Dataset Geography: Mapping Language Data to Language Users

Anonymous ACL submission

Abstract

As language technologies become more ubiquitous, there are increasing efforts towards expanding the language diversity and coverage of natural language processing (NLP) systems. Arguably, the most important factor influencing the quality of modern NLP systems is data availability. In this work, we study the geographical representativeness of NLP datasets, aiming to quantify if and by how much do NLP datasets match the expected needs of the language speakers. In doing so, we use entity recognition and linking systems, also making important observations about their cross-lingual consistency and giving suggestions for more robust evaluation. Last, we explore some geographical and economic factors that may explain the observed dataset distributions.[1]

1 Introduction

The lack of linguistic, typological, and geographical diversity in NLP research, authorship, and publications is by now widely acknowledged and documented (Caines, 2019; Ponti et al., 2019; Bender, 2011; Adelani et al., 2021). Nevertheless, the advent of massively multilingual models presents opportunity and hope for the millions of speakers of under-represented languages that are currently under-served by language technologies.

Broadening up the NLP community's research efforts and scaling from a handful up to the almost 7000 languages of the world is no easy feat. In order for this effort to be efficient and successful, the community needs some necessary foundations to build upon. In seminal work, Joshi et al. (2020) provide a clear overview of where we currently stand with respect to data availability for the world's languages and relate them to the languages' representation in NLP conferences. Choudhury and Deshpande (2021) study how linguistically fair are multilingual language models, and provide a nuanced framework for evaluating multilingual models based on the principles of fairness in economics and social choice theory. Last, Blasi et al. (2021) provide a framework for relating NLP systems' performance on benchmark datasets to their downstream utility for users at a global scale, which can provide insights into development priorities; they also discuss academic incentives and socioeconomic factors that correlate with the current status of systematic cross-lingual inequalities they observe in language technologies performance.

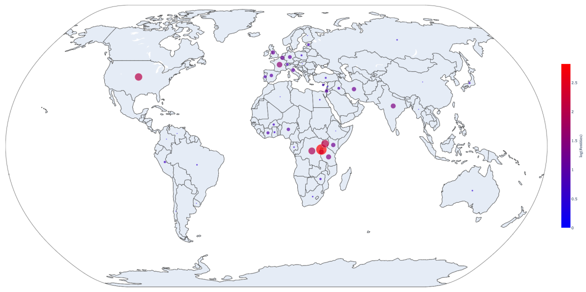


Figure 1: Example of the dataset map our method produces for the Kinyarwanda section of MasakhaNER. Despite its small size, the dataset is generally representative of Kinyarwanda speakers, including entity mentions from Rwanda and neighboring countries.

These works provide insights into current data availability and estimated utility that are paramount for making progress, as well as an evaluation framework for future work. However, there is one missing building block necessary for *real* progress: a way to estimate how representative of the underlying language speakers are our datasets. Any evaluation framework and any utility estimates we build can only be trustworthy as long as the evaluation data are representative.

We propose a method to estimate a dataset's representativeness by mapping it onto the physical space that language speakers occupy, producing visualizations such as Figure 1. Our contributions are summarized below:

[1] We will make our code and data publicly available.

- We present a method to map NLP datasets unto geographical areas (in our case, countries) and use it to evaluate how well the data represent the underlying users of the language. We perform an analysis of the socio-economic correlates of the dataset maps we create. We find that dataset representativeness largely correlates with economic measures (GDP), with geographical proximity and population being secondary.
- We test a simple strategy for performing entity linking by-passing the need for named entity recognition. We evaluate its efficacy on 19 languages, showing that we can get within up to 85% of a NER-informed harder-to-obtain model.
- We highlight the need for evaluating named entity recognition and linking models on parallel data in order to ensure cross-lingual consistency.

2 Mapping Datasets to Countries

Assumptions This work makes two assumptions: that (a) data locality matters, i.e., speakers of a language are more likely to talk about or refer to local news, events, entities, etc as opposed to ones from a different side of the world, and (b) that we can capture this locality by only focusing on entities. Kumar et al. (2019) discuss these *topical correlations* that are present in datasets,[2] noting that they exist and that L1 language identification models tend to pick up on them, i.e. if a text mentions Finland, a L1 langid model is probably going to predict that the speaker is Finnish, because $p(\texttt{Finland}|\texttt{Finnish})$ is generally high. While in that work Kumar et al. (2019) make explicit effort to avoid learning such correlations because they are interested in building models for $p(\texttt{L1}|\texttt{text})$ (i.e. $p(\texttt{Finnish}|\texttt{Finland})$) that are not confounded by the reverse conditional, the mere fact they need to do this confirms that real-world text has such topical confounds.

As for our second assumption that we can capture these topical correlations by only looking at entities, one need only take a look at Table 2 of Kumar et al. (2019), which lists the top topical confounding words based on log-odds scores for each L1 language in their dataset: all lists include either entities related to a country where that language is spoken (e.g. 'Merkel', the name of a former chancellor, for German) or topical adjectives (e.g. 'romanian' for Romanian).

[2] See §2 of their paper.

Approach For a given dataset, our method follows a simple recipe:
1. Identify named entities present in the dataset.
2. Perform entity linking to wikidata IDs.
3. Use Wikidata to link entities to countries.

We discuss each step below.

Entity Recognition Step Standard entity linking is treated as the sequence of two main tasks: entity recognition and entity disambiguation. One approach is to first process the text to extract entities and then disambiguate these entities to the correct entries of a given knowledge base (eg. Wikipedia). This approach relies on NER model quality.

However, to perform analysis on several datasets spanning several low-resource languages, one needs good-quality NER models in all these languages. As we show in Section §4, we can bypass this step if we tolerate a penalty in accuracy. Nevertheless, we revisit NER in our discussion of cross-lingual consistency (Section §5).

Entity Linking Step In this step we map named entities to their respective Wikidata IDs. We further discuss this step in Section §4.

From Entities to Countries We produce maps to visualize the geographical coverage of the datasets we study, discussing their properties and our findings in Section §3.

To link entities to countries,[3] we rely on Wikidata entries, depending on the type of entity:
- for persons, we log their place of birth (P19), place of death (P20), and country of citizenship (P27);
- for locations, we search for their associated country (P17); and
- for organizations, we use the links of the 'located_at' (P276) and 'headquartered_at' (P159) relations.

Since places of birth/death and headquarters are not necessarily at the country level, we perform a second step of associating these locations with countries. In cases where the result does not correspond to a modern-day country (as can often be the case with historical figures), we do not make any attempts to link it to any modern day countries.

3 Dataset-Country Maps

We apply the process described above on several datasets, chosen mostly for their language and typological diversity. Our process is not dataset- or

[3] A single entity can be associated with a set of more than one countries.

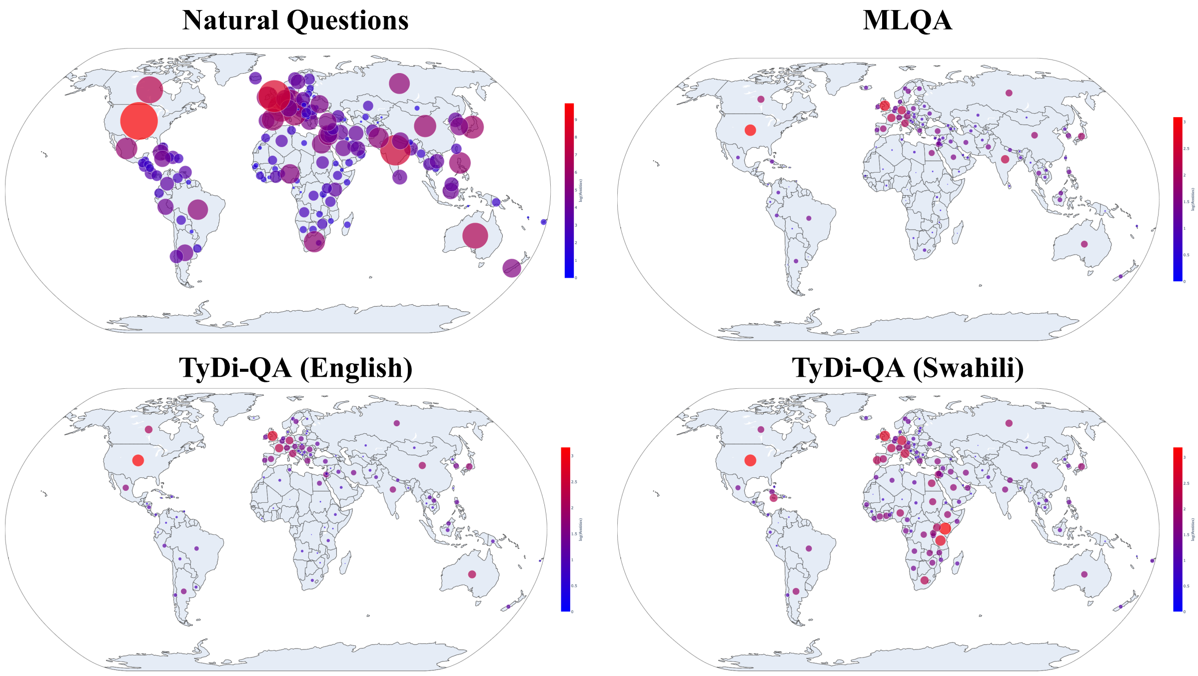


Figure 2: Visualizing the datasets' geography allows easy comparisons of their representativeness.

language-dependent,[4] and could easily be applied on any NL dataset. We briefly describe the datasets we include in our study below, with detailed statistics in Appendix C.

NER Datasets We study the WikiANN dataset (Pan et al., 2017) that is commonly used in the evaluation of multilingual models. We additionally study the MasakhaNER dataset (Adelani et al., 2021), which was created through participatory design (Nekoto et al., 2020a) in order to focus on African languages. Since these datasets are already annotated with named entities, we only need to perform entity linking.

Question Answering We study four question answering datasets (focusing on the questions rather than contexts), namely SQuAD (Rajpurkar et al., 2016), MLQA (Lewis et al., 2020), TyDi-QA (Clark et al., 2020), and Natural Questions (Kwiatkowski et al., 2019, NQ;), which have unique characteristics that lend themselves to interesting comparisons. SQuAD is a large English-only dataset (although it has been translated through efforts like XQuAD (Artetxe et al., 2020)). MLQA is a n-way parallel multilingual dataset covering 7 languages, created by translating an English dataset. TyDi-QA is another multilingual dataset covering 11 languages, but each language portion is derived separately for each language, without translating them. Last, NQ is an English QA dataset created based on real-world queries on the Google search engine for which annotators found relevant Wikipedia context, unlike the other datasets that were created by annotators forming questions *given* a context.

[4] Although it does rely on a decent quality entity linker which we lack for most languages. See discussion.

3.1 Discussion

We show example maps in Figure 1 (for the Kinyarwanda portion of the MasakhaNER dataset) and Figure 2 for NQ, MLQA, and two portions of TyDi-QA (English and Swahili). We provide additional maps for all other datasets in Appendix E.

Starting with the Kinyarwanda example of Figure 1, the utility of our method is apparent. Through the visualization, a researcher can quickly confirm that the dataset seems to reflect the users of the language: most entities indeed correspond to Rwanda, Uganda, Burundi, and to a lesser extent Congo, Tanzania, and Kenya (all neighboring countries). Wealthy or populous countries like USA, France, and India, are also represented, as one would expect. At the same time, the visualization allows a researcher to identify gaps: beyond the neighboring African countries, other African countries as well as central America or central/south-east Asia are clearly under-represented in the dataset.

Comparing datasets The comparison of MasakhaNER to the WikiANN dataset (see

Appendix E) reveals that the former is rather more localized (e.g. more than 80% of the identified entities in the Dholuo dataset are related to Kenya) while the latter includes a smaller portion from the countries were most native speakers reside (between 10%-20%) and almost always also includes several entries that are very European- or western-centric.

The effect of the participatory design (Nekoto et al., 2020b) approach on creating the MasakhaNER dataset, where data are curated from local sources, is clear in all language portions of the dataset, with data being highly representative of the speakers. In Figures 6–7 (App. E) it is clear that the majority of entities in e.g. the Wolof portion are from Cameroon and neighboring countries (as well as France, the former colonial power of the area), and the Yoruba and Igbo datasets are centered on Nigeria.

Figure 2 allows for a direct comparison of different QA datasets (also see maps for SQuAD in Figure 16 and other TyDi-QA languages in Appendix E). The first notable point has to do with NQ, which was build based on real-world English-language queries to the Google search engine. Since such queries happen all over the world, this is reflected in the dataset, which includes entities from almost all countries in the world. Two types of countries are particularly represented: ones were English is an official language (USA, UK, Australia, but also, to a lesser extent, India, Nigeria, South Africa, and the Philippines); and wealthy ones (European, Japan, China, etc). In our view, NQ is an exemplar of a representative dataset, because it not only includes representation of most countries where the language is spoken (with the sum of these entities being the overall majority, as one would expect) but due to its size it also includes entities from almost all countries.

On the other hand, the geographical representativeness of both MLQA and TyDi-QA (their English portion) is lacking. Since these datasets rely on Wikipedia articles for their creation, and Wikipedia is biased towards western countries (Greenstein and Zhu, 2012; Hube and Fetahu, 2018), most entities come from Europe, the US, and the Middle East. Both these datasets underrepresent English speakers from English-speaking countries of the Global South like Kenya, South Africa, or Nigeria, since there are practically almost no entities from these countries. MLQA further under-represents the speakers of all other languages it includes, since all data are translations of the English one. Contrast this to TyDi-QA and its visualized Swahili portion which, even though still quite western-centric, does have a higher representation from countries where Swahili is spoken (particularly ones from Kenya and Tanzania).

This discussion brings forth the importance of being cautious with claims regarding systems' utility, when evaluated on these datasets. One could argue that a QA system that is evaluated on NQ does indeed give a good estimation of real-world utility; a system evaluated on TyDi-QA gives a distorted notion of utility (biased towards western-based speakers and against speakers from the Global South); a system evaluated on MLQA will only give an estimation as good as one evaluated on TyDi-QA, but only on the English portion. We clarify that this does not diminish the utility of the dataset themselves as tools for comparing models and making progress in NLP: MLQA is extremely useful for comparing models across languages *on the exact same data*, thus facilitating easy comparisons of the cross-lingual abilities of QA systems, without the need for approximations or additional statistical tests. But we argue that MLQA should not be used to asses the potential utility of QA systems for German or Telugu speakers.

3.2 Socioeconomic Correlates

In this section we attempt to explain our findings from the previous section, tying them to socioeconomic factors.

Empirical Comparison of Factors We identify socioeconomic factors ϕ that could be used to explain the observed geographic distribution of the entities in the datasets we study. These are:

- a country's population ϕ_{pop}
- a country's gross domestic product (GDP) ϕ_{gdp}
- a country's geographical distance from country/ies where the language is spoken ϕ_{geo}

The first two factors are global and fixed.[5] The third one is relative to the language of the dataset we are currently studying. For example, when we focus on the Yoruba portion of the mTREx dataset, we use Nigeria (where Yoruba is spoken) as the focal point and compute distances to all other countries. The assumption here is that a Yoruba speaker is more likely to use or be interested in entities

[5] We also tested a factor that combines GDP and population: GDP per capita. However, its predictive power was significantly worse than using both factors separately.

	TyDi-QA (11)		MLQA (1)		SQUAD (1)		NaturalQ. (1)	
Factors ϕ	**Expl. Var.**	**MAE**	**Expl. Var.**	**MAE**	**Expl. Var.**	**MAE**	**Expl. Var.**	**MAE**
pop	0.272	0.431	0.317	0.401	0.277	1.230	0.395	1.18
gdp	0.507	0.349	0.561	0.332	0.516	1.023	0.535	1.069
geo	0.075	0.499	0.040	0.495	0.062	1.393	0.030	1.561
pop+gdp	0.477	0.352	0.528	0.336	0.495	1.034	0.528	1.041
pop+geo	0.304	0.417	0.360	0.385	0.347	1.129	0.433	1.137
geo+gdp	**0.550**	**0.333**	**0.579**	**0.321**	**0.552**	**0.932**	**0.550**	1.054
pop+gdp+geo	0.532	0.337	0.548	0.326	0.534	0.940	**0.550**	**1.005**

Table 1: Empirical comparison of factors on QA datasets, averaging over their respective languages (number in parentheses). We report the five-fold cross-validation explained variance and mean absolute error of a linear model.

first from their home country (Nigeria), then from its neighboring countries (Cameroon, Chad, Niger, Benin) and less likely of distant countries (e.g. Argentina, Canada, or New Zealand). Hence, we assume the probability to be inversely correlated with the country's distance. For macro-languages or ones used extensively in more than one country, we use a population-weighted combination of the factors of all relevant countries.

To measure the effect of such factors it is common to perform a correlational analysis, where one measures Spearman's rank correlation coefficient ρ between the dataset's observed geographical distribution and the factors ϕ. It is important to note, though, that the factors are potentially covariate, particularly population and GDP.[6] Hence, we instead compute the variance explained by a linear regression model with factors ϕ as input, i.e., $a\phi_{\text{pop}} + b\phi_{\text{gdp}} + c\phi_{\text{geo}} + d$ with a, b, c, d learned parameters, trained to to predict the log of observed entity count of a country. We report explained variance and mean absolute error from five-fold cross-validation experiments to avoid overfitting.

Socioeconomic Correlates and Discussion The results with different combination of factors for the QA datasets are listed in Table 1.[7] The best *single* predictor is, perhaps unsurprisingly, the GDP of the countries where the language is spoken: all datasets essentially over-represent wealthy countries (e.g. USA or Europe). A combination of geographical distance with GDP explains most of the variance we observe for all datasets, an observation that confirms the intuitions we discussed before based solely on the visualizations. Importantly, the fact that including population statistics into the model deteriorates its performance is further proof that our datasets are not representative of or proportional to the underlying populations. The only dataset that is indeed better explained by including population is the NQ one, which we already argued presents an exemplar of representativeness due to its construction protocol.

Limitations It is important to note that our assumptions are also limiting factors in our analyses. Mapping languages to countries is inherently lossy. It ignores, for instance, the millions of immigrants scattered throughout the world whose L1 language could be different than the dominant language(s) in the region where they reside. Another issue is that for many languages the necessary granularity level is certainly more fine that country; if a dataset does not include any entities related to the Basque country but does include a lot of entities from Spain and France, our analysis will incorrectly deem it representative.

An additional hurdle, and the reason why we avoid providing a concrete *representativeness score* or something similar, is that the ideal combination of factors can be subjective. It could be argued, for instance, that geographic proximity by itself should be enough, or that it should not matter at all. In any case, we share the coefficients of the NQ model, since it is the most representative dataset of those we study: $a = 0.9$ (for ϕ_{pop}), $b = 1.44$ (for ϕ_{gdp}), $c = 0.62$ (for ϕ_{geo}). We believe that ideally GDP should not matter ($b \rightarrow 0$) and that a combination of population and geographic proximity is ideal.

4 Bypassing NER for Entity Linking

We use mGENRE (Cao et al., 2021) for the task of multilingual entity linking, a sequence to sequence system that predicts entities in an auto-regressive manner. It works particularly well in a zero-shot setting as it considers 100+ target languages as latent variables to marginalize over.

Typically, the input to mGENRE can be in-

[6] See previous footnote.

[7] See Appendix F for NER datasets, and Appendix G for a breakdown by language for all datasets.

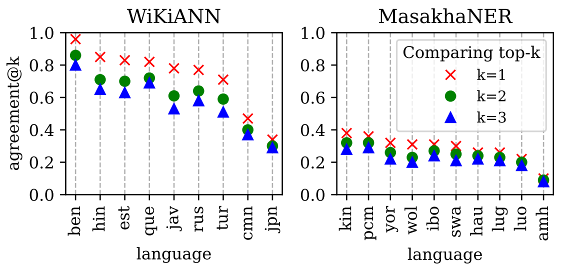


Figure 3: For some languages a NER-Relaxed model is within 60% of a NER-Informed model. agreement@k: ratio of top-k agreement of the models.

formed by a NER model that provides the named entity span over the source. For instance, in the Italian sentence `"[START] Einstein [END] era un fisico tedesco."` (*Einstein was a German physicist.*) the word `Einstein` is enclosed within the entity span. mGENRE is trained to use this information to return the most relevant Wikidata entries.

Due to the plasticity of neural models and mGEBRE's auto-regressive token generation fashion, we find that by simply enclosing the whole sentence in a span also yields meaningful results. In particular, for the previously discussed Italian sentence now the input to mGENRE is `"[START] Einstein era un fisico tedesco. [END]"`.

The advantage of this approach is two-fold. First, one does not need a NER component. Second, exactly because of bypassing the NER component, the EL model is now less constrained in its output; in cases where the NER component made errors, there's a higher chance that the EL model will return the correct result.

Consider the following example from the TyDi-QA Bengali training set: "প্রাগৈতিহাসিক [START] এশিয়ার ভৌগোলিক [END] আয়তন কেমন ছিল ?" (*'What was the [START] geographical [END] area of prehistoric [START] Asia [END]?'*. Our Bengali NER model trained on WikiANN with tuned parameters, returns `Asia` as an entity, as opposed to the, given the context, more appropriate `prehistoric Asia`. As a result, the entity linker fails to link this phrase to the corresponding WikiData entry `(prehistoric Asia, ID: Q4164212)`. When we instead remove these restrictions by simply passing "[START] প্রাগৈতিহাসিক এশিয়ার ভৌগোলিক আয়তন কেমন ছিল ? [END]' to the entity linker, it links to both `(Asia, ID: Q48)` and `(prehistoric Asia, ID: Q4164212)`.

Experiments and Results We conduct experiments to quantify how different a model uninformed by a NER model (NER-Relaxed) will perform compared to one following the typical pipeline (NER-Informed).

Given the outputs of the two models over the same set of sentences, we will compare their average agreement@k, as in the size of the intersection of the outputs of the two models divided by the number of outputs of the NER-Informed model, when focusing only on their top-k outputs.[8] We aggregate these statistics at the sentence level over the whole corpus. We focus on two datasets, namely WikiANN and MasakhaNER, summarizing the results in Figure 3.[9]

Comparing the general performance between these two datasets, it is clear that general agreement is decent. In 7 Out of 9 typologically diverse languages from WikiANN, more than 60% top-1 entities are linked by both models. The African languages from MasakhaNER are low-resource ones yielding less than 40% EL agreement to English in all cases. Given that most of these languages have not been included in the pre-training of BART (the model mGENRE is based on), we expect that using AfriBERTa (Ogueji et al.) or similar models in future work would yield improvements.

5 On the Cross-Lingual Consistency of NER/EL Models

Definition Bianchi et al. (2021) in concurrent work point out the need to focus on consistency evaluation of **language-invariant properties (LIP)**: properties which should not be changed via language transformation models. They suggest LIPs include meaning, topic, sentiment, speaker demographics, and logical entailment We propose a definition tailored to entity-related tasks: cross-lingual consistency is the desirable property that two parallel sentences in two languages, which should in principle use the same named entities (since they are translations of each other), are actually tagged with the same named entities.

5.1 NER Experiments

Models We study two models: SpaCy (Honnibal and Montani, 2017): a state-of-art monolingual library that supports several core NLP tasks; and a mBERT-based NER model trained on datasets from WikiANN using the transformers library (Wolf et al., 2020).

[8] Both models typically output between 1–3 entity links ranked according to their likelihood.

[9] An extensive results table is available in Appendix B.

Model	Greek	Italian	Chinese
Monolingual (SpaCy)	8.6	3.1	14.1
mBERT	**53.4**	**62.9**	**25.5**

Table 2: Using a multilingual NER model leads to significantly higher consistency tested on Eng–X data.

Training To task-tune the mBERT-based model on the NER task we use the WikiANN dataset with data from the four languages we study: Greek (el), Italian (it), Chinese (zh), and English (en).

Evaluation To evaluate cross-lingual consistency, ideally one would use parallel data where both sides are annotated with named entities. What we use instead, since such datasets do not exist to the best of our knowledge, is 'silver' annotations over parallel data. We start with unannotated parallel data from the WikiMatrix dataset (Schwenk et al., 2021) and we perform NER on both the English and the other language side, using the respective language model for each side.

We use the state-of-the-art AWESOME-align tool (Dou and Neubig, 2021) to create word-level links between the words of each English sentence to their corresponding translations. Using these alignment links for cross-lingual projection (Padó and Lapata, 2009; Tiedemann, 2014; Ni et al., 2017, *inter alia*) allows us to calculate cross-lingual consistency, measuring the portion of labels that agree following projection. In particular, we use the cross-lingual projections from the English side as 'correct' and measure precision, recall, and F-score against them.

Results For the three languages we study, the cross-lingual consistency of the monolingual SpaCy models is really low, with scores of 8.6% for Greek–English, 3.1% for Italian–English and 14.1% for Chinese–English. The SpaCy models are independently trained for each language and can produce 18 fine-grained NE labels e.g. distinguishing dates from time, or locations to geopolitical entities. As such, there was no a priori expectation for high cross-lingual consistency. Nevertheless, these extremely low scores reveal deeper differences, such as potentially widely different annotation protocols across languages.[10]

For the mBERT-based model we again label both sides of the parallel data, but now evaluate only on locations (LOC), organizations (ORG) and persons (PER) (the label types present in WikiANN). The mBERT models have significantly higher cross-lingual consistency: on the same dataset as above, we obtain 53.4% for Greek to English, 62.9% for Italian to English and 25.5% for Chinese to English.

Discussion To further understand the source of cross-lingual discrepancies, we performed manual analysis of 400 Greek-English parallel sentences where the mBERT-based model's outputs on Greek and the projected labels through English disagreed.[11] We sampled 100 sentences where the English-projected label was `O` but the Greek one was `LOC` (location), 100 sentences with English-projected as `LOC` but Greek as `O`, and similarly for persons (`PER`).

We performed annotation using the following schema:

- Greek wrong: for cases where only the English-side projected labels are correct
- English wrong: for cases where the English-side projected labels are wrong but the Greek-side are correct
- both wrong: for cases where the labels on both sides are incorrect
- alignment wrong: for cases where the two aligned phrases are not translations of each other, so we should not take the projected labels into account nor compare against them.
- all correct: both sides as well as the alignments are correctly tagged (false negatives).

Encouragingly, the entity alignments were wrong in less than 10% of the parallel sentences we manually labelled. This means that our results our quite robust: a 10%-level of noise cannot account for an almost 50% lack of consistency on the Greek-English dataset.[12] Hence, the system definitely has room for improvement. A second encouraging sign is that less than 2% of the cases were in fact false negatives, i.e. only one of the two sides actually contained an entity.

Going further, we find that mistakes vary significantly by label type. In about 75% of the `O-LOC` cases it was the Greek-side labels that were wrong in outputting `LOC` tags. A common pattern (about 35% of these cases) was the Greek model tagging months as locations. In the case of `O-PER` cases, 62% of the errors were on the English side. A

[10] We note that our evaluation does focus only on labels shared between models/languages.

[11] One of the authors is a fluent speaker of both languages.

[12] It does provide a potential upper bound of around 90% on the consistency we should expect to find.

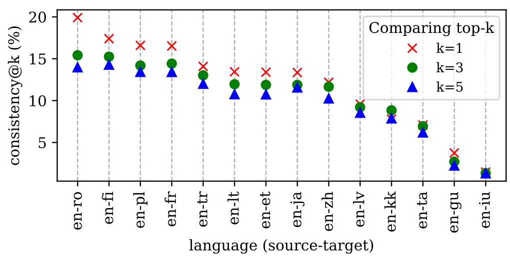


Figure 4: The entity linking cross-lingual consistency is generally low across languages, but especially for low-resource language pairs like English to Inuktitut (iu), Gujarati (gu), or Tamil (ta).

common pattern was the English-side model not tagging persons when they are the very first token in a sentence, i.e. the first token in `Olga and her husband [...].' Appendix I extends this discussion with additional details and examples.

The above observations provide insights into NER models' mistakes, which we were able to easily identify by contrasting the models' predictions over parallel sentences. We argue this proves the utility and importance of also evaluating NER models against parallel data even without gold NER annotations. Improving the NER cross-lingual consistency should in principle also lead to better NER models in general. Potential solutions could use a post-pretraining alignment-based fine-tuned mBERT model as the encoder for our data, or operationalize our measure of cross-lingual consistency into an objective function to optimize.[13]

5.2 Entity Linking Experiments

We now turn to entity linking (EL), evaluating mGENRE's cross-lingual consistency.

Dataset We use parallel corpora from the WMT news translation shared tasks for the years 2014 to 2020 (Bojar et al., 2014, 2015, 2016, 2017, 2018; Barrault et al., 2019, 2020). We work with 14 English-to-target language pairs, with parallel sentence counts in the range of around 1-5k.

Evaluation Unlike our NER experiment settings, we do not need word-level alignments to calculate cross-lingual consistency. We can instead compare the sets of the linked entities for both source and target sentences. In this manner, we calculate and aggregate sentence-level scores for the top-k linked entities for $k = 1, 3, 5$. In Figure 4, we present this score as a percentage, dividing the

[13] We leave this for future work, as it detracts off the main goal of this work (mapping datasets to the language users and measuring their representativeness).

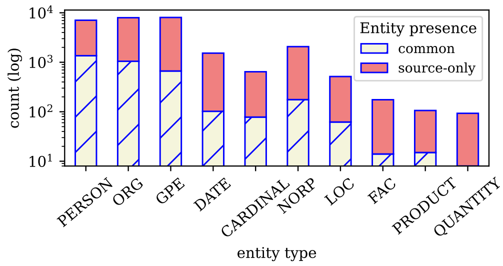


Figure 5: Counts of linked entity types across all WMT language pairs. Notice the y-axis log-scale: many entities are linked differently on non-English input.

size of the intersection (of the source and target sentence outputs) by the number of source sentence entities. Detailed results for all 14 language pairs are also reported in Appendix D.

Results As Figure 4 shows, we obtain low consistency scores across all 14 language pairs, ranging from 19.91% for English-Romanian to as low as 1.47% for English-Inukitut ($k = 1$). The particularly low scores for languages like Inuktitut, Gujarati, and Tamil may reflect the general low quality of mGENRE for such languages, especially because they use non-Latin scripts, an issue already noted in the literature (Muller et al., 2021).

The low percentage consistency scores for all languages makes it clear that mGENRE does not produce similar entity links for entities appearing in different languages. In future work, we plan to address this limitation, potentially by weighting linked-entities according to the cross-lingual consistency score when performing entity disambiguation in a multilingual setting.

Discussion We further analyze whether specific types of entities are consistently recognized and linked across language. We use SpaCy's English NER model to categorize all entities. Figure 5 presents a visualization comparing consistent entity category counts to source-only ones. See Appendix D for additional discussion.

6 Conclusion

We present a recipe for visualizing how representative NLP datasets are with respect to the underlying language speakers, and we analyze entity recognition and linking systems, finding they lack in cross-lingual consistency. We plan to further improve our tool by making NER/EL models robustly handle low-resource languages based on our observations. We will also expand our dataset and task coverage, to get a broader overview of the current utility of NLP systems.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. Masakhaner: Named entity recognition for african languages.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Rachel Bawden, Giorgio Maria Di Nunzio, Cristian Grozea, Inigo Jauregi Unanue, Antonio Jimeno Yepes, Nancy Mah, David Martinez, Aurélie Névéol, Mariana Neves, Maite Oronoz, Olatz Perez-de Viñaspre, Massimo Piccardi, Roland Roller, Amy Siu, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Dina Wiemann, and Lana Yeganova. 2020. Findings of the WMT 2020 biomedical translation shared task: Basque, Italian and Russian as new additional languages. In *Proceedings of the Fifth Conference on Machine Translation*, pages 660–687, Online. Association for Computational Linguistics.

Emily M Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6(3):1–26.

Federico Bianchi, Debora Nozza, and Dirk Hovy. 2021. Language invariant properties in natural language processing.

Damián Blasi, Antonios Anastasopoulos, and Graham Neubig. 2021. Systematic inequalities in language technology performance across the world's languages. arXiv:2110.06733.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Andrew Caines. 2019. The geographic diversity of nlp conferences.

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2021. Multilingual autoregressive entity linking.

Monojit Choudhury and Amit Deshpande. 2021. How linguistically fair are multilingual pre-trained language models? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12710–12718.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Shane Greenstein and Feng Zhu. 2012. Is wikipedia biased? *American Economic Review*, 102(3):343–48.

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

Christoph Hube and Besnik Fetahu. 2018. Detecting biased statements in wikipedia. In *Companion proceedings of the the web conference 2018*, pages 1779–1786.

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.

Sachin Kumar, Shuly Wintner, Noah A. Smith, and Yulia Tsvetkov. 2019. Topics to avoid: Demoting latent confounds in text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4153–4163, Hong Kong, China. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. Mlqa: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, et al. 2020a. Participatory research for low-resourced machine translation: A case study in african languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2144–2160.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020b. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1470–1480.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resource languages.

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36:307–340.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Jörg Tiedemann. 2014. Rediscovering annotation projection for cross-lingual parser induction. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1854–1864.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

A Related Work

One important aspect of our study is the evaluation of cross-lingual consistency while performing multilingual NER or El tasks. In (Bianchi et al., 2021), the authors focus on the consistency evaluation of language-invariant properties. In an ideal scenario, the properties should not be changed via the language transformation models but commercially available models are not prone to avoid domain dependency.

Effective measurement of dataset quality is another aspect of fast-growing significance. Training large language models require huge amount of data and as a result, the inference generated by these pretrained language model as well as the fine-tuned models often show inherent data bias. In a recent work (Swayamdipta et al., 2020), the authors present how data-quality aware design-decision can improve the overall model performance. They formulated categorization of data-regions based on characteristics such as out-of-distribution feature, class-probability fluctuation and annotation-level discrepancy.

Usually, multilingual datasets are collected from diverse places. So it is important to assess whether the utility of these datasets are representative enough to reflect upon the native speakers. We find the MasakhaNER (Adelani et al., 2021) is one such dataset that was collected from local sources and the data characteristics can be mapped to local users as a result. In addition, language models often requires to be truly language-agnostic depending on the tasks, but one recent work shows that, the current state-of-the-art language applications are far from achieving this goal (Joshi et al., 2020). The authors present quantitative assessment of available applications and language-resource trajectories which turns out not uniformly distributed over the usefulness of targeted users and speakers from all parts of the world.

B NER-Informed vs NER-Relaxed model

In this section, we report the detailed results (see table 3) from our experiment with using intermediate NER model vs skipping this step.

C Dataset Statistics

See details in Table 4.

D Cross-lingual consistency experiments

From Figure 5, it is clear that geopolitical entities (`GPE`) are the ones suffering the most from low cross-lingual consistency, with an order of magnitude less entities linked on both the English and the other language side. On the other hand, person names (`PER`) seem to be easier to link. While the most common types of entities are `PERSON`, `ORG` (i.e. organization) and `GPE` (i.e. geopolitical entity), we found that the NER model still failed to correctly categorize entities like (Surat, `Q4629`, `LOC`), (Aurangzeb, `Q485547`, `PER`). However, these entities were correctly linked by the NER-Relaxed pipeline, indicating its usefulness. We hypothesize, and plan to test in future work, that a NER-Relaxed entity further regularized towards cross-lingual consistency will perform better than a NER-Informed pipeline, unless the NER component also shows improved cross-lingual consistency.

Additionally, in Table 5, we report the detailed cross-lingual consistency score percentages for 14 english-language source-target pairs from WMT news translation shared tasks (Bawden et al., 2020).

E Additional Dataset Maps

We present all dataset maps for the datasets we study:

- MasakhaNER languages are available in Figures 6 and 7.
- TydiQA languages are available in Figures 8 and 9.
- WikiANN (panx) languages are available in Figures 10 through 15.
- SQuAD (English) in Figure 16.

F NER Dataset Socioeconomic Factors

Table 1 presents the same analysis as the one described in Section 3.2 for the X-FACTR and the NER datasets. The trends are similar to the QA datasets, with GDP being the best predictor and including population statistics hurting the explained variance.

Language	k=1	k=2	k=3	Dataset
hin	(4239, 761, 0.85)	(6765, 2717, 0.71)	(8377, 4436, 0.65)	
cmn	(9354, 10646, 0.47)	(16015, 23899, 0.4)	(21835, 37346, 0.37)	
jpn	(6739, 13259, 0.34)	(12148, 27820, 0.3)	(17220, 42463, 0.29)	
rus	(15325, 4675, 0.77)	(24663, 13989, 0.64)	(31520, 23051, 0.58)	
est	(16687, 3313, 0.83)	(24413, 10536, 0.7)	(28146, 16459, 0.63)	WikiANN
ben	(9575, 425, 0.96)	(15759, 2541, 0.86)	(20106, 4930, 0.8)	
que	(82, 18, 0.82)	(124, 48, 0.72)	(159, 72, 0.69)	
tur	(14206, 5794, 0.71)	(21165, 14999, 0.59)	(25053, 23597, 0.51)	
jav	(78, 22, 0.78)	(103, 67, 0.61)	(113, 101, 0.53)	
pcm	(549, 994, 0.36)	(955, 2033, 0.32)	(1217, 3030, 0.29)	
kin	(593, 952, 0.38)	(924, 1988, 0.32)	(1112, 2853, 0.28)	
wol	(242, 534, 0.31)	(350, 1158, 0.23)	(435, 1692, 0.2)	
hau	(417, 1178, 0.26)	(747, 2333, 0.24)	(941, 3402, 0.22)	
ibo	(494, 1093, 0.31)	(834, 2225, 0.27)	(1056, 3257, 0.24)	MasakhaNER
amh	(117, 1088, 0.1)	(210, 2184, 0.09)	(289, 3198, 0.08)	
swa	(499, 1175, 0.3)	(819, 2445, 0.25)	(1007, 3678, 0.21)	
lug	(283, 824, 0.26)	(486, 1657, 0.23)	(644, 2362, 0.21)	
yor	(430, 894, 0.32)	(673, 1909, 0.26)	(839, 2893, 0.22)	
luo	(122, 428, 0.22)	(207, 844, 0.2)	(264, 1184, 0.18)	

Table 3: Breakdown of entity extraction count while using NER-informed model. Here for each top k extracted entities, the triplet is the aggregated value of (count of common entities extracted by both ner-informed and ner-relaxed models, count of entities only extracted by ner-relaxed models, ratio of common entity count and total top-k extract by ner-relaxed model)

G Socioeconomic Correlates Breakdown

H NER Models Confusion Matrices

I Greek-English NER Error Discussion

We find that the mistakes we identify vary significantly by label. In about 75% of the `O-LOC` cases it was the Greek-side labels that were wrong in tagging a span as a location. A common pattern we identified (about 35% of these cases) was the Greek model tagging as location what was actually a month. For instance, in the sentence "Τον Μάιο του 1990 επισκέφτηκαν για τέσσερις ημέρες της Ουγγαρία." (*In May 1990 , they visited Hungary for four days.*) the model tags the first two words (“in May”) as a location, while the English one correctly leaves them unlabelled.

In the case of `LOC-O` cases, we found an even split between the English- and the Greek-side labels being wrong (with about 40% of the sentences each). Common patterns of mistakes in the English side include tagging persons as locations (e.g. “Heath” in “Heath asked the British to heat only one room in their houses over the winter.” where “Heath” corresponds to Ted Heath, a British politician), as well as tagging adjectives, often locative, as locations, such as “palaeotropical” in “Palaeotropical refers to geographical occurrence.” and “French” in “A further link [..] by vast French investments and loans [...]”.

Last, in the case of `O-PER` cases we studied, we found that 62% of the errors were on the English side. A common pattern was the English-side model not tagging persons when they are the very first token in a sentence, i.e. the first tokens in “Olga and her husband were left at Ay-Todor.”, in “Friedman once said, ‘If you want to see capitalism in action, go to Hong Kong.’ ”, and in “Evans was a political activist before [...]” were all tagged as `O`. To a lesser extent, we observed a similar issue when the person’s name followed punctuation, e.g. “Yavlinsky” in the sentence “In March 2017 , Yavlinsky stated that he will [...]”.

Dataset	Data-split	Languages	Language count	Sentence count
WikiANN	train	russian, polish, kazakh, bulgarian, finnish, ukrainian, afrikaans, hindi, yoruba, hungarian, dutch-flemish, korean, persian, japanese, javanese, portuguese, hebrew, arabic, spanish-castilian, bengali, urdu, indonesian, tamil, english, malay-alam, tagalog, basque, thai, german, romanian-moldavian-moldovan, chinese, telugu, azerbaijani, quechua, modern-greek, turkish, marathi, georgian, estonian, italian, panjabi, burmese, french, gujarati, malay, lithuanian, swahili, vietnamese	48	658600
TyDi-QA	train	english, korean, japanese, telugu, russian, thai, arabic, finnish, bengali, swahili, indonesian	11	166905
MasakhaNER	train	igbo, wolof, nigerian pidgin, kinyarwanda, amharic, hausa, yoruba, ganda, swahili, dholuo	10	12906
SQuAD	train	english	1	130319
MLQA	dev, test	english, simplified chinese, german, arabic, spanish, hindi, vietnamese	7	12738
Natural Questions	train	english	1	307373

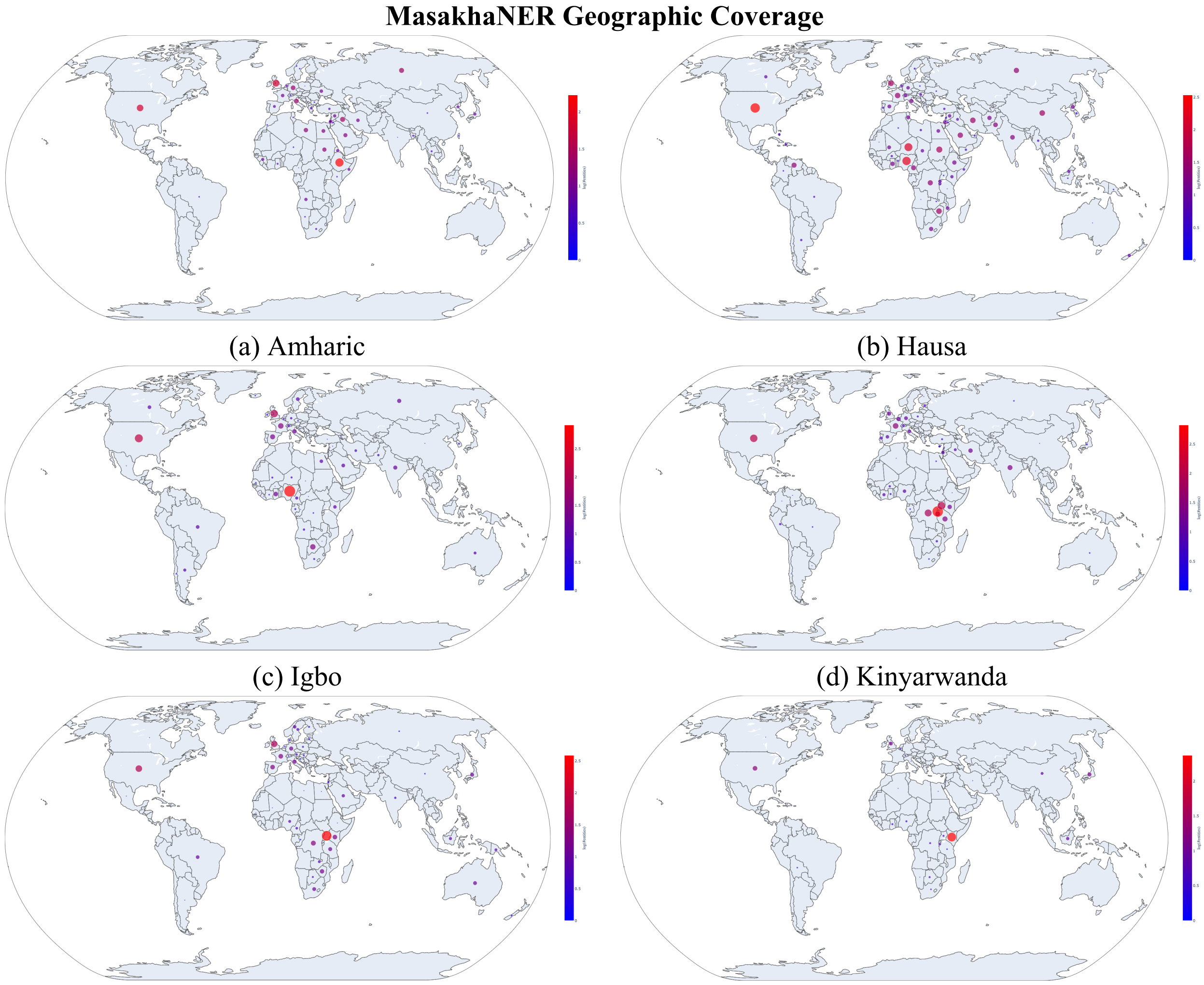
Table 4: Dataset Statistics

source-target	k=1 %	k=3 %	k=5 %	sentence count
en-ro	19.91	15.42	13.98	1999
en-fi	17.40	15.25	14.29	1500
en-pl	16.60	14.19	13.43	2000
en-fr	16.53	14.42	13.42	1500
en-tr	14.09	13.02	12.01	1001
en-lt	13.45	11.96	10.77	2000
en-et	13.40	11.88	10.74	2000
en-ja	13.36	11.88	11.57	1998
en-zh	12.19	11.66	10.26	2002
en-lv	9.59	9.21	8.55	2003
en-kk	7.79	8.84	7.88	2066
en-ta	7.09	6.94	6.19	1989
en-gu	3.75	2.70	2.24	1998
en-iu	1.47	1.34	1.31	5173

Table 5: Cross-lingual consistency score (%) for top-k extracted and linked entities over all source language sentences.

Entity category	Common	Source-only
Unknown	1720	16709
PERSON	1358	5713
ORG	1047	6911
GPE	666	7379
NORP	176	1895
DATE	102	1427
CARDINAL	78	565
EVENT	77	777
LOC	62	453
WORK_OF_ART	20	133
PRODUCT	15	91
FAC	14	161
QUANTITY	8	85
TIME	6	43
MONEY	4	14
LAW	3	113
LANGUAGE	3	80
ORDINAL	2	90
PERCENT	1	3
TOTAL	5362	42642

Table 6: SpaCy NER (Honnibal and Montani, 2017) defined types and counts for consistent linked entities.



(a) Amharic (b) Hausa

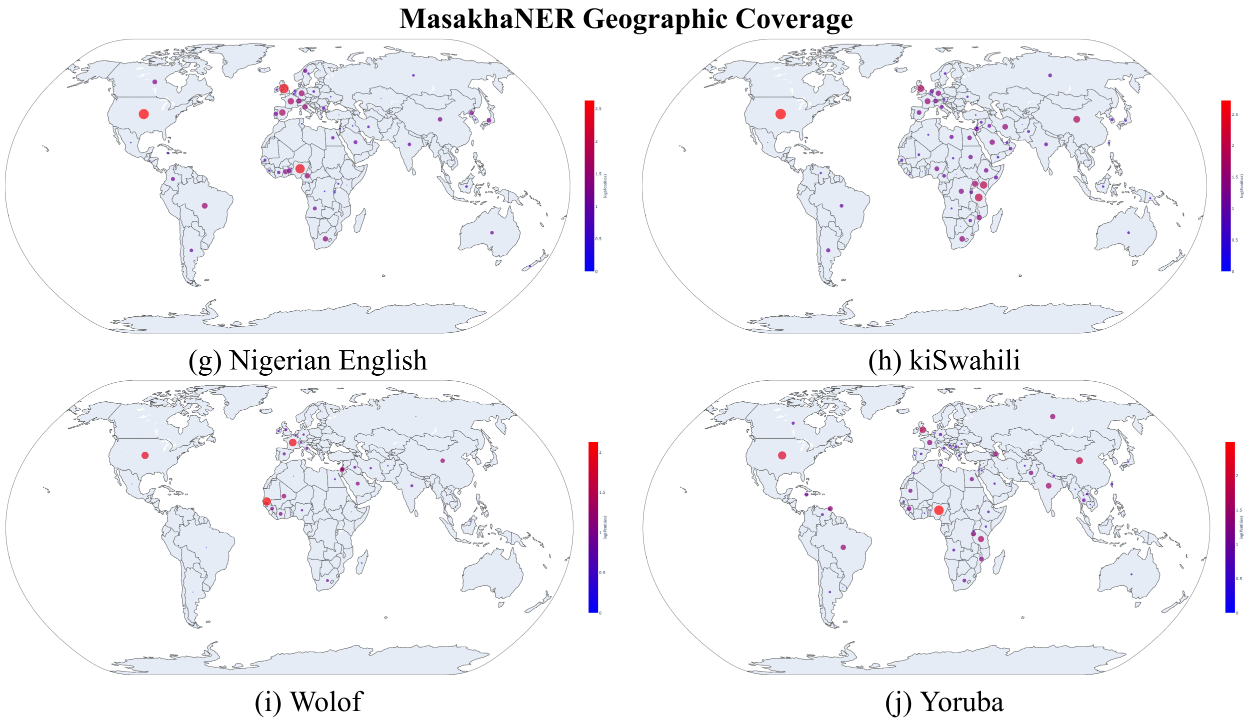
(c) Igbo (d) Kinyarwanda

(e) Luganda (f) Dholuo

Figure 6: MasakhaNER Geographic Distributions (Part 1).

Factors ϕ	X-FACTR (11) Explained Variance	X-FACTR (11) MAE	MasakhaNER (10) Explained Variance	MasakhaNER (10) MAE	WikiANN (48) Explained Variance	WikiANN (48) MAE
pop	0.356	0.457	0.300	0.295	0.387	0.470
gdp	0.516	0.407	0.341	0.295	0.575	0.382
geo	0.022	0.585	0.100	0.359	0.069	0.586
pop+gdp	0.495	0.403	0.348	0.285	0.553	0.388
pop+geo	0.356	0.455	0.369	0.290	0.399	0.467
geo+gdp	**0.521**	**0.398**	**0.443**	**0.284**	**0.591**	**0.376**
pop+gdp+geo	0.504	**0.398**	0.440	0.285	0.572	0.380

Table 7: Empirical comparison of factors on NER datasets, averaging over their respective languages (number in parentheses). We report the five-fold cross-validation explained variance and mean absolute error of a linear model.



(g) Nigerian English (h) kiSwahili (i) Wolof (j) Yoruba

Figure 7: MasakhaNER Geographic Distributions (Part 2).

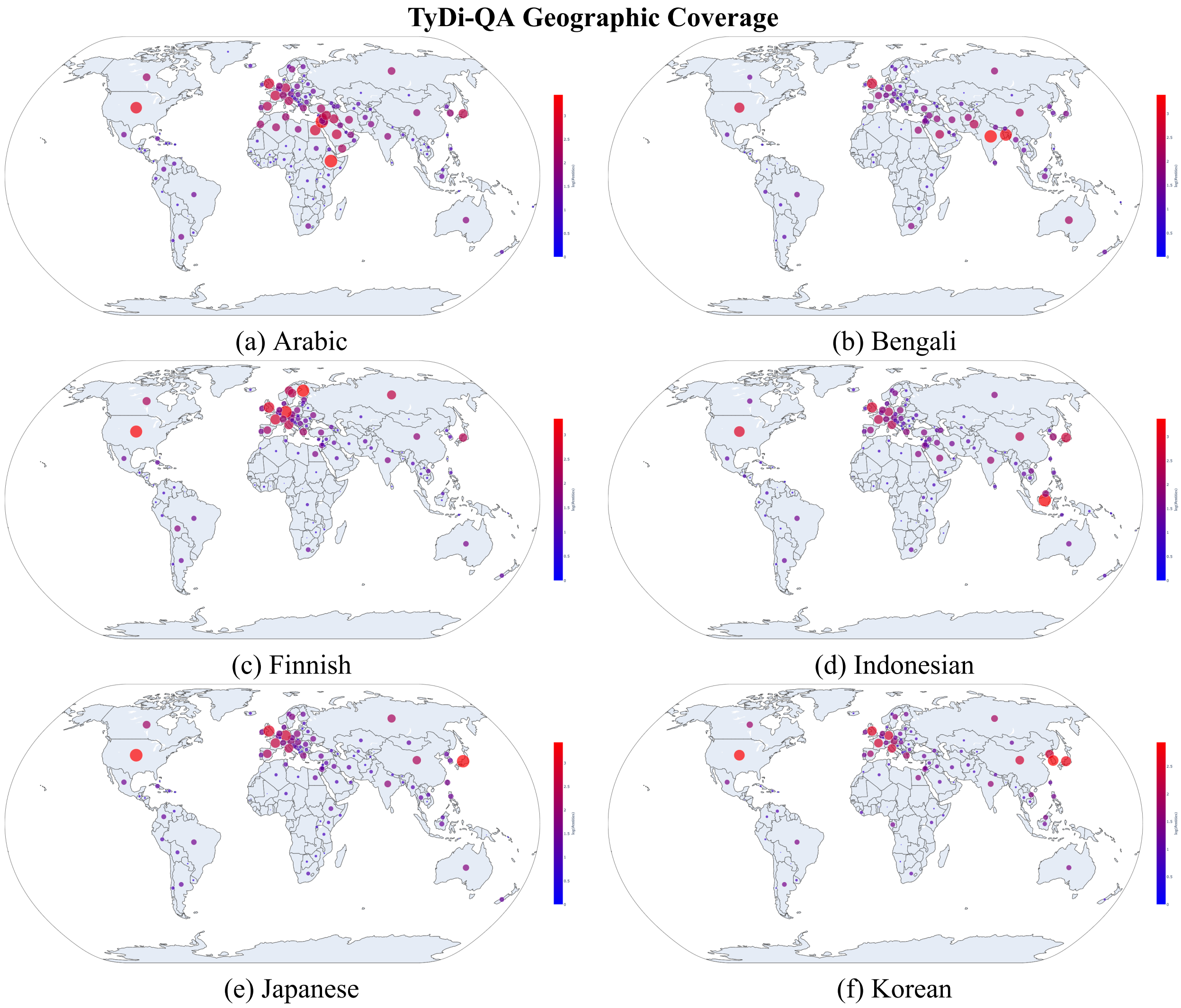


Figure 8: TyDi-QA Geographic Distributions (Part 1).

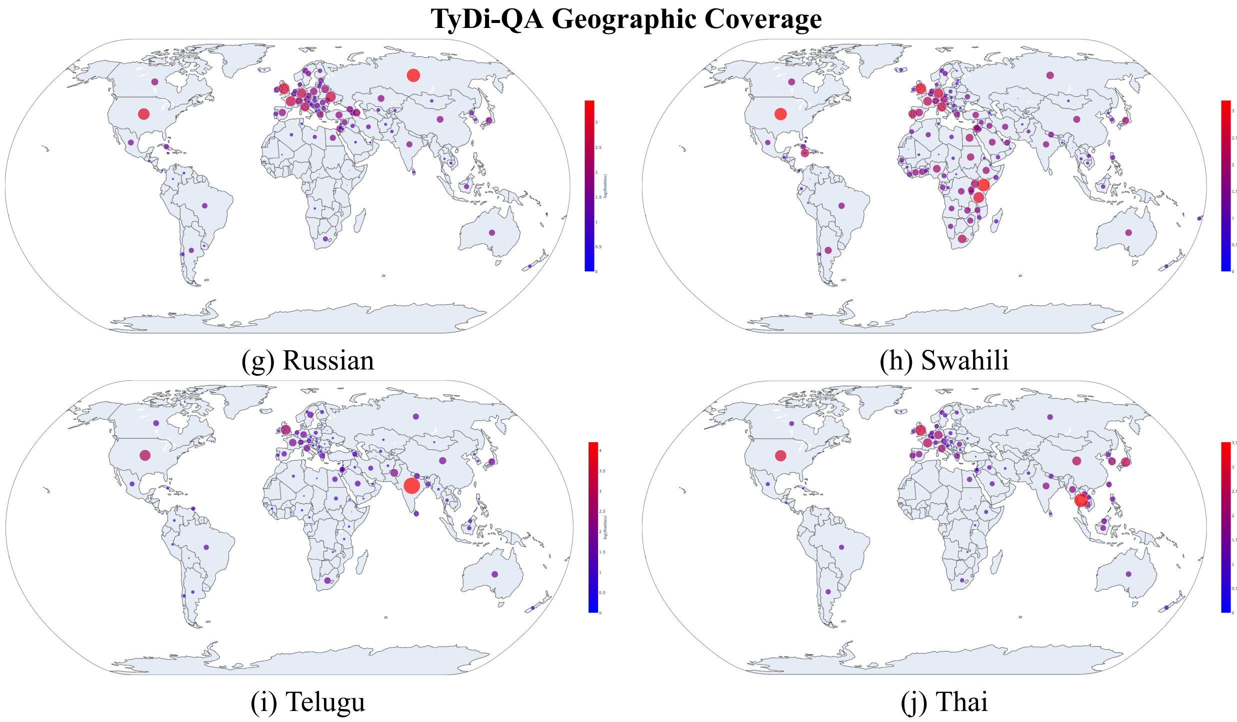


Figure 9: TyDi-QA Geographic Distributions (Part 2).

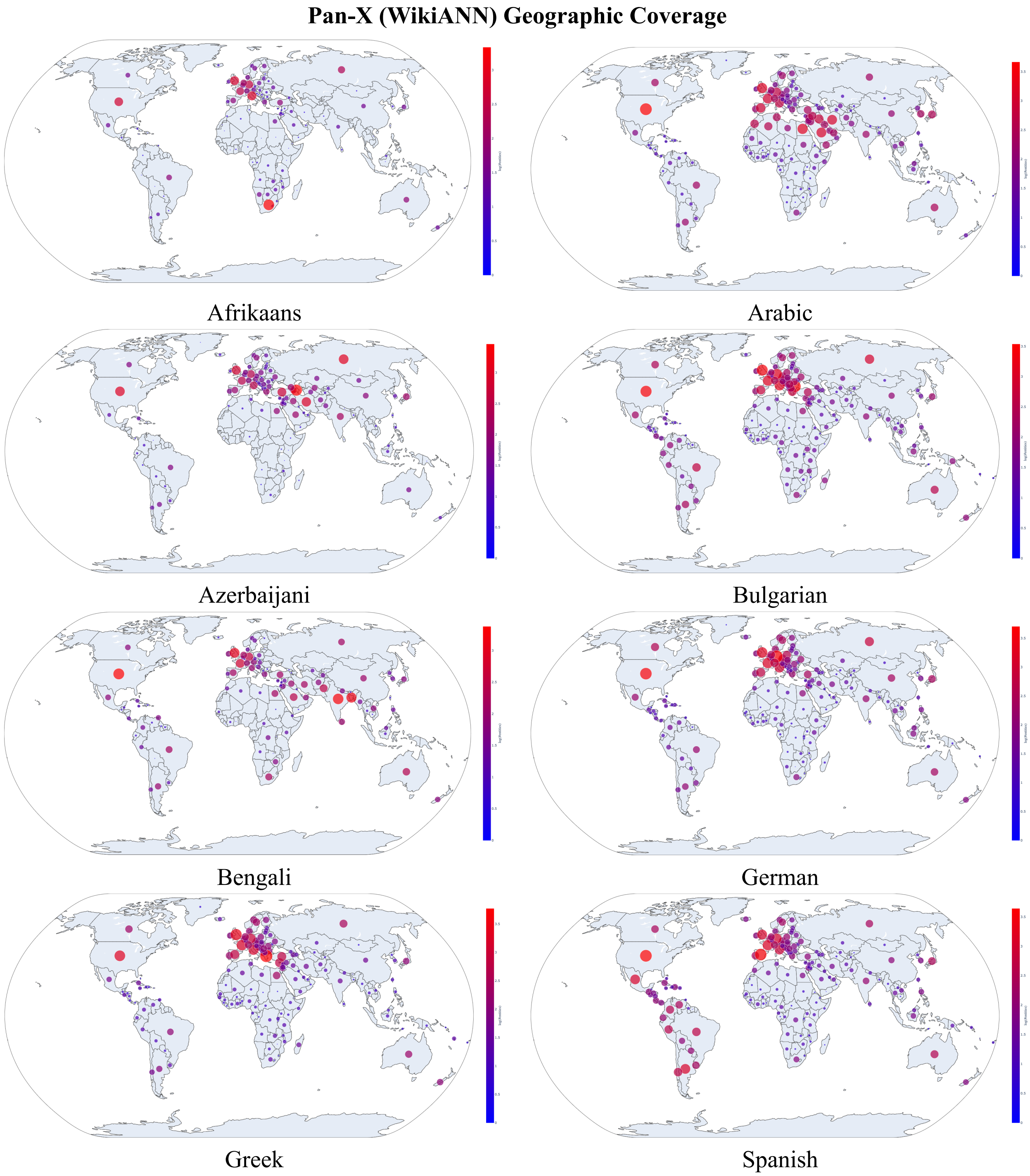


Figure 10: WikiANN Geographic Distributions (Part 1).

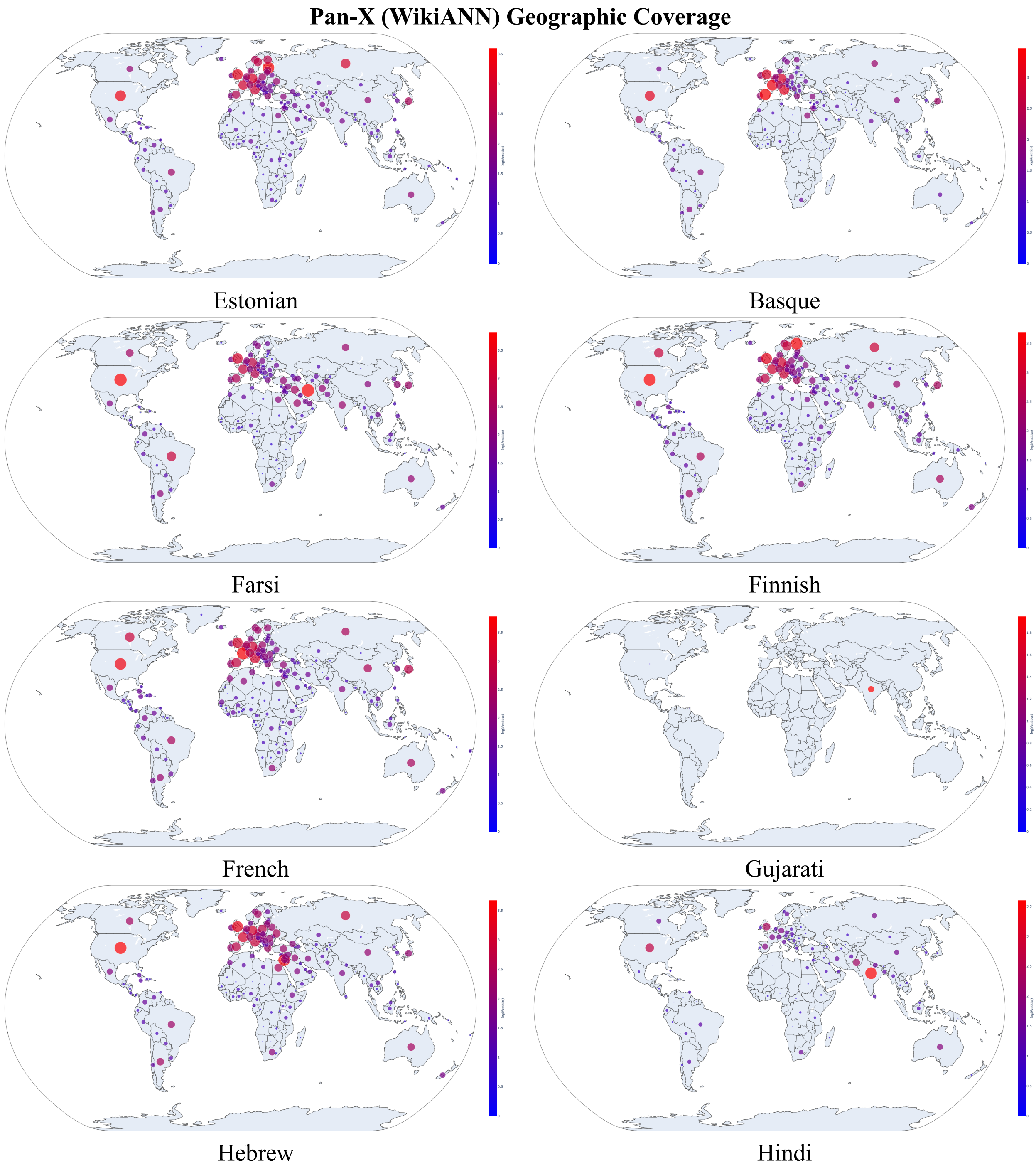


Figure 11: WikiANN Geographic Distributions (Part 2).

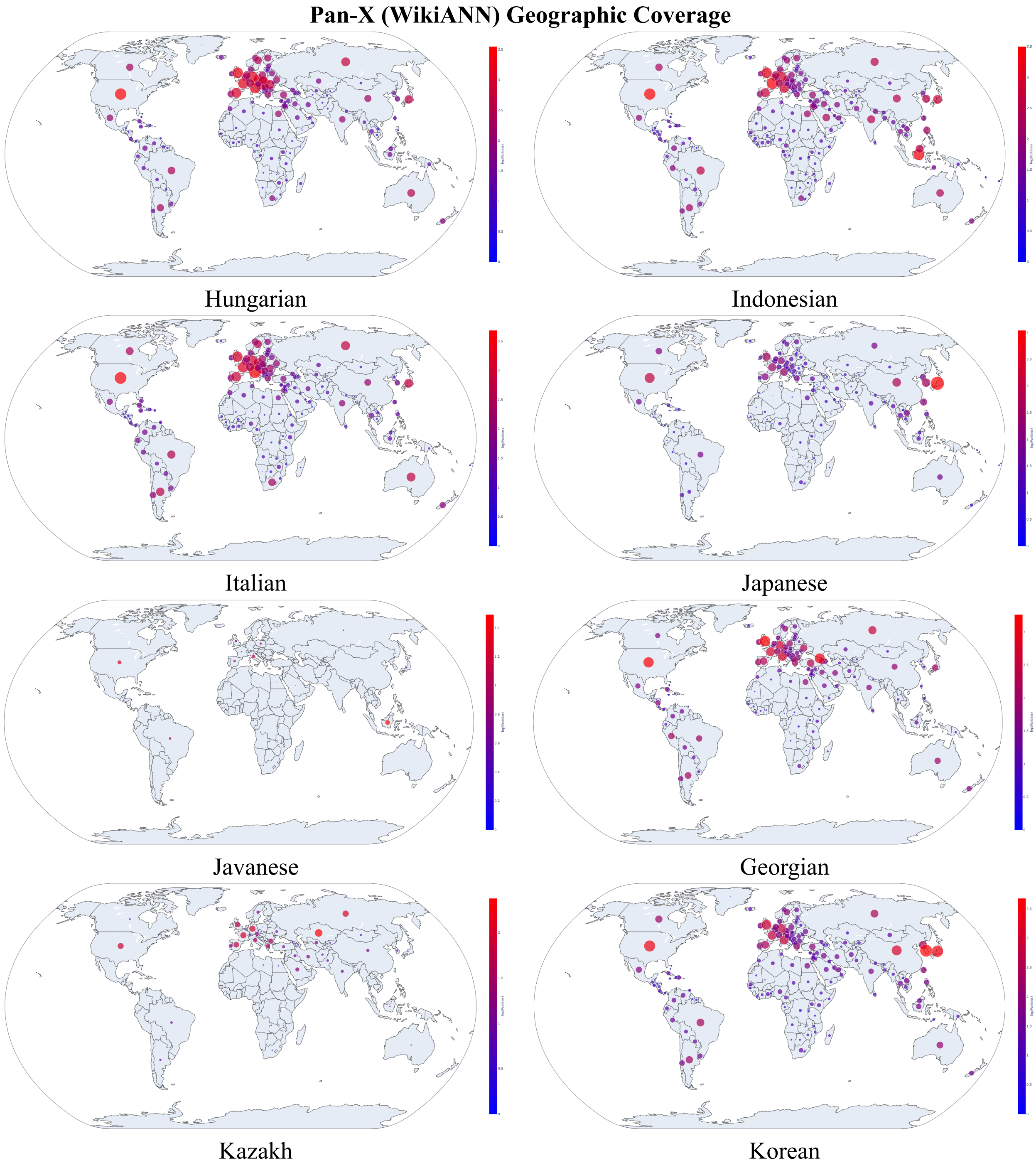


Figure 12: WikiANN Geographic Distributions (Part 3).

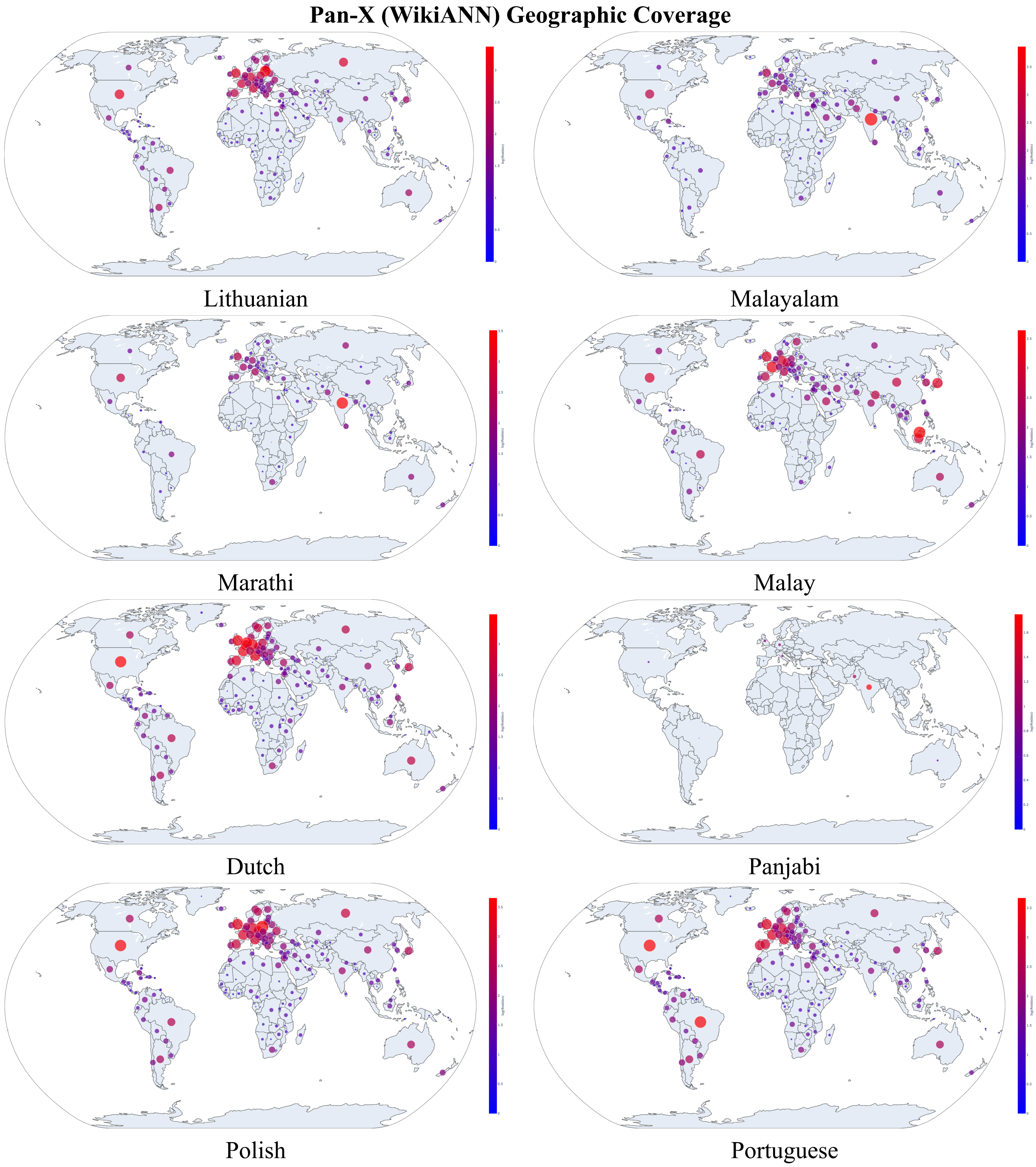


Figure 13: WikiANN Geographic Distributions (Part 4).

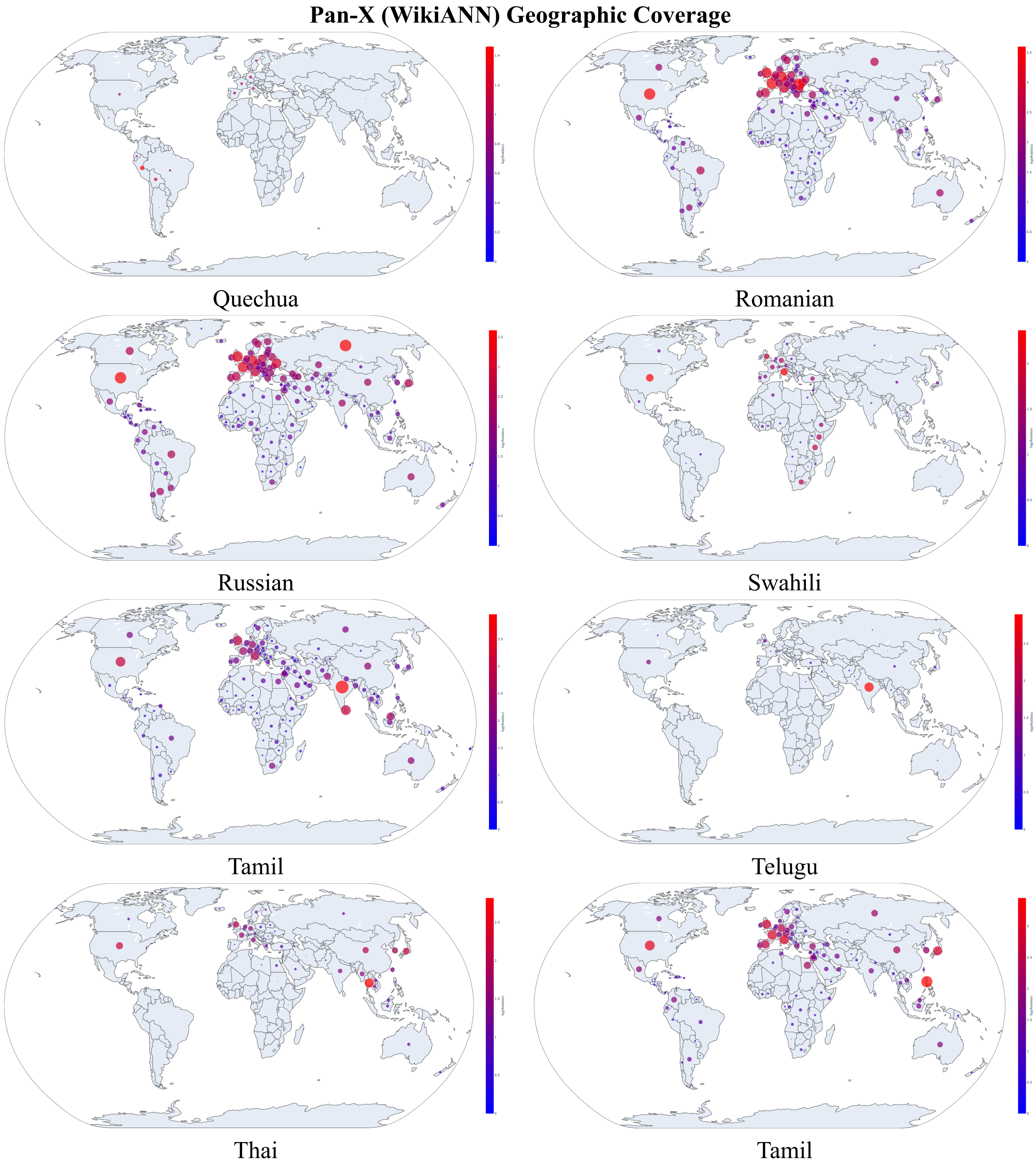


Figure 14: WikiANN Geographic Distributions (Part 5).

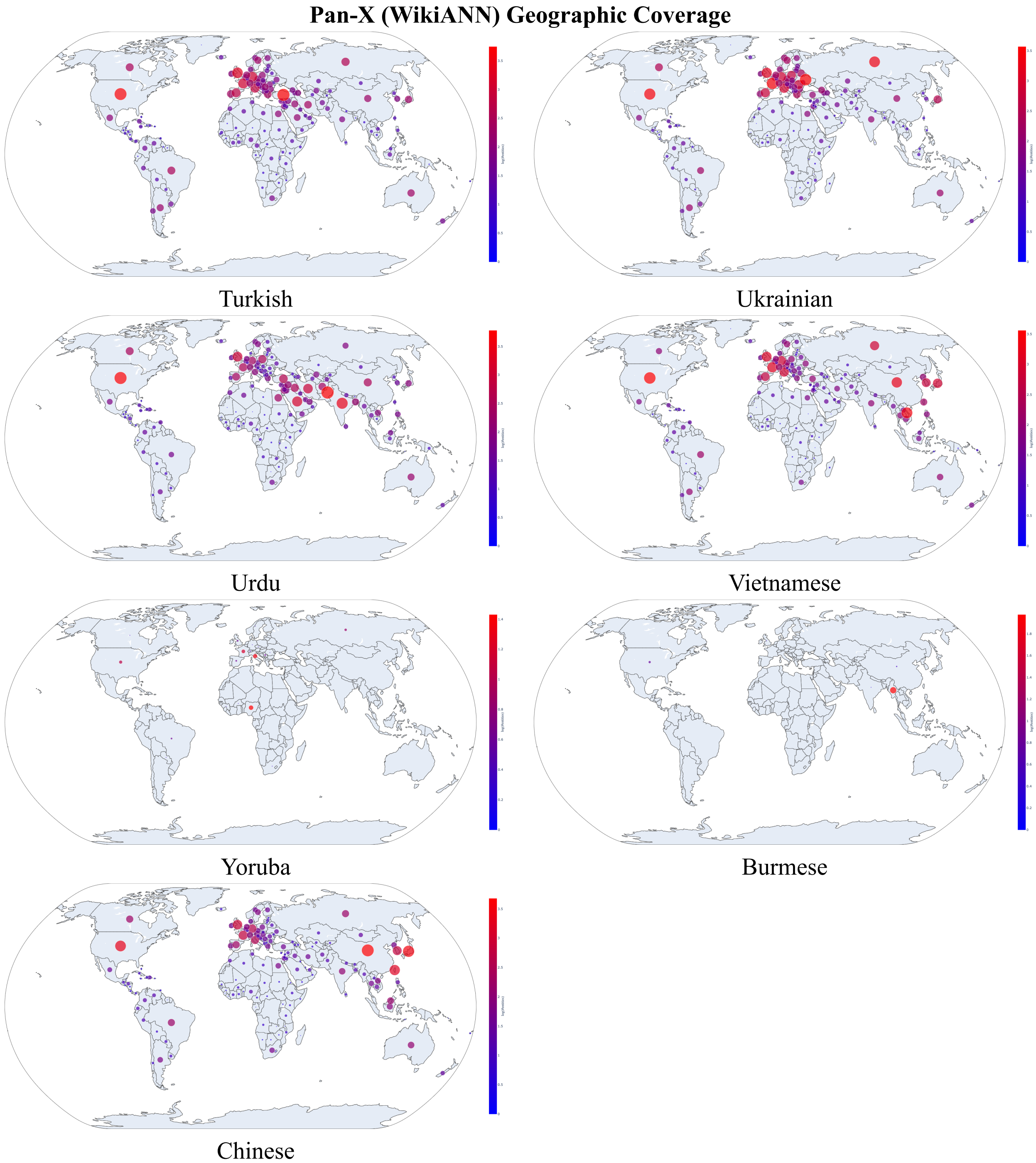


Figure 15: WikiANN Geographic Distributions (Part 6).

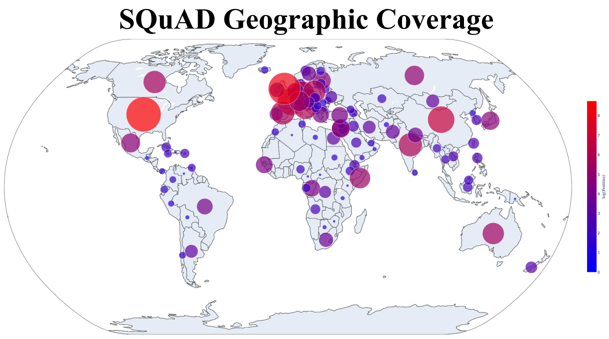


Figure 16: SQuAD Geographic Distributions.

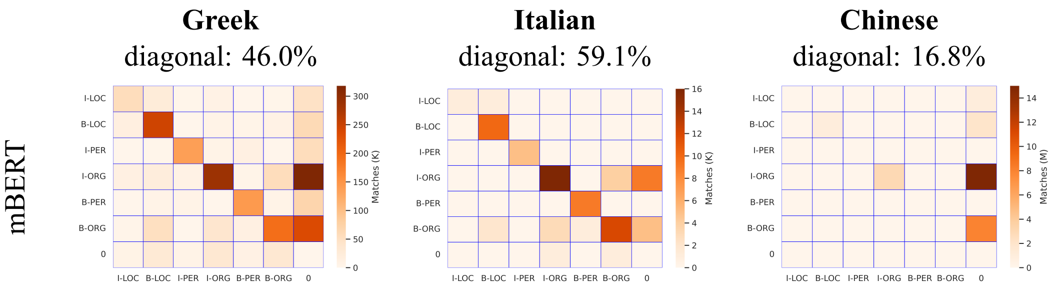


Figure 17: Confusion matrices for Greek, Italian and Chinese.

Language	Country	geo+gdp Expl. Var.	Mean Error
Arabic	SAU	0.501	0.415
Bengali	BGD	0.498	0.385
English	USA	0.562	0.335
Finnish	FIN	0.566	0.376
Indonesian	IDN	0.515	0.387
Japanese	JPN	0.558	0.388
Korean	KOR	0.546	0.336
Russian	RUS	0.522	0.400
Swahili	KEN	0.428	0.469
Telugu	IND	0.534	0.294
Thai	THA	0.550	0.333
Average		0.550	0.333

Table 8: Language breakdown of the most predictive factors (ϕ_{geo} and ϕ_{gdp}) on the TyDi-QA dataset.

Language	Country	geo+gdp Expl. Var.	Mean Error
Amharic	ETH	0.131	0.220
Yoruba	NGA	0.338	0.258
Hausa	NGA	0.321	0.317
Igbo	NGA	0.326	0.207
Kinyarwanda	RWA	0.198	0.229
Luganda	UGA	0.302	0.195
Luo	ETH	0.000	0.110
Nigerian English	NGA	0.493	0.231
Wolof	CMR	0.378	0.160
Swahili	KEN	0.443	-0.285
Average		0.378	0.160

Table 9: Language breakdown of the most predictive factors (ϕ_{geo} and ϕ_{gdp}) on MasakhaNER dataset.

Language	Country	geo+gdp Expl. Var.	Mean Error
af	ZAF	0.497	0.338
ar	SAU	0.570	0.454
az	AZE	0.566	0.395
bg	BGR	0.511	0.475
bn	BGD	0.442	0.502
de	DEU	0.613	0.402
el	GRC	0.484	0.456
es	ESP	0.497	0.462
et	EST	0.565	0.398
eu	ESP	0.565	0.387
fa	IRN	0.589	0.426
fi	FIN	0.590	0.411
fr	FRA	0.597	0.408
gu	IND	0.068	0.030
he	ISR	0.551	0.456
hi	IND	0.529	0.279
hu	HUN	0.563	0.451
id	IDN	0.488	0.442
it	ITA	0.569	0.436
ja	IDN	0.591	0.343
jv	JPN	0.062	0.069
ka	GEO	0.474	0.435
kk	KAZ	0.411	0.205
ko	KOR	0.519	0.423
lt	LTU	0.533	0.395
ml	IND	0.495	0.367
mr	IND	0.530	0.320
ms	MYS	0.496	0.463
my	MMR	0.105	0.038
nl	NLD	0.582	0.435
pa	IND	0.052	0.064
pl	POL	0.584	0.436
pt	PRT	0.567	0.432
qu	PER	0.301	0.090
ro	ROU	0.581	0.436
ru	RUS	0.576	0.435
sw	KEN	0.402	0.223
ta	LKA	0.524	0.367
te	IND	0.351	0.107
th	THA	0.567	0.215
tl	PHL	0.473	0.399
tr	TUR	0.619	0.409
uk	UKR	0.576	0.447
ur	PAK	0.512	0.463
vi	VNM	0.557	0.440
yo	NGA	0.079	0.086
zh	CHN	0.591	0.376
Average		0.591	0.376

Table 10: Language breakdown of the most predictive factors (ϕ_{geo} and ϕ_{gdp}) on the WikiANN dataset.