Cultural and Geographical Influences on Image Translatability of Words across Languages

Nikzad Khani
Boston University
nikzad@bu.edu

Isidora Chara Tourni
Boston University
isidora@bu.edu

Mohammad Sadegh Rasooli
University of Pennsylvania
rasooli@upenn.edu

Chris Callison-Burch
University of Pennsylvania
ccb@upenn.edu

Derry Tanti Wijaya
Boston University
wijaya@bu.edu

Abstract

Neural Machine Translation (NMT) models have been observed to produce poor translations when there are few/no parallel sentences to train the models. In the absence of parallel data, several approaches have turned to the use of images to learn translations. Since images of words, e.g., *horse* may be unchanged across languages, translations can be identified via images associated with words in different languages that have a high degree of visual similarity. However, translating via images has been shown to improve upon text-only models only marginally. To better understand *when* images are useful for translation, we study *image translatability* of words, which we define as the translatability of words via images, by measuring intra- and inter-cluster similarities of image representations of words that are translations of each other. We find that images of words are not always invariant across languages, and that language pairs with shared culture, meaning having either a common language family, ethnicity or religion, have improved image translatability (i.e., have more similar images for similar words) compared to its converse, regardless of their geographic proximity. In addition, in line with previous works that show images help more in translating concrete words, we found that concrete words have improved image translatability compared to abstract ones.

1 Introduction

Neural machine translation (NMT) for low-resource languages has drawn a lot of attention due to the increasing awareness of the lack of linguistic and geographic diversity in NLP research (Joshi et al., 2020; Orife et al.). Since parallel data for these languages is scarce, it necessitates the use of other data to help translation e.g., monolingual texts in unsupervised MT (Lample et al., 2018b,a,c; Artetxe et al., 2018) or images in multimodal MT (Barrault et al., 2018).

Previous works on using images for translation typically accept that images are useful due to their language invariance (Rotman et al., 2018). Since everyday words such as *chair* denote concepts that exist independently of any language, images that ground their meanings should also be invariant to the language. However, to the best of our knowledge, this conjecture on image-language invariance has never been tested. As images' usefulness for translation has only been shown to be marginal (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018), it is important to study this conjecture in relation to the characteristics of languages to understand *when* and to what extent images can aid translation. An alternative view would be that images may be different to some extent in different languages since they reflect the ways different people interact with these concepts; this may depend on *where* they live and the *communities* they live in (Evans and Levinson, 2009). For example, images of the word *breakfast* in different languages may reflect the different cuisines of the communities that speak the languages.

While most multimodal MT datasets are limited to a small set of European languages that come from the same language family, and are spoken by communities that are culturally and geographically close, the Massively Multilingual Image Dataset (MMID) (Hewitt et al., 2018) is constructed specifically to facilitate large-scale multilingual research in translating words via images.

MMID consists of up to 10K words and 100 images per word in 98 languages. This dataset provides an opportunity for us to examine how geographical and cultural relatedness between languages affect translation of words via images. As the use of parallel data from related languages have been found to improve MT for low resource languages (Zoph et al., 2016; Nguyen and Chiang, 2017; Dabre et al., 2017), we want to study if the same extends to translation via images. Specifi-

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 198–209
June 6–11, 2021. ©2021 Association for Computational Linguistics

cally, we want to explore if translatability of words between two languages via images is influenced by the cultural similarity and geographical proximity of their communities. A recent study, (Thompson et al., 2020), has observed such correlations of culture and geography to semantic alignment of word meanings between languages that are measured through similarities in the word embeddings. We hypothesize that the same is true for images, and that the alignment of meanings conveyed via images coincides with culture and geography.

In this work, we primarily define culture as the set of “Language, Norms and Beliefs" of a community (Heather Griffiths, 2015). These elements form our interpretation of cultural closeness between languages, which consist of their common linguistic, ethnic, and religious properties. Our goal is to intrinsically evaluate to what extent images can aid in word translation, for words in languages close in a variety of these characteristics. Assuming that each word is associated with a number of images that convey its meaning, we measure the degree to which images of words that are translations of each other in different languages have similar representations (thus will help in translation). We call this measure *image translatability* or the capacity of word meaning to be transferred from one language to another via images. If images are indeed language invariant, we should observe similar image translatability across different language pairs.

We identify how close word translations are in terms of their image representations (embeddings). Our findings suggest that languages with cultural similarity (defined as a combination of linguistic, ethnicity, or religious similarity of the communities at the cultural centres of the languages by Glottolog (Hammarström et al., 2020)) coincides with their translatability via images, and that the translatability of languages with cultural similarity outperforms that in those with geographical proximity.

Our paper is structured as follows: In section 2 we discuss previous research on image-aided word translation, and how roots, geography and cultural characteristics of languages correlate with semantic alignment of words. In section 3 we describe our dataset and text-image corpora. We also introduce the language pairs we examine and estimate their closeness in culture and geography. In section 4, we present our approach for measuring translatability of words in terms of the similarity of their image representations. Section 5 shows an analysis of our results and how translatability of words via images, which affects the images' fitness for translation, correlates with language properties. In Section 6, we discuss noteworthy examples that illustrate our findings, before concluding in Section 7.

2 Related Work

2.1 Translating Words via Images

This paper extends the work of Hewitt et al. (2018), which introduces a multi-lingual dataset of words in different languages, along with matching images, for word translation. Our goal however, is not to improve on the state-of-the-art methods in word translation using images, but to understand the specific characteristics of languages that influence the quality of translation via images. Hewitt et al. (2018) are the first to create a large-scale multilingual words and images dataset, without a specific part-of speech focus, proposing also a novel method to rate the concreteness of a word to be used in translations. Concreteness (Paivio et al., 1968) identifies tangible concepts and mental images that arise in correspondence to the word. Due to their strong visual representation, concrete words are easier to represent using images. Indeed, the measure of a word's concreteness has been observed to predict the effectiveness of its images to translate the word. (Kiela et al., 2015). A concept synonymous to concreteness is imageability (Kastner et al., 2020).

In terms of word translation, there exists a significant body of work in the area of bilingual lexicon induction, which is the task of translating words across languages without any parallel data (Fung and Yee, 1998; Rapp, 1999). Approaches can be divided into two types, text-based, which aim to find word translations by employing the words' linguistic information, and vision-based that use the words' images as pivots for translation (Bergsma and Van Durme, 2011; Kiela et al., 2015). Additionally, there are works that have incorporated additional signals for translation such as Wikipedia interlingual links (Wijaya et al., 2017).

The core idea in a large number of vision-based methods is using images to learn word and image embeddings that integrate all linguistic and visual information available to improve word translation (Calixto et al., 2017; Gella et al., 2017; Karpathy and Fei-Fei, 2017; Vulić et al., 2016). Recent research in this area extends prior ideas in learning multilingual word-image embeddings, extracting more complex and useful information from images, and applying the methods in few shot scenarios. Singhal et al. (2019) learn multilingual and multimodal word embeddings from weakly-supervised image-text data with a simple bag-of-words-based embedding model that incorporates word context and image information. Similarly, in Chen et al. (2019) the authors suggest mapping linguistic features based on sentence context and localized image features from image regions into a joint space. Aside from translation, multilingual text representations aligned to images has also been used to boost performance in vision-language tasks such as multilingual image-sentence retrieval (Gella et al., 2017; Wehrmann et al., 2019; Kim et al., 2020; Burns et al., 2020).

2.2 Language Characteristics in Translation

The claim that the concepts of language, culture and their geographical affiliations are interdependent, constantly and dynamically evolving and defining each other, has been widely discussed and is well established in the literature. Culture is considered an indistinguishable part of languages when translating from one language to another. The importance of cultural literacy of the translator and his/her awareness of cultural factors, views and tradition, apart from word meaning, for producing high quality translations is indisputable (Nida, 1945; Wakabayashi, 1991; Janfaza et al., 2012).

Despite the importance of language, culture and geography in translation, and findings that parallel data from similar, higher resource languages can help improve MT of low resource languages (Kocmi and Bojar, 2018), no previous work has studied how language similarity may influence translation via images. The most notable recent work in this area, that is most similar to ours, is that of Thompson et al. (2020). The authors predict semantic similarity of words in 41 languages from the NEL dataset (Dellert et al., 2020) and examine the relationships between word semantic similarity (measured via word embeddings) with the cultural, historical and geographical aspects of the communities speaking the language. Their findings, that the role of cultural similarities to this prediction is superior to that of geographical ones, align with ours. However, their methods differ from ours in many aspects. They use word-only embeddings to measure semantic alignment of words and only a small and publicly available set of images (Duñabeitia et al., 2018), for validation of the predicted scores, in a supervised manner, and for a small subset of 6 languages in the Indo-European family.

Language Pair		Similarity			
		Geography	Language	Ethnicity	Religion
az	tr	✓	✓	✓	✓
az	ru	✓	✗	✗	✗
ko	zh	✓	✗	✓	✓
ko	ja	✗	✗	✓	✓
zh	ja	✗	✗	✓	✓
zh	ko	✓	✗	✓	✓
ja	zh	✗	✗	✓	✓
ja	ko	✗	✗	✓	✓
ar	ur	✗	✗	✗	✓
ar	fa	✓	✗	✗	✓
ar	he	✓	✓	✗	✗
ur	ar	✗	✗	✗	✓
ur	hi	✓	✓	✓	✗
es	fr	✓	✓	✓	✓
es	pt	✓	✓	✓	✓
fi	hu	✗	✓	✓	✓
fi	no	✓	✗	✗	✓
af	nl	✗	✓	✓	✓
af	sw	✗	✗	✗	✓

Table 1: The 19 language pairs we explore in this work and the nature of their similarity: Geographical or Cultural: the same Language family, Ethnicity or Religion.

3 Data

3.1 The Massively Multilingual Image Dataset

The dataset we use is the Massively Multilingual Image Dataset (MMID) from Hewitt et al. (2018). It covers 98 languages, containing at most 10,000 words per language and 100 images per word. For each word, in any language, we are given the collected images matching the word meaning, and the word's English translation. They use a language filtering step to ensure that images for each language are collected only from web pages that are identified as containing texts written in the language.

We choose to examine specific language pairs so that for each source language there are two or more target languages whose shared characteristics with the source language differ in zero or more aspects. The shared characteristics between the source and target language include shared culture (i.e., either they are from the same language family[1] or the communities at their cultural centers have the same

[1] en.wikipedia.org/wiki/List_of_language_families

major ethnic group[2] or major religion[3]) or shared geography (i.e., the countries at their cultural centers share land border). For example, for Finnish, we include two target languages: one that has geographical proximity (Norwegian) and another that has ethnolinguistic similarity (Hungarian). In this way, we intend to examine for each source language, which of these groups of characteristics (culture or geography) are more important in image aided word translation, and whether culture or geography dominate one another. We form language pairs from the following 20 languages: Afrikaans (af), Arabic (ar), Azerbaijani (az), Chinese (zh), Dutch (nl), Finnish (fi), French (fr), Hebrew (he), Hindi (hi), Hungarian (hu), Japanese (ja), Korean (ko), Norwegian (no), Persian (fa), Portuguese (pt), Russian (ru), Spanish (es), Swahili (sw), Turkish (tr) and Urdu (ur). We summarize the language pairs and their shared characteristics in Table 1.

3.2 Dataset collection and preparation

We download MMID images[4] for all the source and target languages in our language pairs. In order to get vector embeddings for the images, we scale the images to 224 x 224 pixels, normalize and feed them as input into the ResNet-50 network (He et al., 2015), using network weights pre-trained on ImageNet. We obtain image embeddings from the last average pooling layer of ResNet-50, which gives us a 2048 dimensional vector embedding for each image. For each word, we call the embeddings of the associated images the word's image embedding. Because cosine similarity, which underlies parts of this work and previous works for bilingual lexicon induction via images (Bergsma et al., 2011; Kiela et al., 2015; Hewitt et al., 2018), is non-invariant to translation (Korenius et al., 2007) we treat all vectors with respect to the origin rather than some mean center for each image cluster[5].

Since the MTurk word translations that come with MMID (Pavlick et al., 2014) are limited in coverage and quality i.e., they contain only translations *to* English and the coverage and quality are high ($\geq$70% accuracy) only for a small set (13) of European and Indian languages where many MTurk workers are located; we create translation dictionaries for each of our language pairs using Google Translate, translating all words in the source language to the target language. We compute translatability of words whose translations have associated images in MMID. If a word in the source is translated to a phrase, we use the last word in the phrase to find associated images in the dataset. This heuristic applies to only 10% of the words in the dataset and the Google translations with the majority (80%) of the first word in the phrase translations being indicative of functional words: shared and appearing more than 50 times in the dataset.

4 Methodology

Given images that are associated with the word w_s in the source language s, and w_t in the target language t, we define two measures that determine how well a word can be translated by its images. The first measures whether w_s and w_t have overlapping or disjoint image embeddings. The second measures whether the spread of the image embedding for w_s, and, similarly, for w_t, is tight or loose: such a measure of image dispersion has been found to help predict the usefulness of image representations for translation (Kiela et al., 2014, 2015).

Specifically, when images of w_s and w_t are tight and overlapping in the embedding space, it shows that the images have little diversity (low dispersion) and are similar between w_s and w_t, indicating potentially good translation between them. Conversely, if the images are either spread out or disjoint, it means that the images have greater diversity (high dispersion) or differ between w_s and w_t, indicating potentially poor translation between them. We refer to the degree of overlap between two clusters of images associated with w_s and w_t respectively as their *inter-cluster similarity*, and to the degree of tightness or looseness of the images in each cluster as their *intra-cluster similarity*.

Our conjecture is that this is equivalent to representing image embeddings as samples from some generator distribution G. We can call the generator distribution for a given source word G_s and a generator distribution for a given target word G_t. Two words are translations of each other when $G_s = G_t$, and conversely two words are poor translations when $G_s \neq G_t$. Thus, *inter-cluster similarity* checks to see if an image embedding from G_s could have been produced by G_t. Note that this is a necessary condition but it is not sufficient to say $G_s = G_t$ if inter-cluster similarity is high, because an image embedding from G_s can also be produced by some random image embedding gen-

[2] en.wikipedia.org/wiki/Ethnic_group

[3] pewforum.org/2015/04/02/religious-projection-table/2020/percent/all/

[4] http://multilingual-images.org/

[5] Our code is available at https://github.com/nikzadkhani/MMID-CNN-Analysis

erator G_r with $G_s \neq G_r$. *Intra-cluster similarity* is a measure of how similar samples from a single generator are to each other. This will ensure that G_t and G_s are not random generators and are accurate representations of the word they are generating image embeddings for. In other words, having a high intra-cluster similarity implies that $G_s \neq G_r$ and $G_t \neq G_r$. Thus, we have sufficient conditions when we have high inter- and intra-cluster similarities to say $G_s = G_t$.

4.1 Inter-Cluster Similarity

To measure the degree of overlap (*inter-cluster similarity*) between images associated with the word w_s in the source language, and those associated with the word w_t in the target language, we first cluster their image embeddings with a *k*-means clustering algorithm (*k* = 2). Then, we measure the degree of overlap between images of the two words by the homogeneity score of the resulting clusters, $h_{w_s,w_t} \in [0,1]$ (Rosenberg and Hirschberg, 2007), calculated given the words w_s and w_t as image labels. A homogeneity score of 0 signals that all the image embeddings come from the distribution of a single class, hence represent the same word or concept ($w_s = w_t$). In this case, we say that the images of the two words have high inter-cluster similarity. A score of 1 means that the *k*-means clustering was able to identify two mutually exclusive clusters of images indicative that the images come from two different generators ($G_s \neq G_t$). In other words, the image embeddings were sampled from two different words or senses ($w_s \neq w_t$).

However, if images are highly dispersed (have high diversity, loose clusters), then the inter-cluster similarity may be deceptively high (i.e., low homogeneity) since loose clusters may overlap to some extent. Thus, homogeneity score is only an effective measure of how good an image-aided translation is on the condition that the clusters are sufficiently tight (i.e., have high intra-cluster similarity). In Section 4.5, we discuss how we compute this threshold for intra-cluster similarity.

4.2 Intra-Cluster Similarity

Images of a given word have low intra-cluster similarity when the images have high dispersion, which may be due to the word being abstract (e.g., words like *concept* whose images might be very diverse) or when the word has many different senses (e.g., words like *bug* whose images might represent the different senses of the word). On the other hand, when the intra-cluster similarity is high, it indicates that there is a general consensus on the meaning of the word as represented by the images, which makes for an easier transfer of the word meaning via images (i.e., better image translatability).

The metric we choose for the intra-cluster similarity of a word w is Median Max Cosine Similarity, which, given the set of images associated with the word, I_w, is:

$$\textsc{MedMax}_w = \underset{i \in I_w}{\text{median}} \max_{j \in I_w} \begin{cases} i \neq j : \text{cosine}(i,j) \\ i = j : 0 \end{cases}$$

This is a variation of the Average Maximum Cosine Similarity in Bergsma and Van Durme (2011), using the median to reduce the effect of outliers. Additionally, note that the worst case of this metric giving an undesirable outcome is when we have 50 random pairs of image embeddings for a given word cluster. This will result in a high *intra-cluster similarity* despite the randomness of the overall cluster. However, in our findings this scenario is extremely unlikely. As words have dominant senses, the effect of outliers is mitigated due to the use of the median.

However, intra-cluster similarity on its own is not enough to indicate if the word in the target language w_t is a good translation of the word in the source language w_s. For example, the word *train* may be represented with images of locomotives in one language and with images of people exercising in another, if its meaning differs across languages. Both of the words' images will have high intra-cluster similarity but low inter-cluster similarity, indicating poor translatability via images. Thus, intra-cluster similarity is only an effective measure of how good an image-aided translation is, on the condition that the inter-cluster similarity is sufficiently high. In Section 4.4, we discuss how we compute this threshold for inter-cluster similarity and in Section 4.6 how we combine intra- and inter-cluster similarity for image translatability.

4.3 Concreteness

To study the relationship between image translatability and concreteness of a word, we adopt a method similar to Hewitt et al. (2018) to train a model to predict word concreteness. We use the dataset provided by Brysbaert et al. (2014), consisting of 40,000 words that have been assigned concreteness scores by human judges, on a scale of 1 to 5, from abstract to concrete. We split the

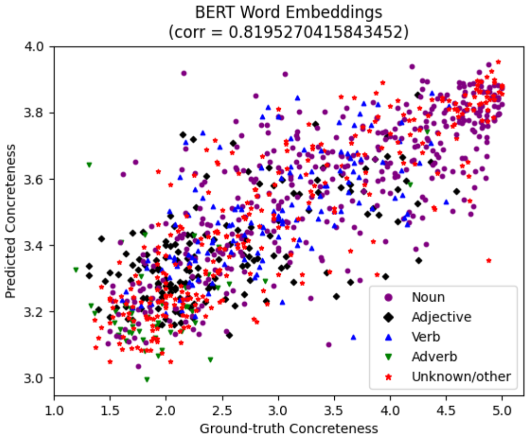

Figure 1: Distribution of concreteness scores predictions on the held-out validation set of 1,000 words from Brysbaert et al. (2014). The Spearman correlation coefficient calculated for ground-truth and predicted concreteness scores is noted.

dataset into train and test sets, randomly picking 39,000 words for training. Similar to Hewitt et al. (2018), our concreteness prediction model is a two-layer perceptron, with one 32-units hidden layer, and a ReLU activation function, trained with an L2 loss. For each word, the model input is the concatenation of the single word embeddings obtained from the top four hidden layers of BERT Devlin et al. (2019), a practice recommended as the best performing feature-extraction method by the authors.

Figure 1 shows the results of our evaluation on the test set of 1,000 words, depicting the distributions of the different part-of-speech categories. We provide the Spearman correlation coefficient between the ground-truth and predicted concreteness scores, which shows the improved effectiveness of our BERT embeddings-based method compared to the Salle et al. (2016) embeddings employed by Hewitt et al. (2018). Using this trained model, we predict the concreteness score of each of the words in our dataset by first translating the word to English and lemmatizing it using spaCy.

Using the predicted concreteness scores, we distinguish words in our dataset to concrete (having a predicted score of > 3) and abstract (≤ 3).

4.4 Inter-Cluster Similarity Threshold

We define a homogeneity score threshold to determine if two words w_s and w_t have sufficiently high overlap in their image embeddings to indicate a good translation. For each language pair, we compute this threshold $h^{\text{thld}}_{s,t}$ by taking the average homogeneity score of clusters of images of 10 randomly chosen word pairs from the source s and the target language t. We take the average of these scores since we want to first be able to compare the threshold with other homogeneity scores and second to be able to capture the skew in negative thresholds as well. These pairs serve as negative examples of translation and we expect their image embeddings to be disjoint. Hence, a word pair with homogeneity score *lower* than this threshold means that the word pair has a good overlap in their image embeddings (i.e., a high inter-cluster similarity), which indicates a good translation.

4.5 Intra-Cluster Similarity Threshold

Similarly, we define an intra-cluster similarity threshold to determine if an image cluster associated with a word w is sufficiently tight. Since intra-cluster similarity is computed for each word (and not word pair), we compute this threshold $\textsc{MedMax}^{\text{thld}}_{l}$ for each language l by constructing a negative example for the language i.e., an image cluster with a high dispersion. We create this negative example by taking five random words from the language and for each word a random sample of 20 images to build a cluster of 100 images (mimicking the typical image cluster size for a word in our dataset). We set the Median Max Cosine Similarity of this image cluster as the intra-cluster similarity threshold. A word that has an intra-cluster similarity *higher* than this threshold would mean that this word has a tight image cluster, a consistent meaning as represented in its images' representations.

4.6 Normalized Score

We define a normalized score $\textsc{Norm}_{w_s,w_t}$ to combine intra- and inter-cluster similarity scores for a word w_s in the source language and its translation in the target language w_t.

Given the intra-cluster similarity of word w_s ($\textsc{MedMax}_{w_s}$) and that of word w_t ($\textsc{MedMax}_{w_t}$); and the maximum and minimum intra-cluster similarities for the source language ($\textsc{MedMax}^{\text{max}}_{s}$ and $\textsc{MedMax}^{\text{min}}_{s}$), and those of the target language ($\textsc{MedMax}^{\text{max}}_{t}$ and $\textsc{MedMax}^{\text{min}}_{t}$); as well as the homogeneity score of the words h_{w_s,w_t} and the maximum and minimum homogeneity scores of words in the language pair ($h^{\text{max}}_{s,t}$ and $h^{\text{min}}_{s,t}$), we compute the normalized score $\textsc{Norm}_{w_s,w_t}$ as:

$$\begin{aligned}\textsc{Norm}_{w_s,w_t} = {} & \textsc{Norm}\ \textsc{MedMax}_{w_s} + \textsc{Norm}\ \textsc{MedMax}_{w_t} \\ & - \textsc{Norm}\ h_{w_s,w_t}\end{aligned}$$

where:

$$\textsc{Norm MedMax}_{w_s} = \frac{\textsc{MedMax}_{w_s} - \textsc{MedMax}_s^{\min}}{2(\textsc{MedMax}_s^{\max} - \textsc{MedMax}_s^{\min})}$$

$$\textsc{Norm MedMax}_{w_t} = \frac{\textsc{MedMax}_{w_t} - \textsc{MedMax}_t^{\min}}{2(\textsc{MedMax}_t^{\max} - \textsc{MedMax}_t^{\min})}$$

$$\textsc{Norm}\ h_{w_s,w_t} = \frac{h_{w_s,w_t} - h_{s,t}^{\min}}{h_{s,t}^{\max} - h_{s,t}^{\min}}$$

For each language pair, we also define a threshold on this normalized score (i.e., $\textsc{Norm}_{s,t}^{\text{thld}}$) by substituting $\textsc{MedMax}_{w_s}$, $\textsc{MedMax}_{w_t}$, and h_{w_s,w_t} with $\textsc{MedMax}_s^{\text{thld}}$, $\textsc{MedMax}_t^{\text{thld}}$, and $h_{s,t}^{\text{thld}}$ respectively in the equation above.

5 Results

In order to compare the image translatability of two language pairs with different characteristics, we compare the ratio of the number of word pairs that are good translations divided by the total number of word pairs in each language pair:

$$\textsc{Ratio} = \frac{\text{\# of word pairs that are good translations}}{\text{\# of word pairs for the language pair}}$$

A word pair w_s and w_t has a good translation via images (or good image translatability) if its homogeneity score h_{w_s,w_t} is lower than the homogeneity threshold $h_{s,t}^{\text{thld}}$ and its Median Max Cosine Similarities i.e., $\textsc{MedMax}_{w_s}$ *and* $\textsc{MedMax}_{w_t}$ are higher than the thresholds $\textsc{MedMax}_s^{\text{thld}}$ and $\textsc{MedMax}_t^{\text{thld}}$, respectively.

The higher this ratio, the more translatable the language pair is via images. When we compare two language pairs that have the same source language but different target languages (with different shared characteristics with the source), we can distinguish how different characteristics such as cultural similarity or geographical proximity affect image translatability. In Table 2, we show image translatability ratios of language pairs with the same source but different target languages side-by-side.

To understand the role of concreteness in translation via images, we also compute how many concrete words have good translations according to our image translatability measures. We consider a word pair, w_s and w_t, to be concrete if w_s has a concreteness score greater than 3. Source word concreteness is taken to as the pair concreteness, considering translation directionality. The ratio of how many concrete words in each language pair have good translations is also shown in Table 2.

Language Pair		**Number of Words**		**Ratio**	
		All	**Concrete**	**All**	**Concrete**
az	**tr**	4538	3470	**0.31**	**0.37**
az	ru	5380	2953	0.17	0.22
ko	zh	338	214	0.18	0.22
ko	**ja**	748	499	**0.69**	**0.72**
zh	**ja**	367	212	**0.56**	**0.58**
zh	ko	310	197	0.36	0.44
ja	zh	212	137	0.39	0.44
ja	ko	741	488	0.67	0.70
ar	ur	4916	3226	0.39	0.44
ar	fa	448	318	0.50	0.55
ar	**he**	2887	1874	**0.69**	**0.73**
ur	**ar**	4243	2466	**0.39**	**0.45**
ur	hi	4588	2817	0.12	0.15
es	fr	6392	3506	0.45	0.58
es	pt	7116	3920	0.40	0.53
fi	**hu**	5615	3190	**0.29**	**0.40**
fi	no	5336	3033	0.17	0.26
af	**nl**	5436	3247	**0.39**	**0.50**
af	sw	4553	2611	0.25	0.31

Table 2: Language pairs along with the numbers of word pairs and their image translatability ratios, for all and concrete word pairs. In boldface we mark the pair that has the highest ratio among pairs with the same source language whose normalized scores are significantly different *and* whose ratio differences are high.

To test whether the difference in image translatability between language pairs that share the same source language (e.g., Finnish to Norwegian vs. Finnish to Hungarian) is statistically significant, we conduct a simple t-test between their normalized score distributions. The resulting p-values signal the difference between their distributions. Low p-values (< 0.05) indicate statistical significance and high variation between the distributions, while higher values suggest low variation and large similarities between the language pairs (Table 3).

From the t-test, we find that the difference in distributions of pairs that share the same source language is almost in all statistically significant (p-value < 0.05, Table 3) except for Japanese to Chinese vs. Japanese to Korean.

Of other pairs whose normalized score differences are statistically significant *and* whose differences in translatability ratios are high (boldfaced, Table 2), we observe that the language pair with the higher image translatability ratio (i.e., Azer-

Language Pair I		Language Pair II		p-value
az	tr	az	ru	2.26×10^{-25}
ko	zh	ko	ja	2.2×10^{-4}
zh	ja	zh	ko	10.6×10^{-5}
ja	zh	ja	ko	**0.43**
ar	ur	ar	fa	9.44×10^{-29}
ar	fa	ar	he	2.39×10^{-13}
ar	he	ar	ur	16.35×10^{-5}
es	fr	es	pt	4.3×10^{-47}
fi	hu	fi	no	2.66×10^{-104}
af	nl	af	sw	10^{-100}

Table 3: p-values of differences between normalized score distributions of language pairs that share the same source language. In boldface, we mark pairs with a high p-value, for which we cannot assume a significant difference in their normalized score distributions.

baijani to Turkish, Korean to Japanese, Chinese to Japanese, Arabic to Hebrew, Urdu to Arabic, Finnish to Hungarian, and Afrikaans to Dutch) is always the pair that shares cultural similarity (i.e., either similar language family, similar major ethnicity, or similar major religion) even when they have little to no geographical proximity. For example, between Arabic and Hindi, Urdu's words are more translatable via images to Arabic (whose speakers share the same major religion as speakers of Urdu), despite Pakistan's geographic proximity to India. Similarly, Finnish words are more translatable via images to Hungarian (whose speakers belong to the same ethnolinguistic group as speakers of Finnish) than to Norwegian, despite Hungary not sharing any land border with Finland. In addition, there may be other language relatedness factors that result in better image translatability between languages, such as the similar writing system of Chinese and Japanese, or the similar grammatical structure of Korean and Japanese; despite their different language families.

From Table 1 we can see that Spanish shares similar characteristics with both French and Portuguese. Similarly, Japanese shares similar attributes with both Chinese and Korean (matching ethnicity and religion, different geography and language family). In such cases, where the two language pairs do not differ in characteristics, we observe that the difference in their translatability ratios is either small (in the case of Spanish to French and Spanish to Portuguese ratios in Table 2) or insignificant (in the case of Japanese to Chinese and Japanese to Korean p-value in Table 3).

On the contrary, Korean to Japanese and Korean to Chinese pairs, and Chinese to Korean and Chinese to Japanese pairs, have at least one difference in their attributes, accounting for the pairs' results statistical significance.

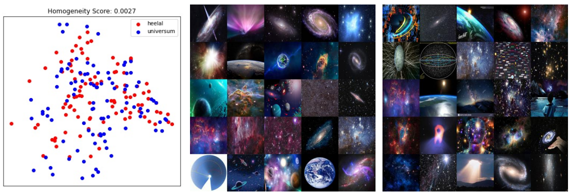

Figure 2: (left) PCA plot of image embeddings of the word *universe* in Afrikaans (*heelal*) and in Dutch (*universum*) showing their tightly overlapping image clusters, (center) images of *heelal* in Afrikaans, (right) images of *universum* in Dutch.

In addition, we observe that concreteness of words largely affects the quality of translations due to the low diversity in its image representations, which facilitate translation between words. On average, across language pairs, 62.4% of words with normalized scores above the threshold are concrete, while only 37.6% are abstract. At the same time, in Table 2 we see that the translatability ratio is considerably higher for concrete word pairs than all pairs. This supports our idea, and other previous works, that concrete words are better represented visually and, so, more likely to have good image-aided translations, compared to abstract ones.

6 Discussion

Our work has identified that language relatedness affect word translations via images. We observe that languages with cultural relatedness have better image translatability; suggesting that cultural relatedness should be taken into account when using images to aid translation. The image translatability measures we have defined can be used to identify a potentially good or poor translation or discover a cultural similarity or disconnect between words in two languages. For example, a word pair that has a high intra-cluster similarity and a high inter-cluster similarity in their image representations indicates that the image clusters are tight and overlapping, signaling a good translation between them. For example, the word *heelal* in Afrikaans and the word *universum* in Dutch have tight and overlapping image clusters (i.e., low homogeneity score) as can be seen in the PCA plot of their image embeddings and in their images (Figure 2).

On the other hand, when a word pair has tight but disjoint image clusters, it can mean that their

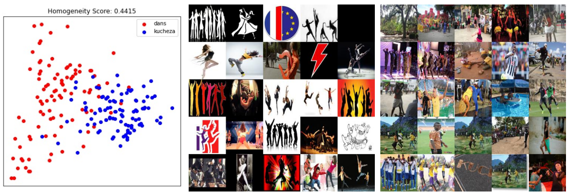

Figure 3: (left) PCA plot of image embeddings of the word *dance* in Afrikaans (*dans*) and in Swahili (*kucheza*) showing their tight but disjoint image clusters, (center) images of *dans* in Afrikaans, (right) images of *kucheza* in Swahili.

images express different meanings of the word. For example, the word *dance*: *dans* in Afrikaans and the word *kucheza* in Swahili have tight but disjoint image clusters (i.e., high homogeneity score) as can be seen in the PCA plot of their image embeddings and in the images (Figure 3), as *kucheza* means both to *play* and to *dance* in Swahili.

We observe that Afrikaans, for example, has a higher image translatability to Dutch due to their cultural (ethnolinguistic) similarity, than to Swahili, despite the relative distances of their cultural centers–South Africa is more distant geographically to the Netherlands than to Tanzania, defined in Glottolog as being the cultural center of the Swahili language. We observe higher visual similarities between words that are translations of each other in Afrikaans and Dutch than in Afrikaans and Swahili. For example, images of the word *park* in Afrikaans are more visually similar to images of the word *park* in Dutch than to images of the word *park* in Swahili (*hifadhi*) (Figures 4, 5). The images of *park* in Afrikaans and in Dutch refer to a Western style park, while its images in Swahili refer more to a wildlife reservation, a culturally different representation of the word *park* that is potentially influenced by how speakers of the different languages interact with the concept of the word. Interestingly, such connotation that is apparent in images may not be apparent in word embeddings, since *hifadhi* is used similarly with *park* in the texts of the language.

7 Conclusions

In this paper, we study when images may be useful for translating words between two languages from the perspective of their cultural and geographical relatedness. We observe that translatability of words via images vary in different language pairs, with language pairs sharing cultural similarities having better image translatability.

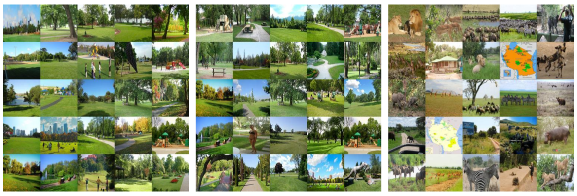

Figure 4: Images of the word *park* in (left) Afrikaans, (center) Dutch, (right) Swahili.

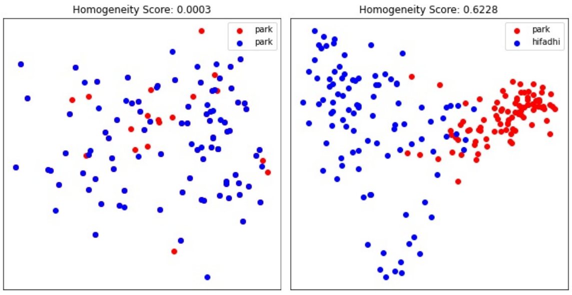

Figure 5: PCA plots for image embeddings of (left) *park* in Afrikaans to *park* in Dutch and (right) *park* in Afrikaans to *hifadhi* in Swahili.

In the future, it will be interesting to study image translatability of more language pairs and their characteristics, including those outside MMID, as well as extend our work to sentence-level image aided translation. It will also be of great value to study if adding considerations of cultural relatedness to image-aided MT can further improve its performance. Additionally, using a different metric for intra-cluster similarity that does not calculate similarity with respect to the origin may be more accurate depending on the application. As many similarity functions, aside from cosine similarity, have been used in the computer vision literature, improving this function could be fruitful future work.

Acknowledgments

We would like to thank the anonymous reviewers for their thoughtful comments. This work is supported in part by the U.S. NSF grant 1838193, DARPA HR001118S0044 (the LwLL program), and the Department of the Air Force FA8750-19-2-3334 (Semi-supervised Learning of Multimodal Representations). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA, the Air Force, and the U.S. Government.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels. Association for Computational Linguistics.

Shane Bergsma and Benjamin Van Durme. 2011. Learning bilingual lexicons using the visual similarity of labeled web images. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1764. Citeseer.

Shane Bergsma, David Yarowsky, and Kenneth Church. 2011. Using large monolingual and bilingual corpora to improve coordination disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1346–1355, Portland, Oregon, USA. Association for Computational Linguistics.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.

Andrea Burns, Donghyun Kim, Derry Wijaya, Kate Saenko, and Bryan A Plummer. 2020. Learning to scale multilingual representations for vision-language tasks. In *European Conference on Computer Vision*, pages 197–213. Springer.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Multilingual multi-modal embeddings for natural language processing. *CoRR*, abs/1702.01101.

Shizhe Chen, Qin Jin, and Alexander Hauptmann. 2019. Unsupervised bilingual lexicon induction from mono-lingual multimodal data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8207–8214.

Raj Dabre, Tetsuji Nakagawa, and Hideto Kazawa. 2017. An empirical study of language relatedness for transfer learning in neural machine translation. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 282–286. The National University (Phillippines).

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jon Andoni Duñabeitia, Davide Crepaldi, Antje S Meyer, Boris New, Christos Pliatsikas, Eva Smolka, and Marc Brysbaert. 2018. Multipic: A standardized set of 750 drawings with norms for six european languages. *Quarterly Journal of Experimental Psychology*, 71(4):808–816.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation*, pages 215–233, Copenhagen, Denmark. Association for Computational Linguistics.

Nicholas Evans and Stephen C. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5):429–448.

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. 2017. Image pivoting for learning multilingual multimodal representations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2839–2845, Copenhagen, Denmark. Association for Computational Linguistics.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2020. *Glottolog 4.3*. Jena.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Nathan Keirns Heather Griffiths. 2015. Introduction to sociology 2e. In *Introduction to Sociology 2e/Culture/elements-of-culture*, chapter 3. OpenStax, Oxford.

John Hewitt, Daphne Ippolito, Brendan Callahan, Reno Kriz, Derry Tanti Wijaya, and Chris Callison-Burch. 2018. Learning translations via images with a massively multilingual image dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2566–2576, Melbourne, Australia. Association for Computational Linguistics.

Elenaz Janfaza, A Assemi, and SS Dehghan. 2012. Language, translation, and culture. In *International Conference on Language, Medias and Culture*, volume 33, pages 83–87.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

A. Karpathy and L. Fei-Fei. 2017. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676.

Marc A Kastner, Ichiro Ide, Frank Nack, Yasutomo Kawanishi, Takatsugu Hirayama, Daisuke Deguchi, and Hiroshi Murase. 2020. Estimating the imageability of words by mining visual characteristics from crawled image data. *Multimedia Tools and Applications*, pages 1–33.

Douwe Kiela, Felix Hill, Anna Korhonen, and Stephen Clark. 2014. Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 835–841, Baltimore, Maryland. Association for Computational Linguistics.

Douwe Kiela, Ivan Vulić, and Stephen Clark. 2015. Visual bilingual lexicon induction with transferred ConvNet features. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 148–158, Lisbon, Portugal. Association for Computational Linguistics.

Donghyun Kim, Kuniaki Saito, Kate Saenko, Stan Sclaroff, and Bryan Plummer. 2020. Mule: Multimodal universal language embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11254–11261.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Brussels, Belgium. Association for Computational Linguistics.

Tuomo Korenius, Jorma Laurikkala, and Martti Juhola. 2007. On principal component analysis, cosine and euclidean measures in information retrieval. *Information Sciences*, 177(22):4893–4905.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018b. Word translation without parallel data. In *International Conference on Learning Representations*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018c. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Eugene Nida. 1945. Linguistics and ethnology in translation-problems. *WORD*, 1(2):194–208.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, et al. Masakhane–machine translation for africa. *arXiv preprint arXiv:2003.11529*.

Allan Paivio, John C Yuille, and Stephen A Madigan. 1968. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of experimental psychology*, 76(1p2):1.

Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The language demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2:79–92.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 519–526, College Park, Maryland, USA. Association for Computational Linguistics.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic. Association for Computational Linguistics.

Guy Rotman, Ivan Vulić, and Roi Reichart. 2018. Bridging languages through images with deep partial canonical correlation analysis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 910–921, Melbourne, Australia. Association for Computational Linguistics.

Alexandre Salle, Aline Villavicencio, and Marco Idiart. 2016. Matrix factorization using window sampling and negative sampling for improved word representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 419–424, Berlin, Germany. Association for Computational Linguistics.

Karan Singhal, Karthik Raman, and Balder ten Cate. 2019. Learning multilingual word embeddings using image-text data. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 68–77, Minneapolis, Minnesota. Association for Computational Linguistics.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.

Bill Thompson, Seán G Roberts, and Gary Lupyan. 2020. Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, pages 1–10.

Ivan Vulić, Douwe Kiela, Stephen Clark, and Marie-Francine Moens. 2016. Multi-modal representations for improved bilingual lexicon learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 188–194, Berlin, Germany. Association for Computational Linguistics.

Judy Wakabayashi. 1991. Translation between unrelated languages and cultures, as illustrated by japanese-english translation. *Meta: journal des traducteurs/Meta: Translators' Journal*, 36(2-3):414–423.

Jonatas Wehrmann, Douglas M Souza, Mauricio A Lopes, and Rodrigo C Barros. 2019. Language-agnostic visual-semantic embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5804–5813.

Derry Tanti Wijaya, Brendan Callahan, John Hewitt, Jie Gao, Xiao Ling, Marianna Apidianaki, and Chris Callison-Burch. 2017. Learning translations via matrix completion. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1452–1463.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.