Title
Typology of topological relations using machine translation

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 45(45)

Authors
Meewis, Floor
Fourtassi, Abdellah
Dautriche, Isabelle

Publication Date
2023

Peer reviewed

Typology of topological relations using machine translation

Tanguy Tiran[1] (tanguy.tiran@etu.univ-amu.fr)
Floor Meewis[1] (floor.meewis@univ-amu.fr)
Abdellah Fourtassi[2] (abdellah.fourtassi@univ-amu.fr)
Isabelle Dautriche[1] (isabelle.dautriche@cnrs.fr)

[1]Aix Marseille Université, CNRS, LPC, Marseille, France
[2]Aix Marseille Université, Université de Toulon, CNRS, LIS, Marseille, France

Abstract

Languages describe spatial relations in different manners. It is however hypothesized that highly frequent ways of categorizing spatial relations across languages correspond to the natural ways humans conceptualize them. In this study, we explore the use of machine translation to gather data in semantic typology to address whether different languages show similarities in how they carve up space. We collected spatial descriptions in English, translated them using machine translation, and subsequently extracted spatial terms automatically. Our results suggest that most spatial descriptions are accurately translated. Despite limitations in our extraction of spatial terms, we obtain meaningful patterns of spatial relation categorization across languages. We discuss translation limits for semantic typology and possible future directions.

Keywords: Semantic typology; machine translation; spatial relations; semantic universals

Introduction

Languages vary in the way they encode thoughts, leading to differences in how – and even which – concepts are expressed. Such differences can be found in the domain of topological relationships, i.e., non-perspectival spatial relations which include relations such as support, containment, and proximity (Levinson & Wilkins, 2006). For instance, whereas English conflates horizontal and vertical support under a single term ("*on*"), Dutch employs a finer-grained distinction and uses different terms ("*op*" and "*aan*", respectively) for both of these contact relationships (Bowerman & Choi, 2001).

How do these semantic variations in spatial language relate to cognition? One possibility is that highly frequent ways of categorizing spatial relations across languages correspond to the natural ways humans conceptualize them; a proposition put forward as the typological prevalence hypothesis (Gentner & Bowerman, 2009). To test this hypothesis, Levinson and Meira (2003) elicited spatial terms from native speakers of 9 languages who were asked to describe 71 pictures covering a wide range of topological relationships (Topological Relations Picture Series or TRPS, Bowerman & Pederson, 1992). They found that, while the 9 languages under study did not share the same basic topological categories, they did tend to organize their semantics around conceptual attractors. Those attractors included notions of attachment, containment, superposition, as well as a category conflating "near" and "under" relations. These results, while incompatible with an absolute universal position, hint nevertheless at the existence of a distributional universal in the domain of topological relationships. Such a distributional universal could be linked to the existence of naturally existing conceptual categories, in line with the typological prevalence hypothesis.

Linguistic universals, or the absence of thereof, are often justified by cross-linguistic analyses. Levinson and Meira (2003) for instance, chose to study 9 languages that belong to distinct language families in order to control for any similarity that could have been accounted for by historical proximity. However, while it is desirable for typologists to investigate genetically diverse languages - as languages exert influence over each other, including as many languages as possible may also be important - as some differences may exist even between closely related languages (as exemplified earlier with the differences between English and Dutch). Yet, while several studies have investigated topological relations (e.g., Beekhuizen & Stevenson, 2015; Bowerman & Choi, 2001; Levinson & Meira, 2003), hardly any has done so with a large set of languages. An important reason behind this shortcoming is the substantial time and monetary expenditure that is required to conduct this type of research.

We aim at bypassing those limits and target a wide range of languages by exploring the use of a novel method in semantic typology: automatic translation. Despite improvements in recent years, machine translation has, to our knowledge, only been used once to study spatial terms before, and on single words only (Strickland & Chemla, 2018).

We aim here at using automatic translation to translate sentences containing spatial descriptions in many languages. Critically, neural-based models use context to translate an utterance and appear to be able to pick up the necessary information from the rest of the sentence to accurately translate a spatial description. For instance, consider the two English sentences "the picture is on the wall" and "the picture is on the table", which use the same term ("*on*") to express two different relations. A machine translation model such as Google Translate distinguishes these two relations when translating both sentences to a language that expresses them differently (producing in Dutch, for example, "De foto hangt *aan* de muur" and "De foto staat *op* tafel").

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

Our goals are (1) to assess whether automatic translation can be a valid tool for semantic typology and if so, (2) to ultimately provide a typology of topological relations for a large set of languages. Such typology could answer whether (distributional) universal conceptual biases govern the way we carve up the semantic space.

To achieve these goals, we plan to (1) elicit spatial descriptions in English with the help of the TRPS, (2) translate the elicited sentences in other languages and extract the spatial terms within those translations, and (3) evaluate in a sample of languages the quality of both translations and extractions. Lastly (4), we will use clustering analyses to see whether different TRPS pictures are described by similar terms across the different languages.

Acquirement of the translated data

Crowdsourcing elicitation data

We conducted an elicitation task in which English participants were asked to describe pictures using spatial terms. The reasons why we gathered spatial elicitations ourselves instead of relying on available data were: (1) that we needed sufficient elicitations for each picture to capture the wide range of ways of describing spatial relations, and (2) that we needed those elicitations to be in a standardized format to be able to translate them efficiently.

The experiment was hosted online. Beekhuizen and Stevenson (2015) have compared data obtained online with data obtained in-person and have shown that crowdsourcing is an appropriate method to obtain data in semantic typology. In total, 509 English speakers agreed to take part in the experiment, which was hosted on the Amazon Mechanical Turk crowdsourcing platform.

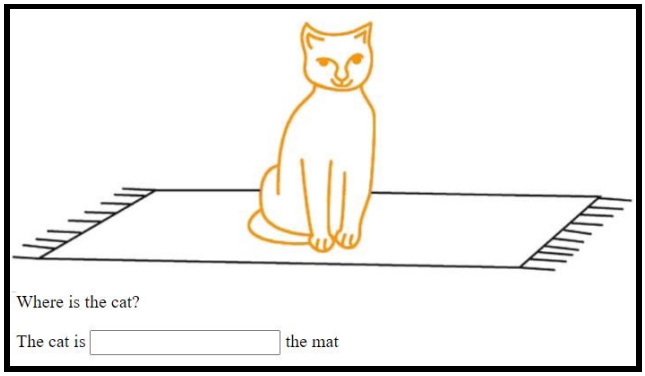

Figure 1: Example of a picture that participants had to describe in the elicitation task.

The original TRPS was used as stimuli for elicitation. The 71 pictures were divided into seven subsets, each containing 10 pictures (except for one subset which contained 11 pictures). Participants saw all pictures of a subset in random order, and were asked for each picture to answer the question "where is the [yellow object/figure]" using the first words that came to their mind (example provided in Fig. 1). Participants typed in their answers, filling in sentences of the kind "The *figure* is [BLANK] the *ground*". This standardized format allowed us to translate sentences efficiently.

Participants' data were rejected when participants wrote the same answer for every picture (N=17), when it was clear they did not do the task (e.g., some forms only contained numbers or random words) (N=15), when they tried participating more than once for the same subset of pictures (N=3) or when participants declared speaking more than one language at the end of the experiment (N = 155), as the experiment was intended for monolingual English speakers only. This left a total of 377 participants (66.5% of the initial pool of participants), yielding on average 55 elicitations for each picture (min=49, max=60).

Data was then cleaned by fixing clear typos (e.g., "hangnig on") and removing elicitations that did not fit grammatically in the sentences (e.g., "The cup is top the table"). Even though we constrained the answers heavily and asked participants to only complete sentences with spatial terms, many participants failed to do the task properly. This is comparable to previous studies which suggest that Mechanical Turk has experienced a decrease in participants responses' quality in recent years (e.g., Chmielewski & Kucker, 2020).

Finally, in order to cross-check participants' answers, two native English speakers assessed whether the elicited terms were semantically appropriate for each picture. The consultants judged at least one spatial expression as correct for each picture (min=1, max=8, mean=3.1). Note that the two consultants agreed on 82% of the proposed spatial terms (Cohen $\kappa = 0.64$) suggesting that there is some variability on how to describe the spatial relations depicted in the TRPS.

Translation of the elicited spatial descriptions

In order to understand how spatial relations are described across languages, the elicited sentences were translated in all languages available on Google Translate (GT) using the GT API in Python 3.9.5, in October 2022. GT is a state-of-the-art translation system which uses recent neural-based techniques (Wu et al., 2016). It was chosen for its accessibility and its extensiveness: the system offers support for more than a hundred languages belonging to 16 different language families.

Extraction of spatial terms within spatial descriptions

Because of the large number of languages in our sample, we sought to extract spatial terms from the translated sentences automatically.

We extracted spatial terms from the translated sentences using AWESoME, a recent alignment model which achieves state-of-the-art performance (Dou & Neubig, 2021). It uses multilingual BERT representations to align sentences of different languages and is available in more than 100 languages.

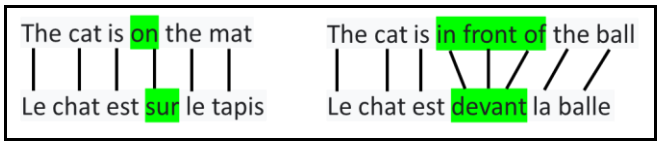

Figure 2: Illustration of alignment and extraction of spatial terms for two sentences.

AWESoME takes as input tokenized sentences (here, sentences split into words) in a source language (in this case, English) and a target language and matches their words (as in Fig. 2). Sentences were tokenized using mosestokenizer (Koehn et al., 2007) through the sacremoses library in python. The final list of investigated languages is the intersection between languages available on GT and languages available on AWESoME (minus Japanese, which was dropped because of tokenization issues). It includes 73 languages covering 14 different language families.

The spatial terms of the translated sentences were retrieved by extracting the words that aligned with the English spatial terms of the original sentences.

AWESoME's alignment performance ostensibly varied across languages. To infer whether it correctly aligned sentences, we used an alignment score for spatial terms for each language (henceforth referred to as the alignment score). We defined this score as the percentage of times AWESoME identified matching words for the English spatial terms in the translated sentences. Most languages revealed high alignment scores (mean≈86%) with alignments for spatial terms found in every sentence (most of these languages belong to the Indo-European family). However, alignment scores were considerably lower in some other languages (min≈57%). Most languages with low alignment performance turned out to be agglutinative, i.e., languages where spatial relations can be expressed through affixes appended to verbs or nouns. While the alignment scores help us hypothesize for which languages alignment could fail, it does not disclose whether the right spatial terms were extracted. Evaluating the extraction of spatial terms is consequently a necessity.

Evaluation of the translated data

We evaluated the sentence translations and the accuracy of the spatial term alignment in a sample of eight languages. Those languages were selected according to their genetic diversity and/or their morphological complexity, as well as native speaker availability.

Native speakers (one for each language) were shown the translated sentences and extracted spatial terms in their languages in a random order. They had to judge (1) if the (automatically translated) sentences were correct descriptions of their matching pictures (an evaluation of our translation tool, GT) and (2) if the spatial terms that were extracted using alignment were indeed correct spatial terms (an evaluation of our alignment tool, AWESoME).

Table 1 presents the evaluation results. In our sample, the evaluators judged most sentences as correct descriptions of their matching pictures in their native languages (except for Hindi) (mean percentage of sentences judged correct = 78%, sd = 14%). This indicates that GT properly translated most sentences in seven of our eight sampled languages.

We obtained more mitigated results on spatial terms extractions (mean of scores = 51%, sd = 23). Spatial terms were on average correctly retrieved in Dutch, Arabic, and to some extent, French. On the other hand, we obtained low scores in our sample's remaining languages.

We also found that patterns of errors varied across languages. In French and Hindi, most errors occurred when extracted terms did not contain all relevant words (e.g., "l'intérieur" instead of "à l'intérieur"/inside). Other languages exhibited the opposite pattern, with most errors corresponding to extractions containing more than only the relevant terms ("tableON" types of error for example). This was especially the case for agglutinative languages (e.g., Turkish or Hungarian). We consider other potential ways to extract spatial terms from translations in the discussion.

Table 1: Results of the translation and alignment evaluations in eight languages. Left: Percentage of translated sentences (by GT) judged correct by native speakers. Right: Percentage of extracted spatial terms (by AWESoME) identified as spatial terms by native speakers.

Language	Sentences (%)	Spatial terms (%)
Dutch	84	81
French	85	66
Hindi	48	28
Arabic	79	76
Hebrew	98	59
Turkish	78	42
Hungarian	78	42
Chinese	78	15

Clustering analysis

We investigated the extent to which different pictures are described by the same terms in each language. The method we follow here can be summarized in three steps.

Step 1: matrices representing the linguistic dissimilarity between pictures' description were computed for each language. Similarity between pictures *i* and *j* was computed as the number of spatial terms shared between them divided by the number of unique terms used for either picture.

$$D_{i,j} = \frac{|st_i \cap st_j|}{|st_i \cup st_j|}$$

Where st_k corresponds to the spatial terms used to describe picture *k*.

This measure resulted in 71 by 71 dissimilarity (or distance) matrices for each language where the value in a cell $D_{i,j}$ corresponded to the extent to which picture *i* and *j* differed in their linguistic treatment, with scores close to 1

meaning they were described using different spatial terms, and scores close to 0 meaning they were described using the same spatial terms.

Step 2: All matrices were weighted to mitigate the effect of genetic relatedness. The matrices of each language were multiplied by the inverse of the number of languages belonging to their family. They were subsequently added together and the final matrix was divided by the number of families in our set, so as to get values ranging between 0 and 1.

Step 3: We ultimately performed two statistical analyses on the final matrix. Hierarchical clustering (using the factoextra library (Kassambara & Mundt, 2017) in R) was applied to find clusters of pictures that were treated similarly across languages. Another statistical tool, Multidimensional Scaling (or MDS), was used to help visualize the linguistic treatment of pictures across languages. MDS takes as input the 71 by 71 matrices and reduces them to 71 two-dimensional points that can then be plotted in a cartesian plane. Because matrices represent distances between pictures, points that will be close on this plot will correspond to pictures that tend to be treated the same across languages, while points that are further apart will correspond to pictures seldomly described using the same terms.

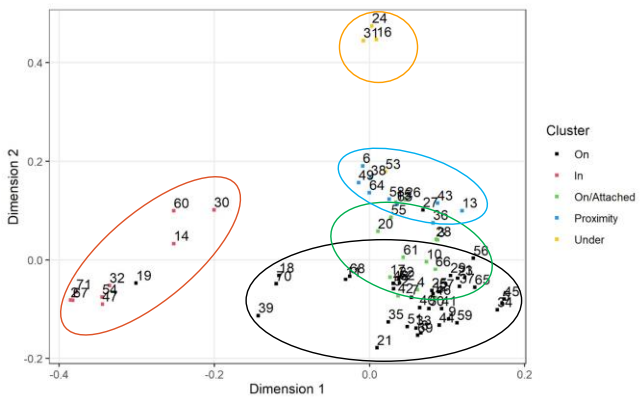

Figure 3: Pictures' similarity between how spatial relations are described across different languages plotted in a two-dimensional plane. Each point corresponds to a picture of the TRPS. Points that are close together indicate that they are described with the same spatial terms in different languages. Visual inspection revealed five clusters of pictures that can be described by similar spatial terms.

From Figure 3, we can see that pictures are not randomly distributed in the 2D space but tend to cluster together. We identified 5[1] clusters corresponding to containment ("In"), "Under", Proximity (e.g., "dog next to a doghouse", "lamp above a table"), Support, and a final cluster of relations – which we label "On/Attached" – where figures are attached to the ground and cannot be easily displaced (e.g., "face on a stamp", "strap attached to a bag"). Despite the poor quality of spatial terms extraction, it thus seems that our method is able to pick up cross-linguistic regularities in the way languages express topological relationships.

Discussion

In this project, we explored the use of machine translation to acquire data in semantic typology. We had the following two objectives: (1) assess whether automatic translation could be used to answer questions in semantic typology, and if so (2) take advantage of machine translation's assets to provide a typology of topological relations for a large set of languages.

Is automatic translation appropriate for semantic typology?

The results on translations are promising in all but one language (Hindi), with four out of five (79%) sentences being correctly translated on average, indicating they are valid descriptions of the TRPS pictures. Yet this leaves around 20% of translated sentences judged as incorrect. It should however be taken into account that only one speaker evaluated the translations for each language. It is thus possible that the consultants judged as incorrect some sentences that other speakers of those languages would have found correct. Recall that the two English consultants who helped construct the final English dataset only agreed on 82% of cases suggesting that inter-individual variability exists and is non-negligeable when expressing spatial relations. Overall, these results suggest that, while not being perfect, Google Translate, our translation tool, successfully managed to use context embedded in the source English sentences to pick appropriate spatial expressions in the target languages. Although we only evaluated translations in eight languages, these languages belonged to different typological families, thus we expect these conclusions to apply to the rest of languages available on GT.

Whilst the scores we obtain on translation are promising, results on the extraction of spatial terms are much more mixed: 51% of spatial terms are correctly extracted over the eight languages. The extraction performance varied greatly between languages and was poor in half of our evaluation sample's languages. Two possibilities may explain this result. First, mBert representations, which are used by AWESoME, are trained unevenly on each language (Wu & Dredze, 2020). AWESoME is thus maximally performant when used between pairs of high-resource languages. Secondly, one limit of our extraction technique is that we are unable to retrieve spatial affixes, which are especially common in agglutinative languages.

We have explored another way to extract spatial terms automatically and bypass the issues mentioned above. This other method took advantage of the constrained format of participants' responses in the crowdsourcing task. We created an additional dataset of English sentences by completing all sentences (e.g., "The cat is … the mat") with various spatial terms. We then translated those additional

[1] An objective method to find optimal number of clusters, the Silhouette method (Rousseeuw, 1987) suggested the data comprised six clusters. As one of the six clusters could not be interpretated, we display here results with five clusters.

sentences. Since for each picture, only the spatial terms varied across sentences, words that were less frequent in the translations were likely to be spatial terms. Importantly, this other extraction method (1) did not rely on a neural model and thus could perform similarly on low-resource and high-resource languages, and (2) allowed us to extract spatial affixes by looking at the frequencies of strings of letters instead of full words. Although this method did help retrieve spatial suffixes in agglutinative languages, it did not yield better results than alignment overall.

Despite its problem, it is noteworthy that our method yields interpretable clusters of topological relations across languages. This suggests that the analysis of categorization patterns could be robust to extraction errors when investigating a large number of languages. The errors may especially be randomly distributed across languages and pictures, and could be canceled out when investigating many languages. In this case, errors brought by our method could thus be compensated by the quantitative scale machine translation allows us to reach.

It is also worth mentioning that our method, while building on elicitations, has ties with another source of data that has been used in typology: parallel corpora, that is, texts which have been translated (usually not automatically) in different languages (e.g., Cysouw & Wälchli, 2007). Parallel corpora have been used to investigate semantic spaces across languages, especially in abstract domains (such as indefinite pronouns, as in Beekhuizen, Watson & Stevenson, 2017) where words (e.g., "someone" or "anything") are difficult to elicit.

In this work, by automatically translating English elicitation data, we have effectively created a parallel corpus of topological relationships. Our study thus relates to parallel corpora usage in semantic typology and bridges this strand of work to traditional elicitation studies. Additionally, machine translation presents potential benefits over these two methods. Firstly, it allows to collect data much faster than elicitation and across many languages. This data could also include abstract words that are difficult to study using elicitations only, and future research could explore in this direction. Secondly, this method is not restricted to scarcely available massively parallel texts (i.e., texts that are already translated in many languages). Moreover, automatically translated texts could be designed to be much more tailored to specific research topics, facilitating appropriate data collection.

However, a potential shortcoming of using automatic translation could be the presence of mistranslations (an already existing risk with parallel corpora studies). Ensuring the translated data is not akin to "*translationese*" and reflect how native speakers would have described the different pictures is thus a necessity, which we discuss in the following section.

Comparison with elicitation data

The translated data can only be used for semantic typology purposes if the translations adequately reflect how native speakers would have described the different spatial relations. Even though, on average, the translations correctly matched their accompanying picture, they might not necessarily correspond to how native speakers may spontaneously describe these pictures.

To test this, we asked native speaker consultants of six different languages (one for each language) to describe the TRPS pictures and compared these elicited descriptions to our translations. The consultants had to type a full sentence describing the picture they were shown and identify themselves the spatial words present in their descriptions. For each language, we computed three measures (recall, precision, and the F-score) comparing the elicited spatial terms with terms extracted manually (by the first author) from the translated sentences[2]. Recall is defined as the proportion of elicited spatial terms that were present in our translated data, while precision corresponds to the proportion of spatial terms in our translated data that were also elicited. F-score is the harmonic mean between these two metrics. An F-score of 1 would mean that the translated data perfectly matches the elicited spatial terms. The results of the comparison between those terms and the elicited terms are shown in Table 2.

Table 2: Comparison between the elicited spatial terms and manually extracted terms from the translations.

Language	Recall	Precision	F-score
Dutch	0.67	0.52	0.58
French	0.70	0.38	0.49
Russian	0.83	0.44	0.58
Turkish	0.57	0.36	0.44
Hungarian	0.71	0.41	0.52
Chinese	0.76	0.43	0.55
Mean	0.71	0.42	0.53

Our results in a sample of six languages suggest the two types of data moderately match (F-score = 0.53). The high recall rate indicates that the translations managed to effectively capture spatial terms that are used by native speakers. Notably, recall was much higher than precision across the six languages. This result can be explained by the fact that the translated data contained on average more sentences by language (M=219) than the elicited data (M=80). It is critical to note that there are some limitations to

[2] We also compared the elicited spatial terms with the automatically extracted spatial terms. Yet, as we described previously, the extraction method worked poorly on some languages, and this unsurprisingly leads to a low average F-score (M=0.30). While this strengthen our conclusions that spatial term extraction is challenging, this comparison cannot be used to evaluate whether our translated sentences match native speakers' spontaneous descriptions of the pictures as the spatial term, while not being extracted correctly, may still be present in the sentence.

the interpretation of these scores. Importantly, the elicitation data was gathered from only one speaker for each language and given the inter-individual variability for describing spatial relations, this might not reflect the full range of topological descriptions used for that given language.

In sum, this suggests that translation data, albeit not perfect, does reflect how native speakers use spatial terms.

Conclusions

In our study, we explored whether machine translation could be a valuable tool to derive a semantic typology of spatial relations.

Our findings indicate that current translation tools can fairly accurately translate elicitations of spatial relations from English to other languages. However, extracting spatial terms automatically from sentences remains a challenge. We found that alignment varies considerably across languages and performs especially badly on less-documented languages (a recurrent challenge for NLP tools) and languages that use spatial affixes. Further research could look into methods of improving extraction of words of interest especially for these kinds of languages.

Regardless of these concerns, our clustering revealed non-random patterns. While these non-random patterns are encouraging for the use of machine translation in semantic typology, they might however appear despite:

- A bias towards the source language (here, English). Translation could indeed be less accurate for languages which categorize spatial relations differently from English. Therefore, it remains unclear if translating from a single source language favors categorization patterns that are similar to that of the source language.
- Occasional translation errors. If randomly distributed across languages and pictures, these errors could nevertheless be canceled out when investigating many languages all together.

A future step we should take before interpretating clusters of pictures in depth would be to investigate these two possibilities.

Based on our study, we conclude that there are still some technological hurdles to take before machine translation can be effectively implemented to create a semantic typology of spatial relations. However, because our exploratory analyses showed some meaningful similarities across languages, we are positive that an improved method could prove a fruitful tool to answer whether universal biases exist in how we conceptually represent space.

Acknowledgments

This work was funded by the French National Agency for Research (ANR-20-CE28-0005).

We would like to thank everyone involved in making this study possible. We especially thank the consultants for their help in evaluating the sentences in English and in our sampled languages, as well as reviewers for their insightful comments.

References

Beekhuizen, B., & Stevenson, S. (2015). Crowdsourcing elicitation data for semantic typologies. In *CogSci*.

Beekhuizen, B., Watson, J., & Stevenson, S. (2017). Semantic Typology and Parallel Corpora: Something about Indefinite Pronouns. In *CogSci*.

Bowerman, M., & Choi, S. (2001). Shaping meanings for language: universal and language-specific in the acquisition of semantic categories. In M. Bowerman & S. C. Levinson (Eds.), *Language acquisition and conceptual development*. Cambridge: CUP.

Bowerman, M., & Pederson, E. (1992). Topological relations picture series. In *Space stimuli kit 1.2* (p. 51). Max Planck Institute for Psycholinguistics.

Chmielewski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, 11(4), 464-473.

Cysouw, M., & Wälchli, B. (2007). Parallel texts: Using translational equivalents in linguistic typology. *Language typology and universals*, 60(2), 95-99.

Dou, Z. Y., & Neubig, G. (2021). Word alignment by fine-tuning embeddings on parallel corpora. *arXiv preprint arXiv:2101.08231*.

Gentner, D., & Bowerman, M. (2009). Why some spatial semantic categories are harder to learn than others. The Typological Prevalence Hypothesis. In J. Guo et al. (Ed.), *Crosslinguistic approaches to the psychology of language. Research in the tradition of Dan Isaac Slobin*. New York: Psychology Press.

Kassambara, A., & Mundt, F. (2017). Package 'factoextra'. Extract and visualize the results of multivariate data analyses, 76.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., ... & Herbst, E. (2007, June). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions* (pp. 177-180).

Levinson, S. C., Meira, S., & The Language and Cognition Group. (2003). 'Natural concepts' in the spatial topological domain – Adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. *Language*, 79(3), 485–516.

Levinson, S. C., & Wilkins, D. P. (2006). The background to the study of the language of space. In *Grammars of space: Explorations in cognitive diversity*. Cambridge University Press.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20, 53-65.

Strickland, B., & Chemla, E. (2018). Cross-linguistic regularities and learner biases reflect "core" mechanics. *Plos one*, 13(1), e0184132.

Wu, S., & Dredze, M. (2020). Are all languages created equal in multilingual BERT?. *arXiv preprint arXiv:2005.09093*.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.