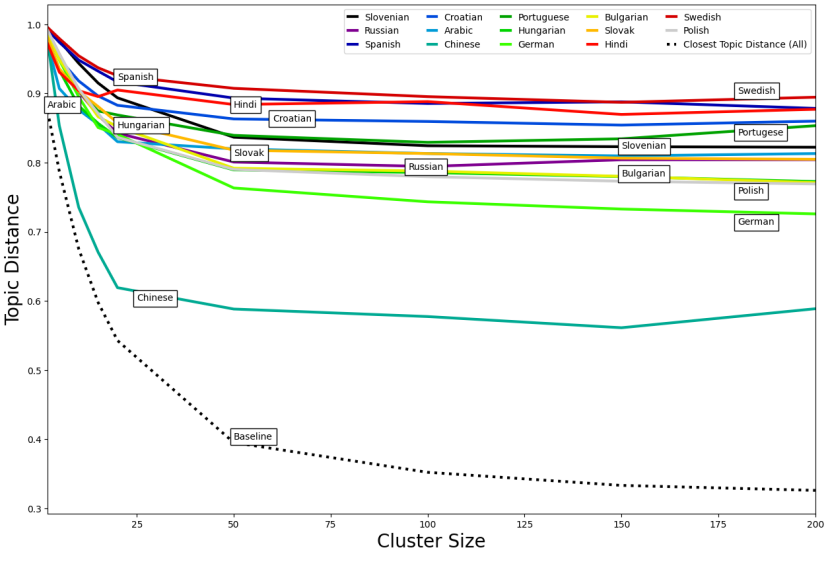

Figure 4: **Similarity between topics of original and back-translated texts increases with the number of topics/granularity but remains stable subsequently.** For a given number of topic clusters, the baseline depicts the mean similarity between a text's original topic and its closest other topic, across all languages. Performances decline steadily up to n = 50, after which they remain fairly stable, and all languages perform well above the baseline

Application 3: Preservation of semantics in embedding space

In the final application of this work, we examine the multilingual preservation of semantic vectors in high-dimensional neural embeddings after machine translation and back-translation.

On average, original language tweets are significantly closer to their back-translations than to other original language tweets in the same collection (Figure 5). Across languages, the average distances of original language tweets from each other are 0.207–0.496 units, their minimum distances from each other are 0.028–0.132, and their distances from their back-translations are 0.041–0.184. Being less than half of the average baseline (except Chinese) and below or slightly above the minimum baseline, we can conclude that the semantic change introduced by the translation algorithm is enough to change the meaning of a back-translated tweet to be mistakable for a

different closely related tweet, but not the typical more distantly related tweet.

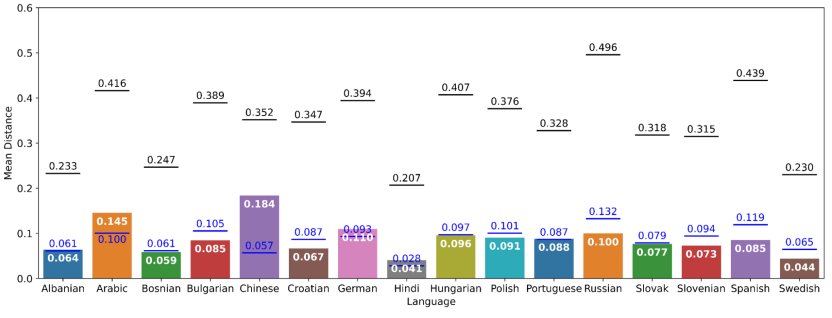

Figure 5: **Tweets are closer to their back-translations on average than to other tweets.** Average (black) and minimum (blue) distance between original language and back-translated sentence embeddings by language. Black lines denote the mean baseline distance and blue lines denote the minimum baseline distance. All 16 languages have mean distances below their mean baseline, meaning that they outperform chance. All languages but Albanian, Arabic, Chinese, German, Hindi, and Portuguese have mean distances below their minimum baseline, meaning that there is no message in the sample closer to a source message than its own back-translation. In these six languages, back-translated tweets are further from their source tweet in meaning than tweets that are very semantically similar to the source. But even tweets in these languages are still consistently closer to their source than the average tweet.

Of the 16 languages involved in the analysis, 6 languages (Albanian, Arabic, Chinese, German, Hindi, and Portuguese) failed the minimum baseline test, with back-translated tweets having greater semantic distance from their originals than the average closest outside tweet. The Albanian, German, and Portuguese corpora failed by small margins (mean distance of 0.064 compared to minimum baseline distance 0.061 in Albanian; distance 0.110 compared to baseline 0.093 in German; 0.088 against 0.087 in Portuguese). But in Arabic, Chinese, and Hindi, embeddings of translations were even further from their original (distance 0.145 against 0.100 in Arabic; 0.184 against 0.057 in Chinese; 0.041 against 0.028 in Hindi). It should be noted that Arabic, Chinese, and Hindi were drawn from the Imran et al. (2022) dataset focused on COVID-19-related tweets, included in our effort to expand this project's analysis beyond languages of European origin. As baseline measures were calculated on the distance between random tweets relative to their distance with all other tweets, and these tweets were semantically more closely related, these languages' baseline measures may have been especially narrow relative to those of the other languages as a result of their shared topic. Although they failed the rigorous minimum distance test, Arabic, Chinese,

and Hindi passed the mean distance test: they were closer in meaning to their original than the average tweet (mean baseline distances 0.416, 0.352, and 0.207, respectively).

Discussion

As the global academic world becomes increasingly interconnected, Communication scholars must meet the challenge of making claims about communication processes more globally relevant. Fortunately, with recent advances in natural language processing, quantitative Communication research has an opportunity to be multilingual by default. Advances that bring equal attention to more of the world's languages will not only provide greater generality of results, but greater attention to the work of Communication scholars from all parts of the world. Standard approaches to large multilingual corpora will also allow the rapid transfer of groundbreaking knowledge to and from the international Communication community.

Of course, these advances have downsides. When multi-language analyses are conducted by scholars who can't speak all of those languages, it becomes harder for them to "gut check" or "sanity check" their results, culturally contextualize those results, and interpret whatever valid cross-linguistic differences that do appear. By encouraging researchers to conduct multilingual studies by default, we are almost necessarily advocating for a circumstance in which scholars are making conclusions about languages that they do not know. Although this approach has some acceptance in other fields, such as large-scale comparative linguistics, it would be understandable to see it as controversial. As novel as this problem may be, the way forward may not be novel at all. Quantitative and qualitative methods have a fundamental complementarity, with the former bringing generality as the latter brings depth and sensitivity to context. By supporting the summary quantitative claims of non-speakers with citations to other work by native speakers and other domain experts, scholars may be able to justify not knowing the languages they are engaging with. This complementary approach will be particularly valuable for understanding outliers. In the case of our research, results sometimes varied widely between languages. Having ruled out explanations that go to the phylogeny, character set, and geography of these languages, domain experts become the best candidates for understanding how and why specific languages deviate from their peers. This illustrates the importance of framing our contribution as a complement to expert-based multi-language Communication research, rather than a substitute.

One surprising result from this work was that the accuracy of senti-

ment detection after translation into English, and in some cases after back-translation, was higher than in the original texts. This echoes the identical finding by Araujo et al. (2016), who also evaluate the effectiveness of machine translation to English for sentiment analysis, for nine languages (Araujo et al., 2016). This finding is easier to understand with an appreciation of how sentiment analysis works in libraries like Polyglot. Polyglot uses the "dictionary" method, in which hundreds to thousands of high-frequency words are given sentiment scores, and the score of a statement is calculated from the sum of scores of the subset of words in the detector's sentiment dictionary. If the dictionary is large, or the text is long, then its assigned sentiment score will be based on many signals. Consequently, this method is less suitable for rarer languages and shorter texts (like tweets), which are less likely to contain scored words. It is also more suitable for texts with more common words since uncommon words are less likely to appear in a language's sentiment dictionary.

Why would translation to English, or back-translation from English, improve task performance? Polyglot's English sentiment lexicon is longer than those of other languages in this analysis, which may hold for dictionary-based sentiment detection lexicons in general. And subsequent back-translation may improve detection performance if it results in uncommon unscored words from the source text being back-translated into more common words that are scored. This result underscores the need to validate Google Translate for each natural language task that it is being used to support.

In this work, we validated the performance of Google Translate by leveraging several source-language tools: sentiment lexicons, word embeddings, and tokenizers. However, as others make use of machine translation, they will not have the comfort of source-language tools, and may feel that they are "flying blind." Although we succeed in showing that translation introduces negligible drift, it may still be uncomfortable to apply it to a new dataset, particularly with text analysis methods beyond the three that we validate here (such as custom classifiers). To address this concern, researchers can use not only our conclusions but also the back-translation method itself to perform partial validations for their case. Most likely, it should be possible to find "home language" tools for at least a handful of languages in a larger corpus. If an author can show satisfactory and stable performance across this subset by comparing original and back-translated texts, they can assure their audience that the method is probably working for other languages as well.

Such back-translation can be used to instill confidence in results despite

the potential influences noted above: the conversion of rare words to common words in the sentiment task, language contamination, and the lack of source-language tools. For example, a scholar could perform iterative back-translations to calculate how many cycles must be introduced for the statistical significance of their result to degrade below a threshold. If it takes a large number of back-translations to degrade a result, readers can have confidence that artifacts introduced by the method are not sufficient to explain those results. Conversely, if machine translation is artificially amplifying a result, scholars can measure this effect with iterated back-translation to suggest an appropriate amount of caution.

Another design choice of this work ensures the generality of the method we introduce. All three applications of this work were performed with tweets. Tweets are short, making them challenging for text analysis methods like sentiment and topic analysis. That our method is effective on challenging text is encouraging for scholars who would extend this method to more typical (longer) texts.

A limitation of our approach is its accessibility. We have argued that Google Translate is very accessible, and this is true in that it requires a small amount of code (that we provide) to translate large quantities of text to English and back. However, our approach is not as financially accessible. The Google Translation API costs $20 USD per million characters. In practice, this translates to roughly $100 USD per 130,000 tweets. Fortunately, free translation tools of comparable quality are increasingly common and can also be validated in practice using back-translation.

Conclusion

There is an unmet need to extend Communication scholars' applications of text analysis to more languages, particularly in the data-rich context of social media studies. Translation tools such as Google Translate can be immensely helpful in meeting this need. We have quantified Google Translate's effectiveness in maintaining sentence meaning in translations to and from English. In doing so, we have demonstrated the flexibility and simplicity of back-translation as an all-purpose tool for validating multilingual text analysis. Across 16 non-English languages, sentiment analysis scores were shown to improve when translated to English, and only diminish marginally when translated back to their original languages. Similarly, both topic and semantic distances are preserved during back-translation. Our findings demonstrate that machine translation is able to preserve semantic content and make non-English datasets legible to English-trained computational

tools. We hope this analysis gives researchers the confidence to use machine translation to simply and economically increase the number of languages involved in their research, and thereby the generality of their findings.

Acknowledgements

We would like to thank Arti Thakur, Communication Ph.D. Candidate at the University of California, Davis, for her assistance with the analysis. All code is available at https://osf.io/jx476/. The authors report there are no competing interests to declare.

References

Ahmed, S., & Danti, A. (2016). Effective sentimental analysis and opinion mining of web reviews using rule based classifiers. *Advances in Intelligent Systems and Computing*, 171–179. https://doi.org/10.1007/978-81-322-2734-2_18

Al-Rfou, R., Kulkarni, V., Perozzi, B., & Skiena, S. (2015). Polyglot-ner: Massive multilingual named entity recognition. *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, British Columbia, Canada, April 30- May 2, 2015*. https://doi.org/10.48550/arXiv.1307.1662

Alshaabi, T., Dewhurst, D. R., Minot, J. R., Arnold, M. V., Adams, J. L., Danforth, C. M., & Dodds, P. S. (2021). The growing amplification of social media: Measuring temporal and social contagion dynamics for over 150 languages on twitter for 2009–2020. *EPJ Data Science*, *10*(1), Art. 1. https://doi.org/10/gjq4qq

Araujo, M., Reis, J., Pereira, A., & Benevenuto, F. (2016). An evaluation of machine translation for multilingual sentence-level sentiment analysis. *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, 1140–1145. https://doi.org/10.1145/2851613.2851817

Blevins, T., & Zettlemoyer, L. (2022). Language contamination helps explain the cross-lingual capabilities of english pretrained models. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates*, 3563–3574. https://doi.org/10.18653/v1/2022.emnlp-main.233

Braithwaite, S. R., Giraud-Carrier, C., West, J., Barnes, M. D., & Hanson, C. L. (2016). Validating machine learning algorithms for twitter data against established measures of suicidality. *JMIR Mental Health*, *3*(2). https://doi.org/10.2196/mental.4822

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

Caswell, I. (2022). Google translate learns 24 new languages [Retrieved December 15, 2022]. https://blog.google/products/translate/24-new-languages/

Chen, Y., & Skiena, S. (2014). Building sentiment lexicons for all major languages. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 383–389. https://doi.org/10.3115/v1/p14-2063

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Dai, Y., Radford, B., & Halterman, A. (2022). Political event coding as text-to-text sequence generation. *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text (CASE)*, 117–123.

De Vries, E., Schoonvelde, M., & Schumacher, G. (2018). No longer lost in translation: Evidence that google translate works for comparative bag-of-words text applications. *Political Analysis*, *26*(4), 417–430. https://doi.org/10.1017/pan.2018.26

Dostert, L. E. (1963). Machine translation and automatic language data processing. *Vistas in Information Handling*, 92–110.

Ervin, S., & Bower, R. T. (1952). Translation problems in international surveys. *Public Opinion Quarterly*, *16*(4, Special Issue on International Communications Research), 595. https://doi.org/10.1086/266421

Featherstone, J. D., & Barnett, G. A. (2020). Validating sentiment analysis on opinion mining using self-reported attitude scores. *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*. https://doi.org/10.1109/snams52053.2020.9336540

Featherstone, J. D., Barnett, G. A., Ruiz, J. B., Zhuang, Y., & Millam, B. J. (2020). Exploring childhood anti-vaccine and pro-vaccine communities on twitter – a perspective from influential users. *Online Social Networks and Media*, *20*, 100105. https://doi.org/10.1016/j.osnem.2020.100105

Frey, S., Donnay, K., Helbing, D., Sumner, R. W., & Bos, M. W. (2018). The rippling dynamics of valenced messages in naturalistic youth chat. *Behavior Research Methods*. https://doi.org/http://doi.org/cwbz

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, *112*(33), 10336–10341. https://doi.org/10.1073/pnas.1502134112

Ganganwar, V., & Rajalakshmi, R. (2022). Mtdot: A multilingual translation-based data augmentation technique for offensive content identification in tamil text data. *Electronics*, *11*(21), 3574. https://doi.org/10.3390/electronics11213574

Gautam, G., & Yadav, D. (2014). Sentiment analysis of twitter data using machine learning approaches and semantic analysis. *2014 Seventh International Conference on Contemporary Computing (IC3)*. https://doi.org/10.1109/ic3.2014.6897213

Gharavi, E., Nazemi, N., & Dadgostari, F. (2020). Early outbreak detection for proactive crisis management using twitter data: Covid-19 a case study in the us. *arXiv*. https://doi.org/arXiv:2005.00475

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv*. https://doi.org/arXiv:2203.05794

Hampshire, S., & Salvia, C. P. (2010). Translation and the internet: Evaluating the quality of free online machine translators. *Quaderns: Revista de Traducció*, 197–209.

Heo, D., & Choi, H. (2023). End-to-end training for back-translation with categorical reparameterization trick. *arXiv*. https://doi.org/arXiv:2202.08465

Hoang, C. D. V., Koehn, P., Haffari, G., & Cohn, T. (2018). Iterative back-translation for neural machine translation. *ACL 2018*, *23*(32.5), 18.

Imran, M., Qazi, U., & Ofli, F. (2022). Tbcov: Two billion multilingual covid-19 tweets with sentiment, entity, geo, and gender labels. *Data*, *7*(1), 8. https://doi.org/10.3390/data7010008

Licht, H., & Lind, F. (2023). Going cross-lingual: A guide to multilingual text analysis. *Computational Communication Research*, *5*(2), 1. https://doi.org/10.5117/CCR2023.2.2.LICH

Lindsey, D. T., & Brown, A. M. (2009). World color survey color naming reveals universal motifs and their within-language diversity. *Proceedings of the National Academy of Sciences*, *106*(47), 19785–19790. https://doi.org/10.1073/pnas.0910981106

Lotz, S., & Van Rensburg, A. (2014). Translation technology explored: Has a three-year maturation period done google translate any good? *Stellenbosch Papers in Linguistics Plus*, *43*(0), 235. https://doi.org/10.5842/43-0-205

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis*, *23*(2), 254–277. https://doi.org/10.1093/pan/mpu019

Maier, D., Baden, C., Stoltenberg, D., De Vries-Kedem, M., & Waldherr, A. (2022). Machine translation vs. multilingual dictionaries assessing two strategies for the topic modeling of multilingual text collections. *Communication Methods and Measures*, *16*(1), 19–38. https://doi.org/10.1080/19312458.2021.1955845

Mariken A.C.G Van Der Velden, M. S., & Baden, C. (2023). Introduction to special issue on multilingual text analysis. *Computational Communication Research*, *5*(2), 1. https://doi.org/10.5117/CCR2023.2.1.VAND

Mocanu, D., Baronchelli, A., Perra, N., Gonçalves, B., Zhang, Q., & Vespignani, A. (2013). The twitter of babel: Mapping world languages through microblogging platforms. *PLOS ONE*, *8*(4), e61981. https://doi.org/10/f4vdv4

Mozetič, I., Grčar, M., & Smailović, J. (2016). Multilingual twitter sentiment classification: The role of human annotators. *PLOS ONE*, *11*(5). https://doi.org/10.1371/journal.pone.0155036

Nausheen, F., & Begum, S. H. (2018). Sentiment analysis to predict election results using python. *2018 2nd International Conference on Inventive Systems and Control (ICISC)*. https://doi.org/10.1109/icisc.2018.8399007

NLLB Team, Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., et al. (2022). No language left behind:

Scaling human-centered machine translation. *arXiv*. https://doi.org/arXiv:2207.04672

Prabhumoye, S., Tsvetkov, Y., Salakhutdinov, R., & Black, A. W. (2018). Style transfer through back-translation. *arXiv*. https://doi.org/arXiv:1804.09000

Proksch, S., Lowe, W., Wäckerle, J., & Soroka, S. (2019). Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches. *Legislative Studies Quarterly*, *44*(1), 97–131. https://doi.org/10.1111/lsq.12218

Reber, U. (2019). Overcoming language barriers: Assessing the potential of machine translation and topic modeling for the comparative analysis of multilingual text corpora. *Communication Methods and Measures*, *13*(2), 102–125. https://doi.org/10.1080/19312458.2018.1555798

Thara, S., & Poornachandran, P. (2018). Code-mixing: A brief survey. *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2382–2388. https://doi.org/10.1109/ICACCI.2018.8554413

Van Der Veen, A. (2023). Word-level machine translation for bag-of-words text analysis: Cheap, fast, and surprisingly good. *Computational Communication Research*, *5*(2), 1. https://doi.org/10.5117/CCR2023.2.8.VAND

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Xue, L., Constant, N., Roberts, A., Aditya, M. K. R. A. R., & Raffel, S. A. B. C. (2020). Mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv*. https://doi.org/arXiv:2010.11934

Yin, J., & Wang, J. (2014). A dirichlet multinomial mixture model-based approach for short text clustering. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 233–242. https://doi.org/10.1145/2623330.2623715

Zhang, H., Huang, H., Gao, J., Chen, Y., Xu, J., & Liu, J. (2022). Iterative constrained back-translation for unsupervised domain adaptation of machine translation. *Proceedings of the 29th International Conference on Computational Linguistics*, 5054–5065.

Original	Back translated	Original Topic by representative words (N=20)	Back-translated Topic by representative words (N=20)	Original Topic by representative words(N=200)	Back-translated Topic by representative words(N=200)
फंस गए पाकिस्तानी सटूडेंट! इमरान खान चीन से नहीं निकाल पाए, हो गया कोरना वायरस	फंसे हुए पाकिस्तानी छात्र! चीन से बाहर नहीं निकल पाए इमरान खान, हो गया कोरोना वायरस	उसक, सकत, हम, कह	उसक, सकत, हम, कह	रन, इमर, me, aaisa	शर, कहत, viralvideo, समदरज	
yeahhh !! está como para una mirada am no??	yeahhh !! ¿Es para echarme un vistazo, verdad?	Buenos, días, que, de	Que, no, me, te	Nunca, disponible, siempre, problema	Gracias, que, día, lo	

Table 2: **Sensitivity to Number of Clusters.** For lower N, topic clusters are more general. Minor text variations due to translations do not affect topic modeling. For higher N, topic clusters are more specific, and depending on how considerable variations are, back-translations may cause topic reassignment. For conciseness, topics are represented in the table using only the top 4 words from the TF-IDF vectors of the respective topic. The first example shows how higher N can sometimes see topic reassignment, while the second example shows an instance of major changes from back translation, as a result of which topics are not preserved across values of N.

Language	Family	Region	Speakers	Script	Word Order	Glottolog
English	Indo-European	Global	1.46B	Latin	SVO	stan1293
Albanian	Indo-European	Eurasia	7.5M	Latin	SVO	alba1267
Bosnian	Indo-European	Eurasia	2.6M	Latin	SVO	bosn1245
Bulgarian	Indo-European	Eurasia	10M	Cyrillic	SVO	bulg1262
Croatian	Indo-European	Eurasia	6.4M	Latin	SVO	croa1245
German	Indo-European	Eurasia	180M	Latin	V2	stan1295
Hungarian	Uralian	Eurasia	17M	Latin	SVO	hung1274
Polish	Indo-European	Eurasia	41M	Latin	SVO	poli1260
Portuguese	Indo-European	Eurasia	260M	Latin	SVO	port1283
Russian	Indo-European	Eurasia	260M	Cyrillic	SVO	russ1263
Serbian	Indo-European	Eurasia	12M	Cyrillic	SVO	serb1264
Slovak	Indo-European	Eurasia	7M	Latin	SVO	slov1269
Slovenian	Indo-European	Eurasia	2.5M	Latin	SVO	slov1268
Spanish	Indo-European	Eurasia	600M	Latin	SVO	stan1288
Swedish	Indo-European	Eurasia	13M	Latin	SVO	swed1254
Arabic	Afro-Asiatic	Afroasia	510M	Arabic	SVO	arab1395
Chinese	Sino-Tibetan	Asia	1.12B	Traditional/ Simplified	SVO	mand1415
Hindi	Indo-European	Asia	610M	Devanagari	SOV	hind1269

Table 1: Languages in this study, with descriptives