Analysing Cross-Lingual Transfer in Low-Resourced African Named Entity Recognition

Michael Beukman and **Manuel Fokam**
School of Computer Science and Applied Mathematics,
University of the Witwatersrand, Johannesburg, South Africa
{mcbeukman,arnolfokam23}@gmail.com

Abstract

Transfer learning has led to large gains in performance for nearly all NLP tasks while making downstream models easier and faster to train. This has also been extended to low-resourced languages, with some success. We investigate the properties of cross-lingual transfer learning between ten low-resourced languages, from the perspective of a named entity recognition task. We specifically investigate how much adaptive fine-tuning and the choice of transfer language affect zero-shot transfer performance. We find that models that perform well on a single language often do so at the expense of generalising to others, while models with the best generalisation to other languages suffer in individual language performance. Furthermore, the amount of data overlap between the source and target datasets is a better predictor of transfer performance than either the geographical or genetic distance between the languages.[1]

1 Introduction

The technique of using a pre-trained Natural Language Processing (NLP) model and fine-tuning it on task-specific data has recently taken the NLP world by storm, achieving state-of-the-art scores in many different tasks (Jiang et al., 2020; Raffel et al., 2020; Hendrycks et al., 2021). Although much of the focus of pre-trained models is on English (Radford et al., 2018; Devlin et al., 2019), there are also monolingual models for other languages (de Vries et al., 2019; Canete et al., 2020) and multilingual models that were trained on a large multilingual corpus (Conneau et al., 2020; Xue et al., 2021).

Generally, the training data of these models mostly consists of higher-resourced languages (i.e., those that have large amounts of available data, such as English and German). This can result in a large discrepancy between the performance of these models on higher-resourced and low-resourced languages (where data is scarce; e.g., many African languages (Alabi et al., 2022)).

A common challenge that arises when working with these models is the lack of task-specific data for the target language (Adelani et al., 2021). Despite this, in many cases, we have access to data from other languages. This presents an opportunity to leverage *cross-lingual transfer*, training a model on the language that we have data for and using it to make predictions for the target language. This is a common scenario, especially for low-resourced languages (Adelani et al., 2021).

Given this opportunity for cross-lingual transfer and the prevalence of pre-trained models, research has begun investigating the properties of these models more deeply. Studies have looked into multilingualism (Pires et al., 2019; K et al., 2020), syntactic transfer (Dhar and Bisazza, 2018), and the effect of linguistic features (Dolicki and Spanakis, 2021) and other attributes (Lin et al., 2019) on transfer performance. Despite this, it is not always clear which language we should transfer from, or which factors affect transfer (Lin et al., 2019).

Inspired by this line of work, we focus on investigating cross-lingual transfer more deeply, specifically in a low-resourced setting. We achieve this by studying the effect of different training schemes and identifying features that are indicative of high transfer performance. We build upon the work of Adelani et al. (2021), who recently introduced a high-quality named entity recognition dataset for ten low-resourced African languages. They also performed some analysis into which pre-trained models perform best and preliminary work into the cross-lingual transfer capabilities of models.

Our results show that adaptively fine-tuning a multilingual model on unlabelled monolingual data can improve performance on the target language, while often diminishing transfer performance by overfitting to this language. This effect is exac-

[1] We publicly release our code and models at https://github.com/Michael-Beukman/NerTransfer.

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 199–224
November 1–4, 2023.

erbated if the monolingual dataset is large. Furthermore, we find that when the source and target dataset contain many shared tokens, then transfer performance is generally higher. In particular, the number of overlapping tokens between datasets is a stronger predictor of transfer performance than many other features, including the geographic distance between where the languages are spoken, and the genealogical distance between the languages.

2 Background and Related Work

2.1 Named Entity Recognition (NER)

Named Entity Recognition is a token classification task in which the objective is to classify each token (or word) as one of a few classes, person, location, date, organisation, or no entity. NER is an impactful field (Sang and Meulder, 2003; Lample et al., 2016) with many applications (Marrero et al., 2013), including information retrieval and spell-checking (Adelani et al., 2021). In NER, performance is predominantly measured using the F1 score (Sang and Meulder, 2003; Adelani et al., 2021), which balances precision and recall.

2.2 Transfer Learning

Transfer learning is a technique that is often used in NLP to improve performance while requiring less task-specific data (Ruder et al., 2019). In one common form of transfer, we start by training a large language model on a massive corpus of unlabelled data, using these learned weights as the starting point for a specific problem, and fine-tuning further on task-specific labelled data (Ruder, 2021). This approach has become the dominant paradigm in NLP, especially for low-resourced languages, due to its high performance when fine-tuning on small datasets (Adelani et al., 2021). The idea is that the pre-training process instills knowledge into the model about how language behaves on a general level, which then does not need to be learned from scratch using the smaller amount of task-specific data (Radford et al., 2018; Devlin et al., 2018).

If the pre-training data is in a substantially different domain from the target task, we often use *adaptive fine-tuning*. This fine-tunes the pre-trained model on unlabelled data in the domain of the target task using a (masked) language modelling loss (Gururangan et al., 2020). A related approach, *language adaptive fine-tuning* (LAFT), fine-tunes a pre-trained model on unlabelled data in the target language, which can result in improved performance on the target language (Pfeiffer et al., 2020).

Recent work has also explored learning different pre-trained base models, tailored to particular languages. For instance, Ogueji et al. (2021) pre-train a BERT-style model on less than 1GB of text from African languages, and find that this performs well on downstream tasks, compared to massively-multilingual models that were trained on much larger datasets. Ogundepo et al. (2022) extend this by pre-training a T5-based model, expanding the applications to more general sequence-to-sequence tasks such as translation. Overall, these works contribute new, Africa-centric pre-trained models and provide initial benchmarks showing that these models can perform well on downstream tasks in the languages they pre-train on (after appropriate fine-tuning). While this is useful for advancing the field of low-resourced NLP, they generally do not deeply investigate cross-lingual transfer learning, which is the focus of our work.

2.3 Analysis

While approaches such as fine-tuning and cross-lingual transfer have been empirically shown to work well, there has been a recent trend that attempts to understand these techniques more deeply. For instance, Lin et al. (2019) focus on finding a way to choose the best language to transfer from, and develop a model that takes in a wide range of features, such as linguistic distance, entity overlap, etc., and predicts the transfer performance. Dolicki and Spanakis (2021) also consider features relevant to transfer and find that this depends on the task – no single feature can explain transfer performance well across tasks. Malkin et al. (2022) instead focus on the effect of the pre-training language and find that some languages, called *donors*, transfer well to others, while others, denoted *recipients*, benefit from transfer. Other work investigates how fine-tuning a pre-trained model alters its representations of words (Hsu et al., 2019). For example, Zhou and Srikumar (2022) study the effect that fine-tuning has on the representations of a multilingual model and find that this process often clusters together the representations that correspond to the same label, thereby making the classification task easier.

3 Methodology

The primary goal of this paper is to gain a deeper understanding of transfer learning in low-resourced settings. To achieve this, we focus on language

adaptive fine-tuning and cross-lingual transfer.

First, we investigate the effect of LAFT on transfer performance. Due to the cost of annotation, we often have more unlabelled data than labelled, task-specific data, making LAFT very applicable.

Secondly, we examine cross-lingual transfer in order to understand which languages transfer well to others and why. This is particularly relevant in cases where data is scarce in the target language but available in other languages, a common occurrence in low-resourced NLP. Knowing which features to consider when choosing a transfer language will be immensely useful to NLP practitioners faced with the choice of transfer language (Lin et al., 2019).

We therefore consider a low-resourced NER task, using the MasakhaNER dataset (Adelani et al., 2021). We fine-tune models on this dataset, and evaluate the effect of adding LAFT and transferring from different languages in Section 5.

Next, in Section 6 we investigate how much the transfer performance correlates with various language- and dataset-based features such as data overlap and linguistic distance. This helps us to understand which features should be considered when choosing a source language to transfer from.

3.1 Data

We consider all ten languages from the MasakhaNER dataset. We choose these languages for three reasons: firstly, they are all low-resourced compared to high-resourced languages such as English (Conneau et al., 2020), allowing us to study transfer learning in the important low-resourced setting. Secondly, there exists a high-quality dataset for these languages, in contrast to many other low-resourced languages. Finally, Adelani et al. (2021) already performed extensive baseline analysis on this dataset.

Information about the languages, including family, the region where it is spoken and dataset size, is contained in Table 1, with additional details in Appendix A. We do note that all the languages use the Latin script, except Amharic, which uses the Fidel script. Igbo, Wolof and Yorùbá use diacritics, which are symbols attached to some letters (e.g. in "ẹ"), which affect the pronunciation of the word.

4 Experiments

4.1 Experimental Setup

Our experiments largely consist of fine-tuning pre-trained language models on NER data and evaluating their cross-lingual transfer performance. We perform each experiment 5 times with different random seeds and report the mean performance. We note that the standard deviation across the different seeds is often quite large when performing transfer, i.e., when the fine-tuning and testing language are not the same. More details are in Appendix C.3.

We use the MasakhaNER implementation[2] and use the same hyperparameters and language codes as Adelani et al. (2021). All metrics reported are overall F1 scores on the test set (to compare against prior work), using the "begin" repair strategy as specified by Palen-Michel et al. (2021). More details regarding the training and evaluation procedures can be found in Appendix B.

4.2 Models

We mainly consider two types of models, the first being `xlm-roberta-base`, denoted as "base". Secondly, we consider LAFT models, obtained by fine-tuning `xlm-roberta-base` on unlabelled monolingual data from a specific language. We choose to use `xlm-roberta-base` as the base model due to its high performance and fast training (Adelani et al., 2021). This model was pre-trained on a large corpus consisting of data from 100 languages, including Amharic, Hausa and Swahili.

We then fine-tune these models on the NER data of a specific language. For clarity, we contract the training procedure of a model, for example, `base` $\rightarrow$ `hau` $\rightarrow$ `wol` is the `xlm-roberta-base` model that performed language-adaptive fine-tuning on Hausa, followed by NER fine-tuning on Wolof. More information about these models and the LAFT process is contained in Appendix B.

5 Cross-lingual Transfer

Here we investigate the zero-shot transfer performance of `xlm-roberta-base` and the language-adaptive models. For each pair of languages X, Y, we take the model fine-tuned on NER data from language X and evaluate its performance on language Y. To evaluate the effect of LAFT, we use both `base` and `base` $\rightarrow X$ (the latter model being obtained by performing LAFT on language X).

This experiment simulates the scenario where we do not have ample labelled data in the source language, but we possess task-specific data in a different language. Here, we must choose the best language to transfer from. This is a common setup in

[2] https://github.com/masakhane-io/masakhane-ner/

Table 1: Language details, partially reproduced from Adelani et al. (2021), with permission. The *NER* and *LAFT Size* columns contain the number of sentences in the NER training dataset and the unlabelled LAFT dataset, respectively. *Country* is the top one or two countries with the most speakers of the language, from Eberhard et al. (2020).

Language	Lang. Code	Family	Country	Region	Speakers	NER Size	LAFT Size
Amharic	amh	Afro-Asiatic-Ethio-Semitic	Ethiopia	East	33M	1,750	3.1M
Hausa	hau	Afro-Asiatic-Chadic	Nigeria, Niger	West	63M	1,903	3.1M
Igbo	ibo	Niger-Congo-Volta-Niger	Nigeria	West	27M	2,233	1.1M
Kinyarwanda	kin	Niger-Congo-Bantu	Rwanda, Uganda	East	12M	2,110	726K
Luganda	lug	Niger-Congo-Bantu	Uganda	East	7M	2,003	506K
Luo Nilo	luo	Saharan	Kenya	East	4M	644	160K
Nigerian Pidgin	pcm	English Creole	Nigeria	West	75M	2,100	207K
Swahili	swa	Niger-Congo-Bantu	Tanzania, Kenya	Central & East	98M	2,104	12.6M
Wolof	wol	Niger-Congo-Senegambia	Senegal	West & NW	5M	1,871	42K
Yorùbá	yor	Niger-Congo-Volta-Niger	Nigeria	West	42M	2,124	910K

NLP, particularly for low-resourced languages (Lin et al., 2019; Pfeiffer et al., 2020; Adelani et al., 2021). Leveraging cross-lingual transfer can lead to useable models in a data-scarce setting, where no data is available for the target language.

5.1 Results

These results are shown in Fig. 1, with the y-axis representing the evaluation language, while the x-axis represents either the language we performed NER fine-tuning on (Fig. 1a), or both the LAFT and NER fine-tuning language (Fig. 1b). In Fig. 1a, as expected, the diagonal is brighter than the off-diagonal elements, as fine-tuning on the same language one evaluates on improves scores significantly. The best zero-shot transfer language generally performs well, obtaining 10-20 F1 lower than training on the target language. When evaluating on Yorùbá, Igbo, Luo and Amharic, however, transfer performance is significantly lower. Igbo and Yorùbá's use of diacritics, or that Amharic has a different script, may be the cause of this. Luo's low performance could be because it has a large number of entities that occur only in its dataset. In addition, `base` did not train on any Luo data, and Luo is from a different family to all of the other languages.

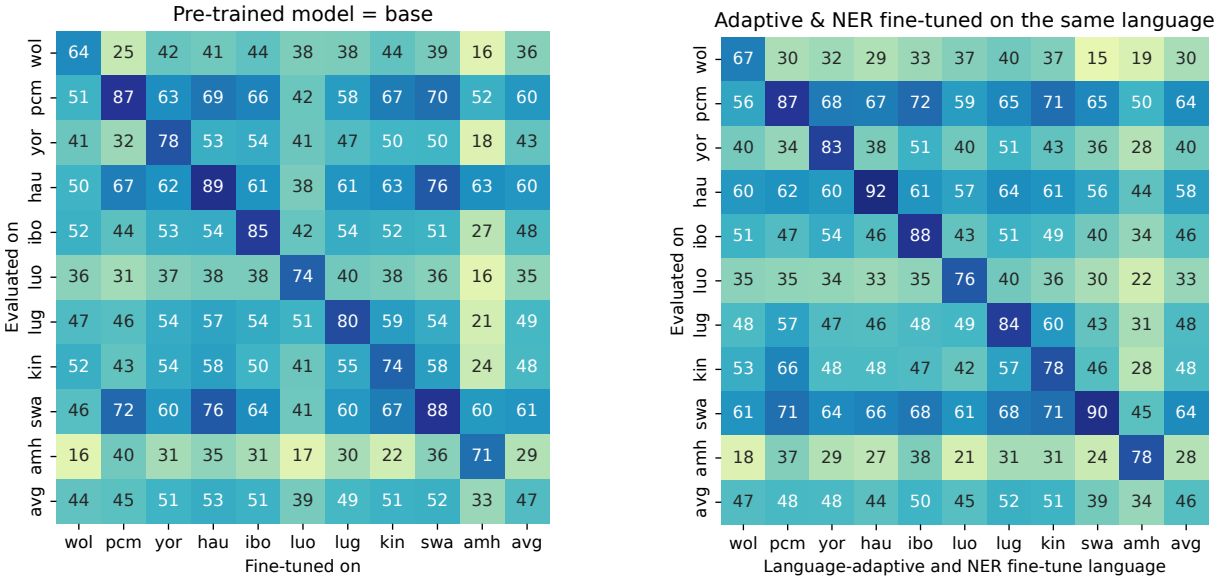
Furthermore, while Amharic transfers poorly on average, it transfers reasonably well to Swahili, Hausa and Nigerian Pidgin. The reason may be that Amharic, Swahili and Hausa were included in the base model's pre-training data, while Nigerian Pidgin shares many similarities with English, another pre-training language. Thus, the pre-trained model may have some link between its representations for Amharic and the other languages it jointly pre-trained on. Fine-tuning on Amharic changes these shared representations, leading to improved transfer results (see Appendix C.10 for details).

Observation: LAFT on the target language improves downstream performance

Comparing the diagonals in Fig. 1a and Fig. 1b, we can see that the LAFT models usually perform much better than the base model after subsequent NER fine-tuning. Of particular interest is the large improvement we see in Yorùbá and Amharic, where the language adaptive models outperform the base models by $+5$ and $+7$ F1, respectively. This could be because Yorùbá contains diacritics, and Amharic does not use the Latin script, making the language adaptive fine-tuning phase crucial to adapt the model to the specific characteristics of these languages. On average, by using language adaptive fine-tuning on the target language, we can improve the F1 performance by approximately 3 F1 after subsequent NER fine-tuning.

Observation: Performing LAFT on a large dataset can diminish transfer performance

While performing LAFT improves performance on the language we fine-tune on, transfer performance often shows a mixed result. For some language pairs, using a model that has been subject to language-adaptive fine-tuning on the same language as one fine-tunes on helps (e.g. the *pcm* column and *kin* row), but for others, this effect is minor, or even negative (e.g. *yor* transferring to *lug*). For some languages, notably Swahili and Hausa, using adaptively fine-tuned models (and then fine-tuning on NER data from the same language) significantly diminishes the transfer capabilities from these languages, possibly indicating overfitting. This is similar to what Pfeiffer et al. (2020) found when performing adaptive fine-tuning on the source language – transfer performance generally decreased. We investigate this further (more details in Appendix C.1) and find that those languages with fewer sentences in the language adaptive datasets



(a) base → X-axis (no LAFT)

(b) base → X-axis → X-axis

Figure 1: Heatmaps indicating the average performance over 5 seeds of specific models on specific languages (y-axis) after being fine-tuned on another language's NER data (x-axis). *avg* indicates the average transfer performance per row or column, respectively. This calculates the average of the entire row or column excluding the diagonal.

transfer better on average after performing LAFT and NER fine-tuning. There is a statistically significant correlation, with Pearson's $R = -0.82$, between the number of sentences in the LAFT dataset and the average improvement in transfer performance when using a LAFT model compared to using the base model. This suggests that larger LAFT datasets result in more overfitting and less transfer.

Recommendation

- Use LAFT on the target language prior to fine-tuning on NER data in the same language.
- If transfer performance is the priority, however, LAFT on a large dataset in the NER fine-tuning language should be avoided.

6 Explaining Transfer Performance

To explain some of the results shown in the previous section, here we examine other dataset and language features, determining whether they have any correlation with the transfer performance.

6.1 Data Overlap

The first feature we consider is the word overlap between the different languages' datasets. We do so because in NER, a token classification task, a model would benefit greatly from previously encountering an entity. Thus, if languages X and Y share tokens, a model trained on X would perform well on the already-seen tokens in language Y.

We call a token overlapping when the same token is labelled as the same entity type in two different datasets (e.g. `John[NAME]` would overlap with `John[NAME]`, but would not overlap with `John[ORG] Deere[ORG]`). To calculate the overlap between source language X and target language Y, we find all of the named entities (i.e., all tokens that are not labelled as "Other") that occur in both datasets and count the total number of times each token occurred in either dataset (e.g. if `John` occurred twice in X and three times in Y, we count it five times). We do not distinguish between tokens that are at the beginning of an entity or in the middle thereof (i.e., we consider B-PER and I-PER to be the same for this experiment). We also consider the entire dataset, i.e., train + dev + test, to obtain a more representative sample.

There are alternative ways to calculate overlap, such as only taking into account unique entities (which we avoid as one entity overlapping multiple times is relevant), or determining the fraction of overlapping tokens instead of the absolute number (Lin et al., 2019). However, we find that these alternative methods generally produce similar results and lead to similar conclusions, so the specific calculation method does not have a significant impact. In Appendix C.5, we provide a more in-depth explanation of these methods and their results.

6.1.1 Results

Fig. 2a shows the overlap between each pair of languages, with the diagonal being proportional to the number of entities for each language. Wolof and Luo have much less data than the other languages, and thus much less overlap, potentially explaining why these two performed poorly in previous experiments. In particular, Wolof has around three times more sentences than Luo, but fewer entities, indicating that entities are sparsely distributed throughout its sentences. Moreover, Amharic, due to it being written in a different script than all of the other languages, does not have any lexical overlap. Finally, there seems to be a large amount of data overlap in general, e.g., Swahili and Hausa have around 33% of their tokens overlapping.

Observation: Data Overlap Strongly Correlates with Transfer Performance

We see a strong correlation (Pearson's $R = 0.73$) between how many tokens overlap and the performance in Fig. 2b. The procedure here was simply to compute the correlation between the data overlap (as in Fig. 2a) and the performance when fine-tuning on one language and evaluating on another, starting from the pre-trained base model (as in Fig. 1a). We do not consider the diagonal elements, as they contain the performance of evaluating on language X after fine-tuning on language X and are thus not considered transfer learning.

These results do not imply a causal relationship, however, as previous work has shown that lexical overlap has a negligible impact on transfer performance, and word order, model depth and other attributes contribute more (Pires et al., 2019; Tran and Bisazza, 2019; K et al., 2020). This might be specific to the task under consideration, however, as other work still has shown that, for some tasks, the word and subword overlap between languages is a useful proxy for expected performance when performing cross-lingual transfer (Lin et al., 2019). Additionally, NER (and other token classification tasks) may be particularly sensitive to word overlap, as the classification happens on a per-word or a per-token basis. Finally, Amharic, due to its different script, has no overlap with any other language, while still displaying some transfer, indicating that more intricate mechanisms are at play.

Observation: Most of the overlap is in English

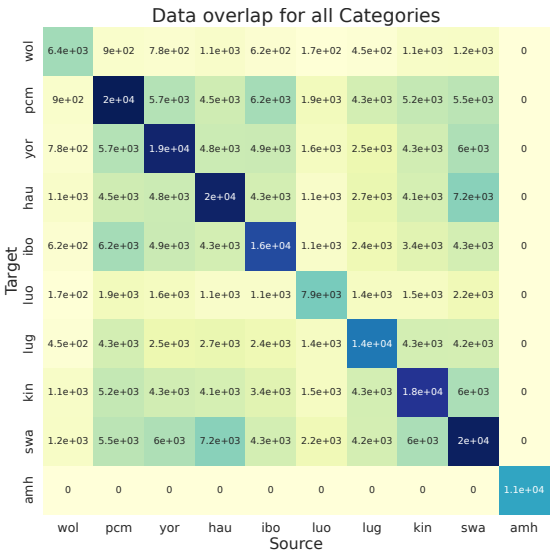
Having shown that data overlap has such a large correlation with transfer performance, we now investigate this deeper, to see which types of entities overlap. To aid in this, we classify a token as "international" if it falls into one of the following categories: (1) place names, such as "Africa", "Washington", "Nairobi"; (2) numbers, such as those found in dates, e.g., 2016; (3) the names of people in English, e.g. Paul, Jean; (4) punctuation marks found in the middle of entities; and (5) common words/companies such as "December", "Christmas", "Monday" and "Google". All of these are written in English. See Appendix C.4 for more details about these categories. Over all of the entity tokens across all languages, around 35% of these correspond to international tokens. However, when only considering the overlapping tokens, international words are the majority, around 69%. This seems to indicate that instead of overlapping tokens representing shared words between languages, they often represent international entities written in English. This holds even when considering only the distribution of unique tokens (instead of taking into account the number of times each token occurs). In this case, international tokens make up 28% of all tokens compared to 64% of overlapping tokens. This could be a factor present mostly in news-based data such as MasakhaNER, however, as news articles often cover globally relevant topics, leading to these types of entities being shared across datasets. Our findings in this section may partially explain why Adelani et al. (2021) obtained poor performance when transferring from Wikipedia. Finally, we note that the correlation results (Fig. 2b) are similar when only considering the local or international tokens, see Appendix C.6.

6.2 Additional Features

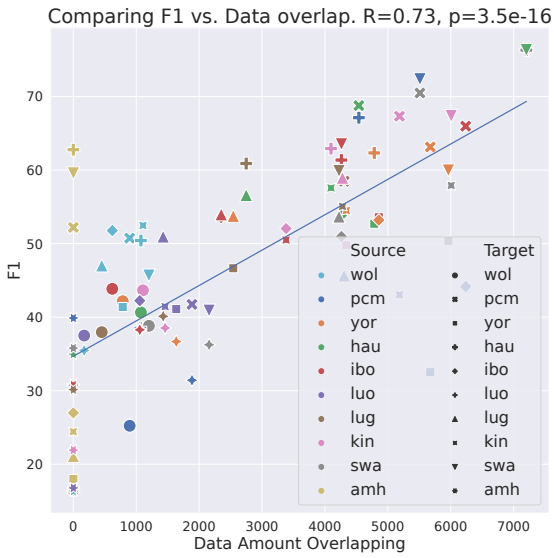
While we primarily focus on data overlap in this work, other features may also influence transfer performance between languages. We specifically consider the features used by Lin et al. (2019). This includes various language-based features, such as genetic and syntactic distance, dataset-based features, such as source dataset size, as well as the geographical distance between the countries where the languages are spoken.

6.2.1 Language-Based Features

The language-based features are largely based on the URIEL database of language properties (Littell et al., 2017). These features are: *Genetic*, *Inventory*, *Syntactic* and *Phonological* distance. *Genetic Distance* is how different two languages are



(a) Data Overlap. Row i, column j indicates the overlap if j was the source (i.e., fine-tuning language) and i was the target (i.e., evaluation language).



(b) Correlation between data overlap and F1 performance when performing zero-shot transfer. 0.73 Pearson's correlation coefficient with $p < 0.05$.

Figure 2: (a) Data overlap and (b) its correlation with F1. R in (b) is similar without Amharic, see Appendix C.7.

based on their language families. The other distances measure the cosine distance between vectors representing each language's syntax or phonology, derived from various linguistic databases (Lewis, 2009; Dryer and Haspelmath, 2013).

6.2.2 Dataset-Based Features

The dataset-based features are *Source Dataset Size*, representing the number of sentences in the source dataset and; *Source over Target Size Ratio*, the number of sentences in the source dataset divided by the number of sentences in the target dataset. We add two similar features, the number of named entities in the source dataset, and the ratio between this and the number of entities in the target dataset.

6.2.3 Geographical Distance

The geographic distance is calculated as a normalised distance between the geographic center of where the language's speakers reside (Littell et al., 2017; Hammarström et al., 2018). This, however, is potentially problematic, especially if a language is spoken in multiple countries and is somewhat spread out. For instance, Swahili is spoken across Kenya, Tanzania, and other countries, but the "geographic center" for Swahili is marked as a point in Southeastern Tanzania (Hammarström et al., 2018). As this may not be particularly accurate, we also experiment with a different approach of calculating the geographic distance between two languages, that of the shortest distance between all of the countries where each language is spoken (obtained from Hammarström et al. (2018)). If these countries share a border, the distance is zero. However, this new method results in distances that are closely correlated with those obtained previously and result in similar conclusions, so we do not consider it further. Finally, *Featural Distance* is the cosine distance between vectors consisting of the four language-based features and the geographical distance.

Observation: Language-based features do not correlate strongly with transfer performance

Similarly to data overlap, we compute the correlation between these features and the transfer performance, with results in Table 2 (see Appendix C.8 for plots). We find that most of these features exhibit a poor correlation with the transfer performance, and not all of the correlations are statistically significant. The genetic distance is the language-based feature with the highest correlation with transfer performance, at Pearson's $R = -0.30$ (i.e., closer languages tend to transfer better). This is still much weaker than the data overlap's correlation of $R = 0.73$. Overall, this suggests that the most important feature for transfer performance is the overlap between the source and target datasets, instead of how close the languages are. However, it may still be best to consider several different factors instead of any single one (Lin et al., 2019).

Table 2: Pearson's correlation coefficient and the corresponding p-value for features used by Lin et al. (2019). The data overlap row is the same as in Fig. 2. The first five features are not statistically significant, as $p \geq 0.05$.

Feature	Type	R	p
Featural Distance	Linguistic	-0.00	1
Phonological Distance	Linguistic	-0.02	0.84
Inventory Distance	Linguistic	-0.06	0.55
Source Over Target Size Ratio	Dataset	-0.20	0.056
Geographic Distance	Geographic	-0.21	0.05
Source Dataset Size	Dataset	0.23	0.029
Syntactic Distance	Linguistic	-0.23	0.028
Source Number Of Entities	Dataset	0.29	0.0063
Source Over Target Entities Ratio	Dataset	-0.30	0.0044
Genetic Distance	Linguistic	-0.30	0.0041
Data Overlap	Dataset	**0.73**	$\mathbf{3.5 \times 10^{-16}}$

Recommendation

- Choosing a source language for NER based on its data overlap with the target is promising.
- Other features have small correlations with transfer performance, much less than data overlap, and should not be used as a primary reason for choosing a specific language.

7 Discussion

Our work follows a recent trend of analysing empirical results more deeply, attempting to better understand the underlying phenomena (Lin et al., 2019; Zhou and Srikumar, 2022). We specifically consider cross-lingual transfer for low-resourced languages, investigating the effect of LAFT, the choice of transfer language, and which features are indicative of high transfer performance.

In line with recent work, we find that LAFT can improve performance on downstream tasks (Adelani et al., 2021; Alabi et al., 2022). We also discover that performing this process on a large dataset can inhibit transfer performance. This motivates having more general models instead of overspecialised ones, which would ideally be more robust. The work of Alabi et al. (2022) may be particularly relevant here, as they perform LAFT on multiple source languages, and encounter less of a loss in generalisation performance compared to our case of performing LAFT on only one language.

We further find that data overlap between the source and target languages correlates strongly with transfer performance, and that this may provide a way to choose a promising language to transfer from. Many other language-specific features have a much lower correlation with transfer performance. This suggests that, for token classification tasks such as NER, data overlap is potentially more important than language similarities.

This may not always be the case, however. Performance in other tasks, such as machine translation, may be less influenced by the amount of data overlap. Furthermore, the MasakhaNER dataset largely consists of annotated news articles. This type of data may skew more towards discussing international entities than, say, local history or fact-based text such as Wikipedia. In these cases, geographical or linguistic distance may contribute more to transfer than data overlap. Thus, while we highlight some important results, they may not necessarily apply to other tasks and domains. This should be investigated in future work.

One promising avenue of investigation for future work is to examine transfer performance when all international tokens are removed, to determine if this would diminish the correlation with data overlap, resulting in other features becoming more important for transfer. Using a more sophisticated strategy than only counting overlapping words when they exactly match would be promising, potentially resulting in the identification of similar, but slightly different, words between related languages. Finally, while we considered ten low-resourced African languages, it would be valuable to extend this study to other languages and regions to determine how well our conclusions generalise.

8 Conclusion

In this paper, we conduct a thorough examination of transfer learning in low-resourced African languages, focusing on language-adaptive fine-tuning and cross-lingual transfer. We find that language-adaptive fine-tuning on a large dataset can lead to improved performance on the target language, but at the cost of reduced transfer performance.

We further demonstrate that data overlap between the source and target datasets is a powerful predictor of transfer performance in NER, surpassing other factors such as geographical or genealogical distance. This, however, does not necessarily imply that data overlap is the cause of transfer performance, as Amharic, without any overlap, still displays some transfer. We also find that English words make up the bulk of the overlapping tokens.

Ultimately, while more work is needed, we hope that our analysis could inform some of the experimental decisions and transfer considerations when dealing with lower-resourced languages, thereby improving the quality of NER models for these languages.

Acknowledgements

Computations were performed using High Performance Computing infrastructure provided by the Mathematical Sciences Support unit at the University of the Witwatersrand. We thank the reviewers for their helpful and insightful comments, which helped to strengthen the final version of this paper. Finally, thanks to Devon Jarvis, Jade Abbot, Steven James and Benjamin Rosman for useful input.

Limitations

While we believe that our work is valuable, it has several shortcomings that could be addressed in future work. First, our focus is solely on one task—NER. While this enables us to perform detailed experiments and analysis, the disadvantage is that our results may not be general to all NLP tasks. Furthermore, as mentioned in the discussion section, our overlap results may be quite particular to NER. Therefore, it would be particularly promising to extend our work to other tasks, such as machine translation, sentiment classification, text classification, etc.

Second, our work only focuses on a subset of ten African languages. While Africa exhibits a large amount of linguistic diversity, and has several low-resourced languages, our conclusions may not necessarily be applicable to all low-resourced languages, or languages in other regions, such as Asia, Latin America, etc. It would be beneficial to extend our work to other languages and regions by using some of the more recent datasets for low-resourced languages (Prabhakar et al., 2022; Adelani et al., 2022a,b; Ebrahimi et al., 2022; Patil et al., 2022). Relatedly, we only considered one dataset, MasakhaNER. While this dataset is of high quality, it is also relatively small. It would be valuable to investigate whether our results hold on lower-quality datasets, as low-resourced languages often lack high-quality datasets such as the one we considered in this work.

Finally, to isolate the training procedure, we focused only on one pre-trained model, `xlm-roberta-base`. Again, this enabled us to perform in-depth analysis, but it would be valuable to extend this work to other models, such as mBERT (Devlin et al., 2019), AfriBERTa (Ogueji et al., 2021) and AfriTeVa (Ogundepo et al., 2022), to determine if our results generalise to other models.

References

David Ifeoluwa Adelani, Jade Z. Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, et al. 2021. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9.

David Ifeoluwa Adelani, Jesujoba O. Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Hassan Muhammad, Guyo Dub Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Wambui Gitau, Jade Z. Abbott, Mohamed Ahmed, Millicent Ochieng, Aremu Anuoluwapo, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022a. A few thousand translations go a long way! leveraging pre-trained models for african news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070. Association for Computational Linguistics.

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba O. Alabi, Shamsuddeen Hassan Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Ouoba Kabore, Chris Chinenye Emezue, Aremu Anuoluwapo, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Elvis Mboning, Tajuddeen Gwadabe, Tosin P. Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo L. Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Adeyemi, Gilles Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu Ngoli, and Dietrich Klakow. 2022b. Masakhaner 2.0: Africa-centric transfer learning for named entity recognition. pages 4488–4508.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to african languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 4336–4349. International Committee on Computational Linguistics.

José Canete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online*. Association for Computational Linguistics.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model. arXiv:1912.09582.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*. Association for Computational Linguistics.

Prajit Dhar and Arianna Bisazza. 2018. Does syntactic knowledge in multilingual language models transfer across languages? In *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018*. Association for Computational Linguistics.

Blazej Dolicki and Gerasimos Spanakis. 2021. Analysing the impact of linguistic features on cross-lingual transfer. *CoRR*, abs/2105.05975.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig (eds.). 2020. Ethnologue: Languages of the world. twenty-third edition. http://www.ethnologue.com.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John E. Ortega, Ricardo Ramos, Annette Rios, Iván Vladimir Meza Ruíz, Gustavo Giménez Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando A. Coto Solano, Ngoc Thang Vu, and Katharina Kann. 2022. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 6279–6299. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Harald Hammarström, Robert Forkel, and Martin Haspelmath. 2018. Glottolog 3.0. *Max Planck Institute for the Science of Human History*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. *CoRR*, abs/2103.03874.

Tsung-Yuan Hsu, Chi-Liang Liu, and Hung-yi Lee. 2019. Zero-shot reading comprehension by cross-lingual transfer learning with multi-lingual language representation model. 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 2019, Hong Kong, China*. Association for Computational Linguistics.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. SMART: robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online*. Association for Computational Linguistics.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: an empirical study. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA*. The Association for Computational Linguistics.

M Paul Lewis. 2009. *Ethnologue: Languages of the world Sixteenth Edition*. SIL international.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019.

Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy*. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Dan Malkin, Tomasz Limisiewicz, and Gabriel Stanovsky. 2022. A balanced data approach for evaluating cross-lingual transfer: Mapping the linguistic blood bank. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*. Association for Computational Linguistics.

H. B. Mann and D. R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1).

Mónica Marrero, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato, and Juan Miguel Gómez-Berbís. 2013. Named entity recognition: fallacies, challenges and opportunities. *Computer Standards & Interfaces*, 35(5).

Steven Moran, D McCloy, and R Wright. 2014. Phoible online. max planck institute for evolutionary anthropology, leipzig.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Odunayo Ogundepo, Akintunde Oladipo, Mofetoluwa Adeyemi, Kelechi Ogueji, and Jimmy Lin. 2022. AfriTeVA: Extending "small data" pretraining approaches to sequence-to-sequence models. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 126–135, Hybrid. Association for Computational Linguistics.

Chester Palen-Michel, Nolan Holley, and Constantine Lignos. 2021. SeqScore: Addressing barriers to reproducible named entity recognition evaluation. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Parth Patil, Aparna Ranade, Maithili Sabane, Onkar Litake, and Raviraj Joshi. 2022. L3cube-mahaner: A marathi named entity recognition dataset and BERT models. *CoRR*, abs/2204.06029.

Jonas Pfeiffer, Ivan Vulic, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: an adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online*. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy*. Association for Computational Linguistics.

Akshara Prabhakar, Gouri Sankar Majumder, and Ashish Anand. 2022. CL-NERIL: A cross-lingual model for NER in indian languages (student abstract). In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 13031–13032. AAAI Press.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*

Sebastian Ruder. 2021. Recent Advances in Language Model Fine-tuning. http://ruder.io/recent-advances-lm-fine-tuning.

Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada*. ACL.

Samuel Sanford Shapiro and Martin B Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4).

Ke M. Tran and Arianna Bisazza. 2019. Zero-shot dependency parsing with pre-trained multilingual sentence representations. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP, DeepLo@EMNLP-IJCNLP 2019,*

Hong Kong, China, November 3, 2019. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*. Association for Computational Linguistics.

Yichu Zhou and Vivek Srikumar. 2022. A closer look at how fine-tuning changes BERT. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022, Dublin, Ireland*. Association for Computational Linguistics.

Appendix

A Dataset

As mentioned in the main text, we used the MasakhaNER dataset (Adelani et al., 2021).[3] Information about the dataset, including the number of sentences and data sources, broken down by language, is shown in Table 3. Adelani et al. (2021) discuss the characteristics of the languages in more depth.

Most of the data was sourced from various news websites around the same time, with e.g. the Swahili and Hausa data both coming from the VOA website. While the authors of Adelani et al. (2021) do not know for certain whether the Hausa and Swahili data are translations of each other, it is quite likely that the events covered are similar, as the data is from around the same period.

B Hyperparameters and Reproducibility

We make our code available to reproduce our experiments. Table 4 contains the hyperparameters that we used when training the models. We used the same hyperparameters and base code as Adelani et al. (2021). For all experiments, in total, we used around 1000 GPU hours, on an internal cluster. The model we use, `xlm-roberta-base`, has 270M parameters.

B.1 Language Adaptive Fine-tuning Procedure

The language adaptive models, introduced by Adelani et al. (2021), had the following procedure: Take the `xlm-roberta-base` model as a starting point and fine-tune this on unlabelled, monolingual data for one language (e.g. Swahili, Wolof, etc.) using a masked language modelling loss. This was done separately for each language, resulting in ten separate language-adaptive models. The data, including its source and the number of sentences, used for this process is described in Table 10 of Adelani et al. (2021).

B.2 Language Split of XLM-Roberta

XLM-Roberta (Conneau et al., 2020) is a multilingual large language model trained on a large corpus consisting of 100 languages. In particular, it was trained on 8 African languages, including 3 languages contained in the MasakhaNER dataset,

[3] Available at https://github.com/masakhane-io/masakhane-ner

Table 3: Information about the different data sources and breakdowns of the NER data per language. Reproduced from Adelani et al. (2021), with permission.

Language Data	Source	Train/ dev/ test	#Anno	PER	ORG	LOC	DATE	% of Entities in Tokens	#Tokens
Amharic	DW & BBC	1750/ 250/ 500	4	730	403	1,420	580	15.13	37,032
Hausa	VOA Hausa	1903/ 272/ 545	3	1,490	766	2,779	922	12.17	80,152
Igbo	BBC Igbo	2233/ 319/ 638	6	1,603	1,292	1,677	690	13.15	61,668
Kinyarwanda	IGIHE news	2110/ 301/ 604	2	1,366	1,038	2,096	792	12.85	68,819
Luganda	BUKEDDE news	2003/ 200/ 401	3	1,868	838	943	574	14.81	46,615
Luo	Ramogi FM news	644/ 92/ 185	2	557	286	666	343	14.95	26,303
Nigerian Pidgin	BBC Pidgin	2100/ 300/ 600	5	2,602	1,042	1,317	1,242	13.25	76,063
Swahili	VOA Swahili	2104/ 300/ 602	6	1,702	960	2,842	940	12.48	79,272
Wolof	Lu Defu Waxu & Saabal	1871/ 267/ 536	2	731	245	836	206	6.02	52,872
Yorùbá	GV & VON news	2124/ 303/ 608	5	1,039	835	1,627	853	11.57	83,285

Table 4: Hyperparameters for the fine-tuning experiments

Hyperparameter	**Value**
Number of Seeds	5
Fine-tuning Epochs	50
Maximum Sequence Length	200
Batch Size	32
Learning Rate	5e-5

Amharic, Swahili and Hausa. The amount of data for these African languages is quite small, however, consisting of 68M tokens for Amharic, 275M for Swahili and 56M for Hausa, compared to e.g. 55B for English and 10B for French and German. Thus, while the model was trained on some African languages, they only make up a small fraction of the entire pre-training dataset.

Additionally, it is important to question whether XLM-Roberta was either pre-trained or adaptively fine-tuned on, for example, the test dataset for some of the languages. Adelani et al. (2021) do not know this for certain. It is unlikely, however, as the data of XLM-Roberta was extracted from the CommonCrawl 2018 snapshot, whereas Adelani et al. (2021) created and annotated the MasakhaNER dataset in 2020 and 2021 from current (at the time) news data.

B.3 NER and its evaluation

In Named Entity Recognition, we often have a distinction between the start of a multi-word entity, and the continuation of one. For instance, John Deere would be labelled as B-ORG I-ORG (denoting the **b**eginning and **i**nside of an entity). However, in some cases, the gold standard, "correct" labels often have invalid transitions, such as I-ORG being immediately after O, bypassing the required B-ORG label (Palen-Michel et al., 2021). Relatedly, the output of a model may also contain some of these invalid transitions. This complicates evaluation, which has resulted in methods being developed to correct these problems. One common approach is the "begin" repair method, where any invalid "I-" is replaced by a "B-" (Palen-Michel et al., 2021). After the label sequence has been repaired, the standard evaluation procedure is then used, comparing the predicted output to the ground-truth annotations. We used this begin repair strategy in our experiments.

C Additional Results and Analysis

This section contains additional experiments and results. In particular, we consider the correlation between overfitting (i.e. transferring worse to other languages) after performing LAFT and the size of the LAFT corpus in Appendix C.1. Appendix C.2 contains more LAFT experiments, considering the effect of performing LAFT on a language other than the fine-tuning one. In Appendix C.3 we have additional transfer results, specifically considering the different NER categories in isolation, and expanding upon the increased variance we found when performing transfer. We next detail the process when classifying tokens as international in Appendix C.4. Appendix C.5 covers the various other overlap calculations we consider, confirming that they all lead to similar conclusions. Appendix C.6 considers correlation results when splitting the data and performance into international and local subsets. Appendix C.7 performs the correlation analysis between performance and data overlap without considering Amharic, which has no overlap with any other language. Appendix C.8 contains more in-depth definitions of the other features, besides data overlap, that we considered, as well as plots showing the correlation of each feature with transfer performance. In Appendix C.9 we consider

the effect of training on a combination of datasets. Finally, in Appendix C.10 we consider the representations of the pre-trained models, how they are changed by fine-tuning and how this may explain some of our transfer results.

C.1 Overfitting vs Dataset size

In this experiment, we evaluate the effect of the size of the language adaptive dataset on the transfer performance of a model trained on downstream data. To do this, we take the (a) base $\rightarrow X \rightarrow X$ models, for each language X; i.e., those that performed language adaptive fine-tuning on language X and additional NER fine-tuning on the same language. We then evaluate these models on the 9 other languages. We do the same for the (b) base $\rightarrow X$ models (i.e. the models that took the base pre-trained model and performed NER fine-tuning on language X). We subtract the average transfer performance of the (b) models from the (a) ones, to obtain the performance gain (or loss) after performing language adaptive fine-tuning. We then plot this quantity against the number of sentences in the language adaptive fine-tuning datasets (obtained from Table 10 in Adelani et al. (2021)) in Fig. 3. We see a strong negative correlation (Pearson's $R = -0.82$) that is statistically significant ($p < 0.05$). This seems to indicate that the larger our language adaptive fine-tuning dataset is, the worse a downstream model will transfer.

Furthermore, we find that this result still holds if we omit the three languages included in `xlm-roberta-base`'s pre-training dataset (Hausa, Swahili, Amharic), with $R = -0.89, p < 0.05$.

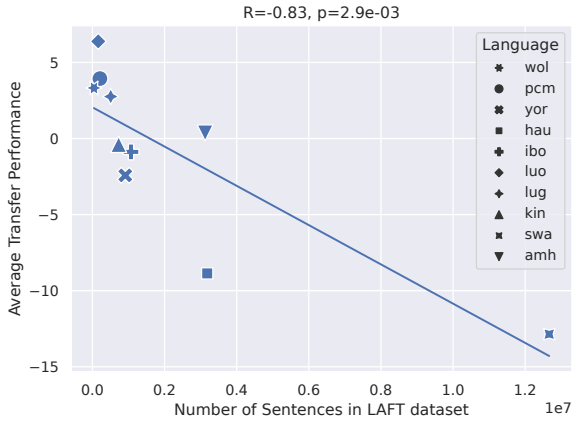


Figure 3: The correlation between the number of sentences in the LAFT sentences on the x-axis and the transfer performance delta of using this model compared to the base for fine-tuning.

C.1.1 Pre-training Size

In Fig. 1b, we find that, on average, only Swahili and Nigerian Pidgin perform better as target languages after performing LAFT on the source language. Hausa, for instance, does not. One reason for this may be the size of the `xlm-roberta-base` pre-training dataset: Swahili had 275M tokens, English 55B (Nigerian Pidgin is an English creole), Amharic 68M and Hausa only 56M. Pre-training on a large dataset may make the model less susceptible to generalising worse after performing LAFT on another language.

C.2 Additional LAFT Experiments

C.2.1 Experiment

For each language X, we compare four different models: base, base $\rightarrow$ Swahili, base $\rightarrow$ Hausa and base $\rightarrow$ X, where the latter three were subject to further language-adaptive fine-tuning on their respective languages. The base model acts as a baseline that does not perform adaptive fine-tuning at all, just fine-tuning on NER data. The base $\rightarrow$ X model shows the benefits of using adaptive fine-tuning on the target language. The base $\rightarrow$ Swahili and base $\rightarrow$ Hausa models provide information on how downstream performance is affected by language-adaptive fine-tuning on a different language. We chose Swahili as it was the language with the most speakers and the largest dataset out of the 10 available ones (Adelani et al., 2021), making it a promising language to transfer from. It is also spoken in Eastern Africa, like many of the other languages we consider. Hausa is chosen as a baseline, as another language with many speakers and a relatively large dataset. Hausa is predominantly spoken in Western Africa, in contrast to Swahili. We fine-tune each of these models on NER data and report the results when evaluated on the test set.

C.2.2 Results

We find that performing LAFT on Swahili outperforms the base model. When using Hausa as the LAFT language, however, we do not see any significant increase in performance compared to the base model when averaging over all languages. This may be due to the fact that the Hausa LAFT dataset is around 4 times smaller than the Swahili one. Besides Swahili and Hausa themselves, four languages have a significant (more than the standard deviation) difference in performance between the

Table 5: Performance of different models after fine-tuning and evaluating on NER data. We use a Mann-Whitney U test (Mann and Whitney, 1947) as some data failed a Shapiro Wilks normality test (Shapiro and Wilk, 1965). * indicates a statistically significant difference ($p < 0.05$) between the base model and the one under consideration, **bold** implies * and being the maximum per language. The leftmost column shows the model we started with before fine-tuning on language-specific NER data, while the other columns indicate the NER fine-tuning and evaluation language. For example, base $\rightarrow X$ is the language adaptive model for each column.

Starting point for NER fine-tune	wol	pcm	yor	hau	ibo	luo	lug	kin	swa	amh	avg
base	64.2 (1.3)	87.3 (0.9)	77.9 (0.3)	89.5 (0.4)	84.9 (0.7)	74.5 (1.3)	80.2 (0.7)	73.7 (0.7)	87.8 (0.5)	70.7 (1.1)	79.1 (0.2)
base $\rightarrow X$	66.9 (1.7)	87.1 (0.8)	**83.3 (0.3)***	**91.6 (0.4)***	**87.9 (0.5)***	76.2 (1.2)	**84.5 (0.5)***	**78.3 (1.0)***	**89.6 (0.6)***	**78.2 (0.8)***	**82.4 (0.2)***
base $\rightarrow$ swa	**67.3 (1.3)***	88.0 (0.8)	78.3 (1.0)	88.8 (0.2)*	84.3 (0.8)	77.2 (1.4)	82.0 (0.5)*	75.2 (1.0)	**89.6 (0.6)***	68.9 (0.9)	80.0 (0.5)*
base $\rightarrow$ hau	66.1 (2.0)	88.3 (1.0)	78.7 (0.7)	**91.6 (0.4)***	85.5 (0.4)	75.2 (1.0)	79.8 (0.6)	72.1 (0.8)*	87.6 (0.6)	68.4 (0.5)*	79.3 (0.3)

Swahili and Hausa models, Igbo, Luo, Luganda and Kinyarwanda. Hausa only outperforms Swahili on Igbo, which is predominantly spoken in West Africa, whereas Swahili outperforms Hausa on the other three languages, largely spoken in East Africa. This suggests that the region of the source and target language has an impact on the effectiveness of LAFT.

C.3 Additional Transfer Results

In Fig. 5, we show more transfer results. The first row contains transfer performance when fine-tuning on the x-axis and evaluating on the y-axis. Figs. 5a and 5b were contained in the main text, and Fig. 5c contains the results when using a $base \rightarrow$ swa language adaptive model. The second row shows the standard deviations of the F1 score over 5 seeds for each model in the first row. In particular, looking at Fig. 5d, the results indicate that the standard deviation is generally higher when performing transfer (off-diagonal elements) compared to performing standard evaluation on the fine-tuned language (diagonal elements). This suggests that transfer exhibits a lack of robustness to random initialisations. This effect is more pronounced when fine-tuning on Luo, as it has significantly less data than the other languages. When transferring from other languages to Amharic, the spread is higher than average, likely due to Amharic's different script.

C.3.1 Performance varies wildly across the different entity types

We also consider the above results in slightly more detail by looking at each NER category individually, to see if any perform much better or worse than the others. These results are shown in Fig. 4. We generally see that dates transfer poorly, over most languages, particularly for Luo. This could be caused by the differences in writing dates across these languages. Organisations transfer poorly for Amharic, possibly caused by its different script.

C.4 International Tokens in Overlap

We have a few separate categories of international tokens, described shortly. The full data we use can be found in our source code base.

Names This contains a list of common English names and surnames.

Places This list contains continents such as Africa, countries such as Russia, cities such as London and states such as Texas. We additionally have the four cardinal directions (North, South, East, West) and the 10 000 cities with the largest population.[4].

Companies A list of popular companies and organisations, such as Twitter, Youtube, Boeing, etc. We additionally use a list of the Fortune 1000 companies.

Numbers/Punctuation This category contains numbers and punctuation marks.

General English General English words, such as the names of the 12 months and the 7 days, words such as "International", "Hospital", "Christmas", etc. This category had the fewest occurrences on average, around 1% of tokens.

We generally find that places, names and numbers made up most of the overlapping tokens. Punctuation, the names of companies and general English words make up the smallest fraction, with less than 10% of the tokens.

[4] https://public.opendatasoft.com/explore/dataset/geonames-all-cities-with-a-population-1000/download/

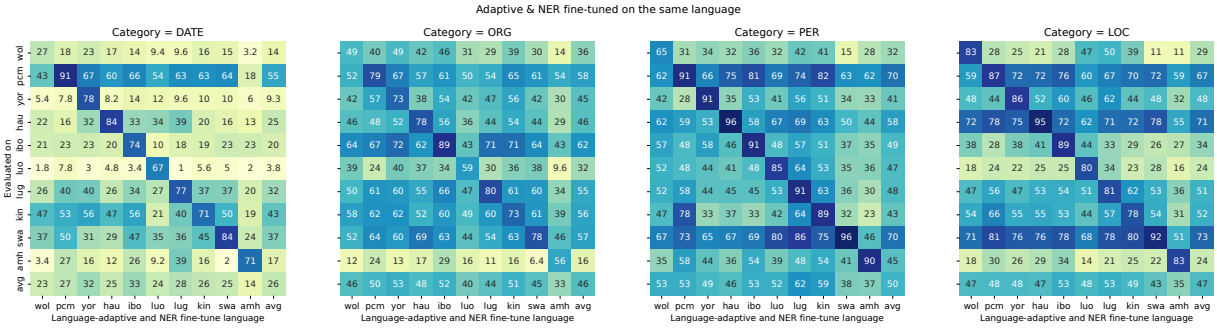


Figure 4: Heatmaps for the language-adaptive fine-tuned model (Fig. 5b), broken down by category.

C.5 Alternate Overlap Calculations

While we used just a single method for calculating overlap in the main text, here we show the results when using other, reasonable techniques. Overall, we find that the conclusions are the same, with overlap correlating strongly with transfer performance. In particular, the variations we consider are:

Unique Entities Only count the number of unique overlapping entities between the datasets.

Only Train Consider only the training dataset.

Source/Target/Sum When counting the number of times a token overlaps, use the number of occurrences in the source dataset, or the target dataset, or the sum of these two values.

Normalise Whether or not to normalise the overlap by dividing by either (1) the total number of entities, (2) the number of entities in the source, or (3) target language, etc.

Considering "O" Whether or not to consider the "Other" entities as well when calculating overlap, or just using the named entities.

Without Labels Whether to consider two entities overlapping if they have different labels.

Overall, we find that using the number of overlapping tokens as the number of occurrences in the source dataset has the lowest correlation, with R around 0.5. If we do not consider this approach of calculating overlap, then all correlation coefficients are at least 0.6, ranging up to 0.7. All correlations are statistically significant, with $p < 0.05$.

This shows that regardless of the overlap method used, there is a strong correlation between the number of overlapping tokens and the transfer performance in NER. Some of the results for different calculation methods are shown in Fig. 6. In particular, in the bottom row of this figure, we show results similar to those in the main text, but considering only the number of tokens present in the target dataset. We also calculate the fraction of overlapping tokens instead of the absolute number.

C.6 Splitting Overlap into Local and International

See Fig. 7 for the overlap results (similar to Fig. 2), split up into international and local tokens. The results here are similar to the ones in the main text (which was averaged over all tokens). The correlation is slightly lower for local tokens, but it is still positive and statically significant.

C.7 Overlap Correlations without Amharic

Fig. 8 contains the correlation results when not considering Amharic.

C.8 Additional Features

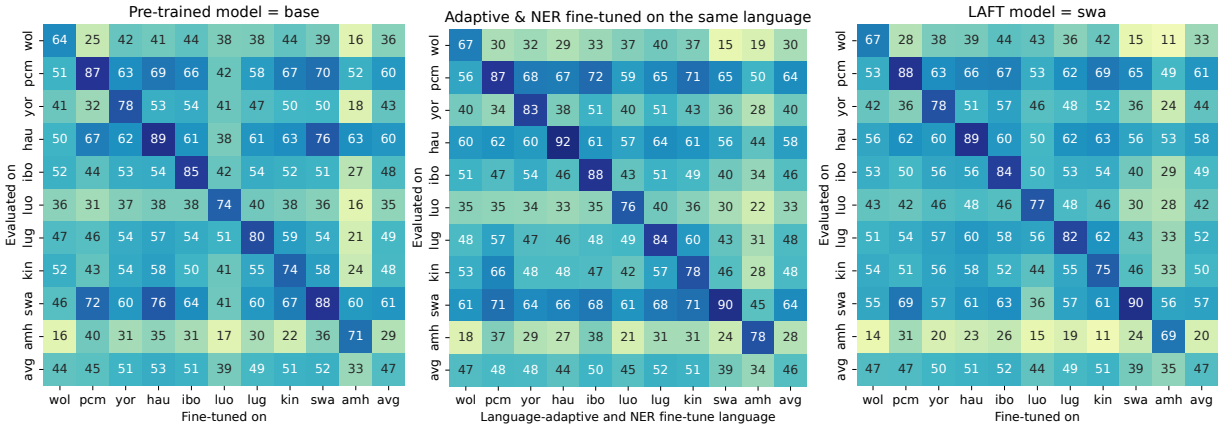
We additionally consider the other features from Lin et al. (2019). These results are shown in Figures 12-21. In particular, we consider the following features:

Geographic distance The distance between where the different languages are spoken, based on data from Glottolog (Hammarström et al., 2018).

Genetic distance The genealogical distance between the languages based on the Glottolog language tree.

Inventory distance The cosine distance between the feature vectors from the PHOIBLE database (Moran et al., 2014).

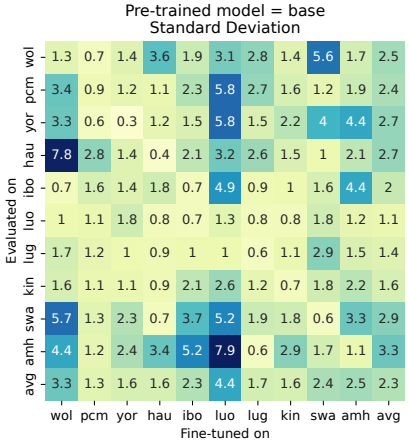
Syntactic distance The cosine distance between the feature vectors that represent the syntactic



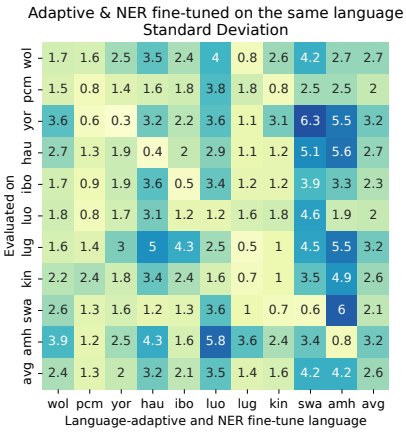
(a) base → X-axis (no LAFT)

(b) base → X-axis → X-axis

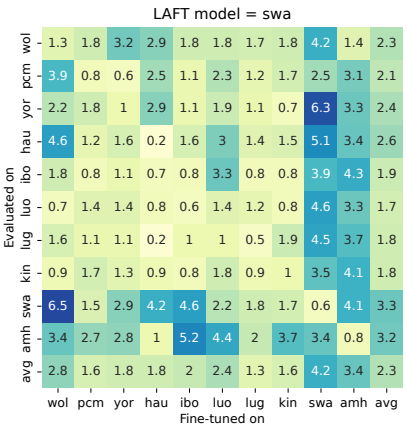
(c) base → swa → X-axis



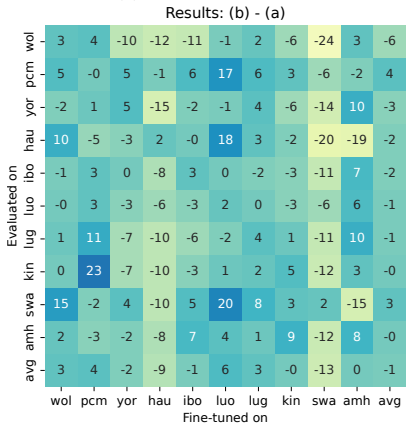
(d) Standard Deviation of (a)



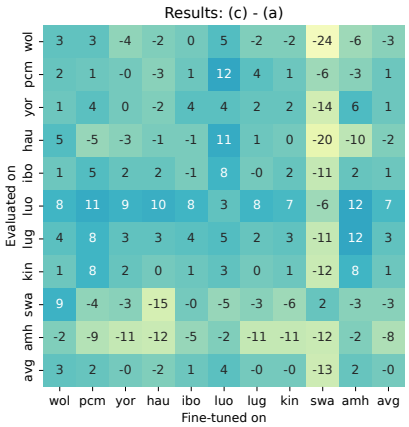
(e) Standard Deviation of (b)



(f) Standard Deviation of (c)



(g) Performance difference when adding language-adaptive fine-tuning. Swahili and Hausa transfer worse on average, while Luo improves.



(h) Performance difference between using a Swahili adaptively fine-tuned model and no language-adaptive fine-tuning after subsequent fine-tuning on NER data. Hausa fine-tuning performs much worse when evaluated on Swahili.

Figure 5: Heatmaps indicating the average performance over 5 seeds of specific models on specific languages (y-axis) after being fine-tuned on another language's NER data (x-axis). In general, we notice a large standard deviation, indicating that this process is unreliable. The bottom row shows the difference between one technique, and base, i.e. how much improvement this new model gives over using the base model. avg indicates the average transfer performance per row or column, respectively. Note that this calculates the average of the entire row or column excluding the diagonal, to be able to see the overall transfer performance at a glance.

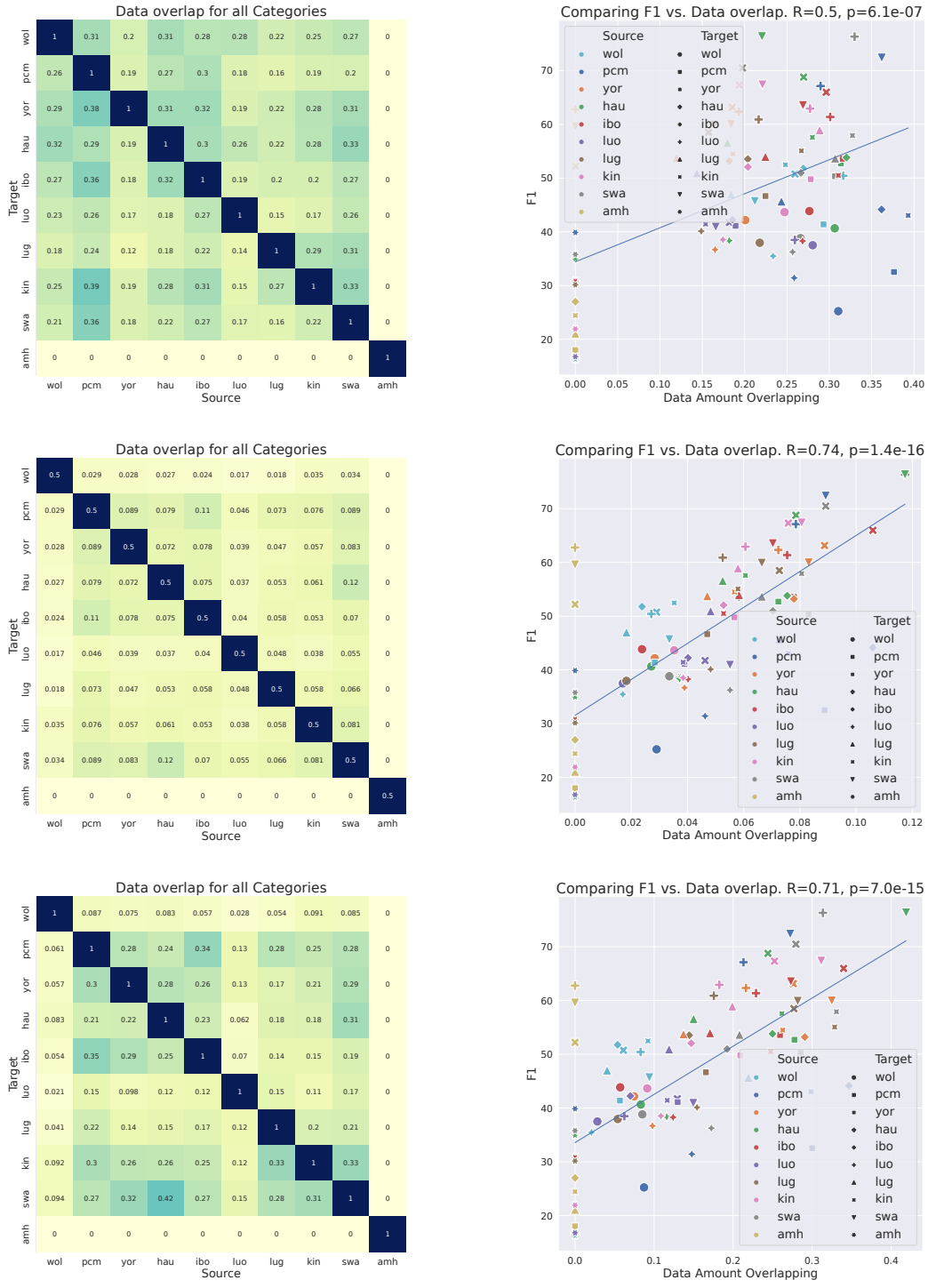
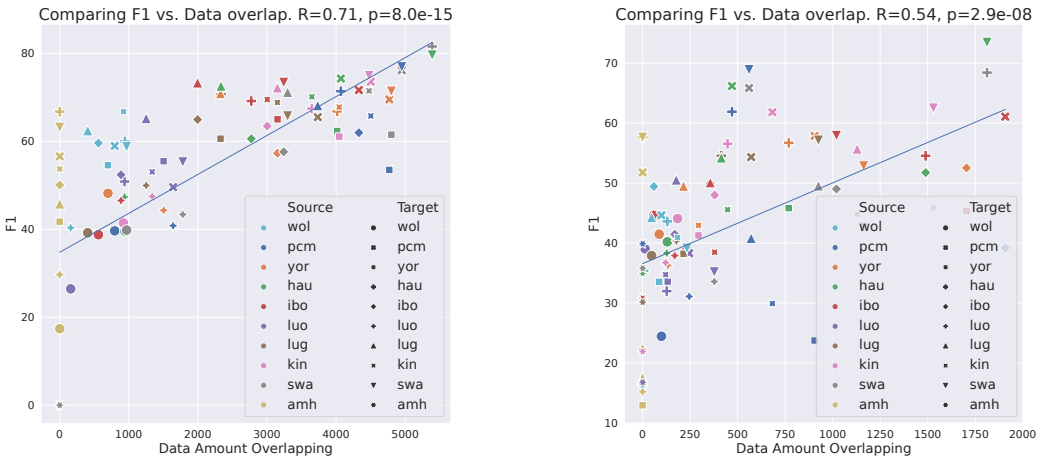


Figure 6: Overlap and correlation plots, for the (top) smallest and (middle) largest correlation coefficients, respectively. The bottom row contains the results when calculating overlap as the fraction of overlapping tokens in the target dataset, to contrast against the main text that used the absolute number. The top row used just the training dataset, counted overlap with respect to the number of occurrences in the source datasets, without considering labels. The middle row also used all of the unlabelled data, but calculated $\frac{|E_s \cap E_t|}{|E_s|+|E_t|}$ where E_s and E_t are the sets of unique entities for the source and transfer languages respectively (Lin et al., 2019)



(a) International (b) Local

Figure 7: Showing the correlation between overlap and performance when only considering (a) International and (b) Local tokens. Here, both the performance and overlap calculations only took these subsets of tokens into account, for instance, comparing the number of overlapping international tokens with the performance on international tokens.

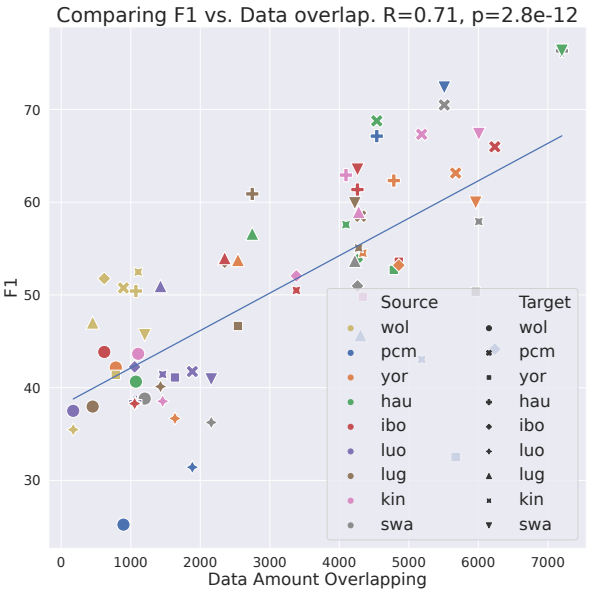


Figure 8: This shows the correlation between data overlap and performance for Amharic, as it has a different script and may thus be considered an outlier. The results are very similar to Fig. 2b.

properties of the languages, from the WALS database (Dryer and Haspelmath, 2013).

Phonological distance The cosine distance between the phonological feature vectors obtained from WALS and Ethnologue databases (Lewis, 2009).

Featural distance The cosine distance between feature vectors consisting of the 5 above features.

Source language dataset size The number of sentences in the source language's dataset.

Source Over Target Size Ratio The size (in number of sentences) of the source dataset divided by the size of the target dataset.

Source language number of entities The number of named entities in the source language's dataset.

Source Over Target entity Ratio The number of entities in the source dataset divided by the number of entities in the target dataset.

Overall, we find that data overlap has the highest correlation with transfer performance, with many other features not having a statistically significant correlation or a very small positive or negative correlation.

C.9 Combining Datasets

As an additional experiment, we train models on a combination of datasets, to see if this has any effect. The two options we consider here are: (1) training on the concatenation of all of the datasets; and (2) training on the concatenation, excluding the target language. We consider (1), as it involves training a single model, and we would like to investigate how well this model performs across all languages. For (2), we measure the effect of transferring from the nine other languages, as opposed to only the single-language transfer we have considered in the

rest of this paper. These results are shown in Table 6. Training on the target language, or on the concatenation of all languages performs quite well. The latter option also has the advantage of only being one model, whereas we need one model per language if we fine-tune only on data from one language. Fine-tuning the base model on the best transfer language performs worse than training on all of the datasets, excluding the target language. Finally, using a LAFT model for the target language and fine-tuning on all datasets except the target performs much better, and is the best transfer option we have considered.

Table 6: Here we show the results when training on a combination of datasets, compared to training on the best transfer language, or the target language itself. $cat - \{X\}$ indicates that the model trained on a combination of all of the datasets excluding the target language.

	wol	pcm	yor	hau	ibo	luo	lug	kin	swa	amh	avg
base $\rightarrow X$	64	87	78	89	85	74	80	74	88	71	79
base $\rightarrow$ X $\rightarrow X$	67	87	83	92	88	76	85	78	90	77	82
base $\rightarrow cat$	65	89	81	91	86	77	81	75	87	71	80
base $\rightarrow$ swa $\rightarrow cat$	66	89	80	91	85	80	81	76	89	69	81
base $\rightarrow$ X $\rightarrow cat - \{X\}$	57	80	68	78	76	45	73	68	74	49	67
base $\rightarrow cat - \{X\}$	43	78	60	74	60	44	63	59	75	33	59
base $\rightarrow best$	44	70	54	76	54	40	59	58	76	40	57

C.10 Representations

This additional experiment follows prior work by Hsu et al. (2019) by investigating the contextual word embeddings from the different models, specifically looking into how these embeddings change as we perform different fine-tuning operations. We take the last 4 layers from the language model (i.e. not the dense final layer) and use the sum of these hidden states to obtain a word vector (of size 768). We use the sentences from the dataset, and only extract the 4 different NER categories for computational reasons. We compute the mean vector per category, which we use in the following. To visualise the data, we show the results after performing PCA.

C.10.1 Variability

We found a large amount of variability when fine-tuning the models on different random seeds (see Figure 1 in the main text), so we next investigate the effect of different initialisations on the embeddings.

Fig. 9 shows the results for a few languages pairs, and immediately we can see that Fig. 9a has clusters corresponding to the different categories, even when using different seeds. Figs. 9b to 9d on the other hand cluster more toward seeds, so the categories differ when using different seeds. This could indicate that the Swahili model is more consistent and robust to random initialisations, and learns roughly the same embeddings for each seed. On the other hand, when fine-tuning from Kinyarwanda, Luo or Wolof, there is no clear clustering of categories (despite a relatively large amount of data overlap between Kinyarwanda and Hausa), suggesting that these models cannot distinguish Hausa categories very well (possibly substantiated by the poorer results shown in the main text).

Now, the above analysis is somewhat impacted by the final linear layers in the models – it is entirely possible that two models that have different embeddings also have different final layers and end up classifying examples exactly the same. We can, however, still use these experiments to extract some qualitative information about the embeddings of different languages. Furthermore, Figs. 9e and 9f – in which the language being investigated is the same as what the models trained on – contain results where the clustering is predominantly towards categories, bolstering the validity of this approach.

C.10.2 Different Languages and Models

Here we consider the same model and analyse the differences in embeddings from different languages, and how this evolves. For example, in Fig. 10a we see that for Nigerian Pidgin (which transferred well previously), the predominant clusters are again categories and not languages.

We next examine different models on the same language, specifically looking at what happens to these embeddings when a model is further fine-tuned. Fig. 10b shows that performing fine-tuning on models does affect the embeddings quite significantly, although there does still seem to be a similar relative positioning between the categories - almost as if in PCA, one principal component was the model used, and another was the category.

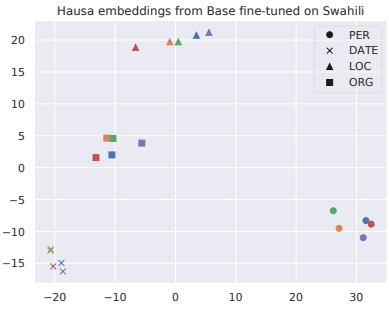
C.10.3 Transfer when fine-tuning on Amharic

In the main text, we observed that Amharic transferred quite well to Hausa, Swahili and Nigerian Pidgin. We now plot the embeddings of different languages, using *base* fine-tuned on Amharic in Fig. 11. In the top row, we have Hausa, Swahili – languages that Amharic was jointly pre-trained with – and Nigerian Pidgin, which is similar to English. In the bottom row, we have three other languages, not contained in the pre-training dataset. Clearly, the top row is clustered significantly more

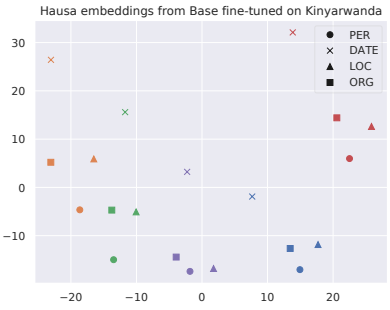
towards categories – indicating that the model manages to transfer knowledge from Amharic to these other languages. The bottom row demonstrates a clear clustering around the random seed – indicating no real information is transferred.

C.10.4 Summary

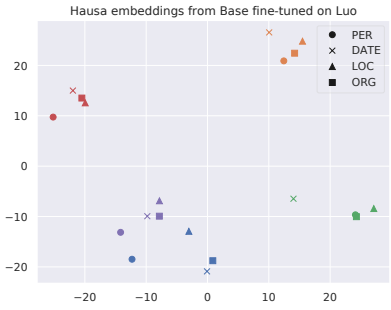
In summary, plotting the embeddings can shed some light on the representations learned by the model which, in many cases, provides some explanation for the results we obtained. Examining the embeddings can shed some light on this.



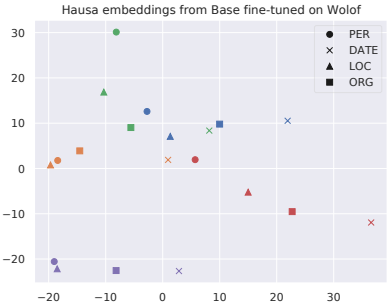
(a) Clear category clusters.



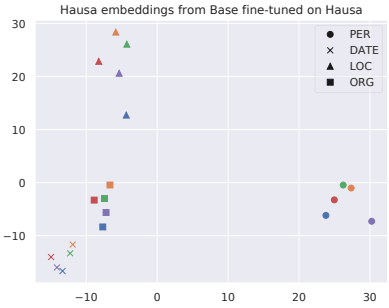
(b) Less clear clustering, but largely revolving around colours.



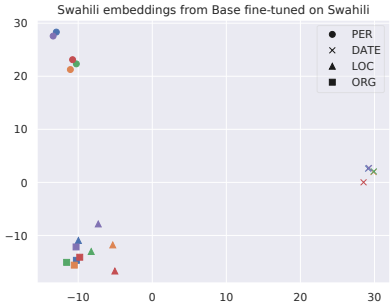
(c) Clear colour clusters.



(d) Clear colour clusters.

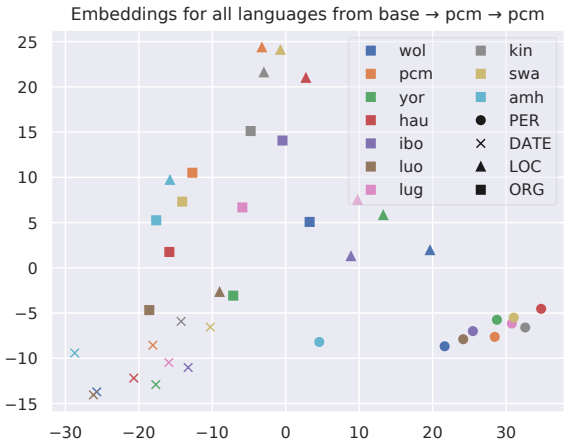


(e) Hausa embeddings from a model that was fine-tuned on Hausa – with clear category clusters.

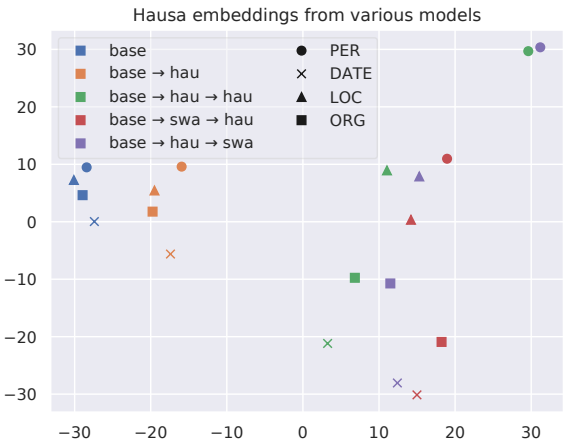


(f) Swahili embeddings from a Swahili fine-tuned model – with clear category clusters, although LOC and organisations are grouped close together.

Figure 9: Scatter plots of embeddings from different models, languages and categories. The shapes indicate different categories, whereas the colours indicate different starting points, i.e. seeds.



(a) Embeddings for different languages from a base → pcm → pcm model. This clustering is quite similar across different seeds.



(b) Hausa embeddings from different models.

Figure 10: Embeddings of (a) multiple languages with one model and (b) Hausa embeddings from different models after performing PCA.

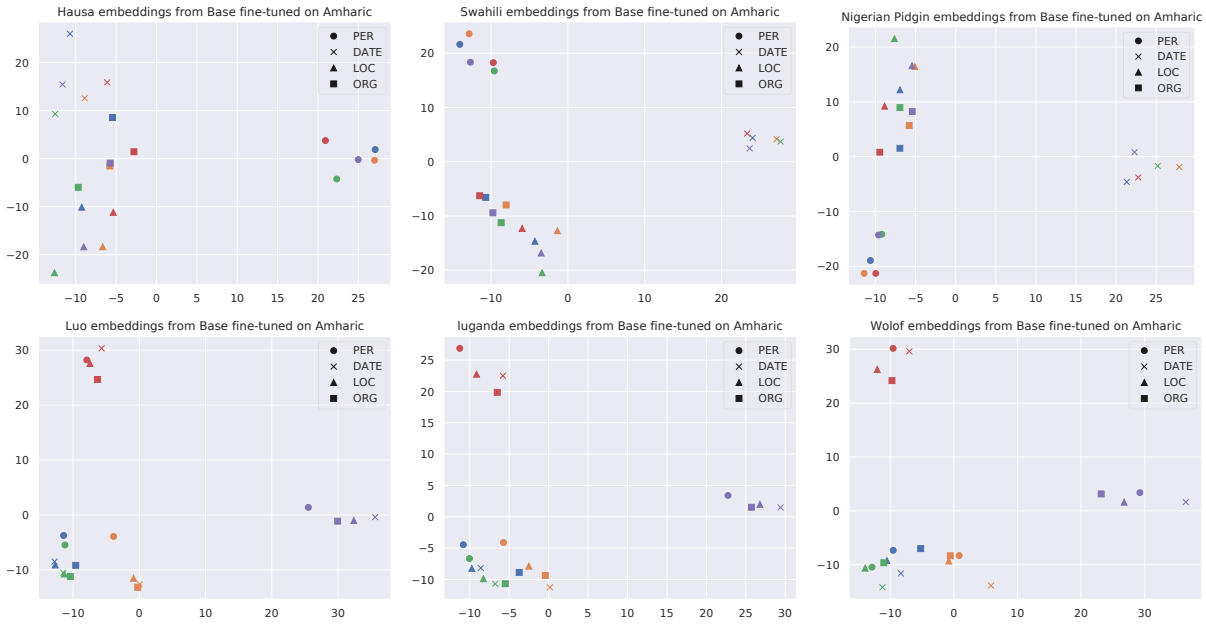


Figure 11: Showing embeddings of various languages, obtained from base → amh.

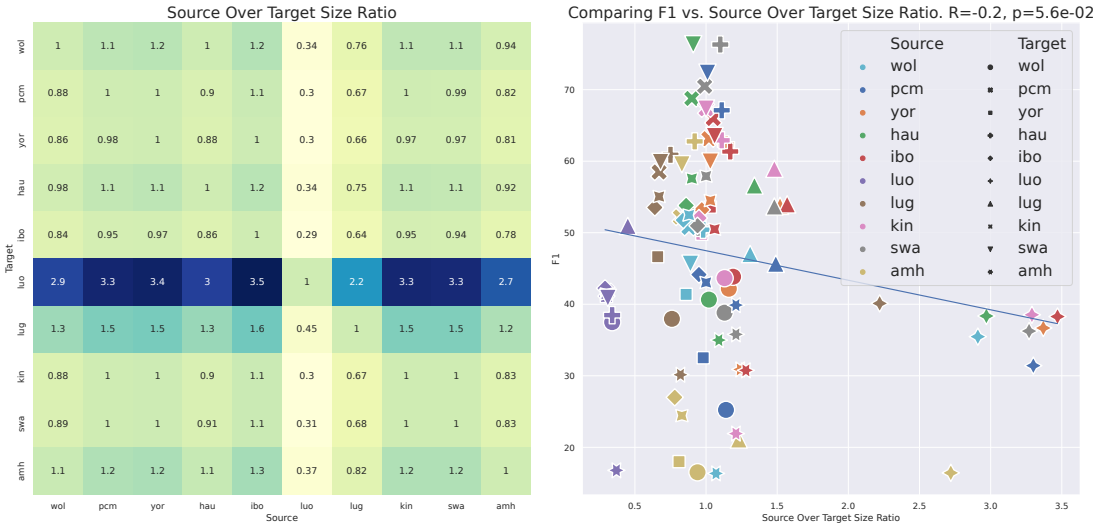


Figure 12: Source over Target Size Ratio

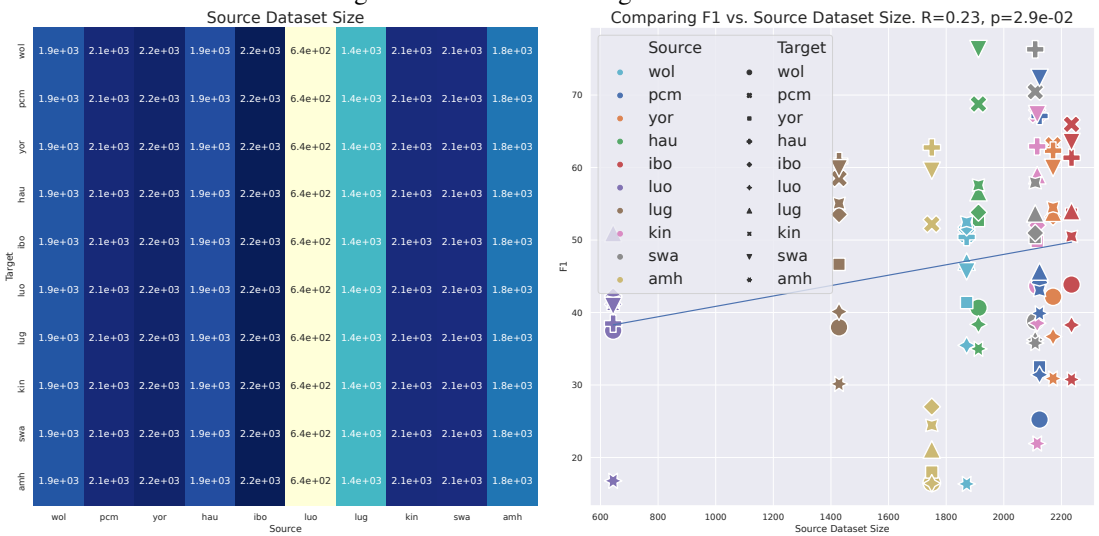


Figure 13: Source Language Dataset Size

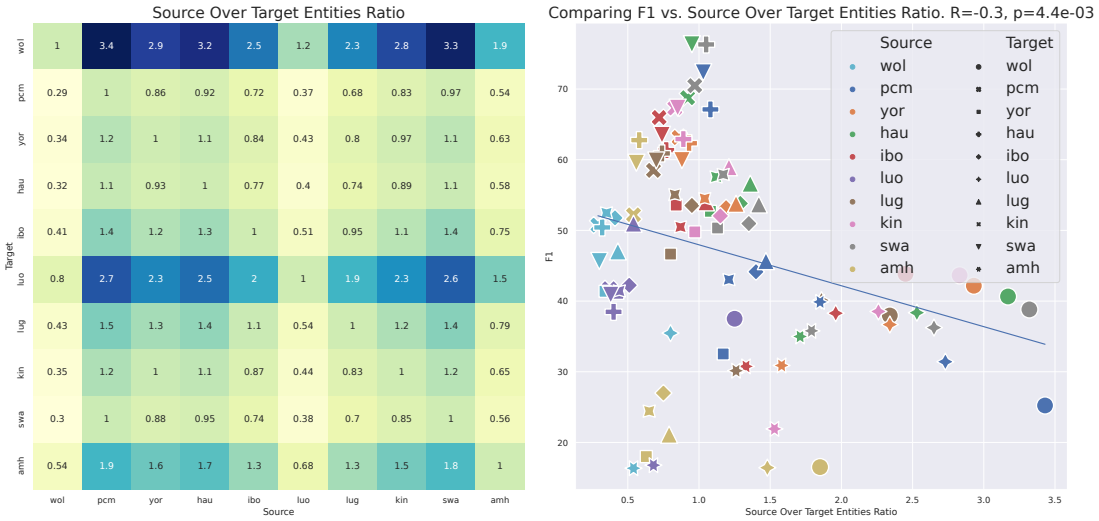


Figure 14: Source over Target Entities Ratio

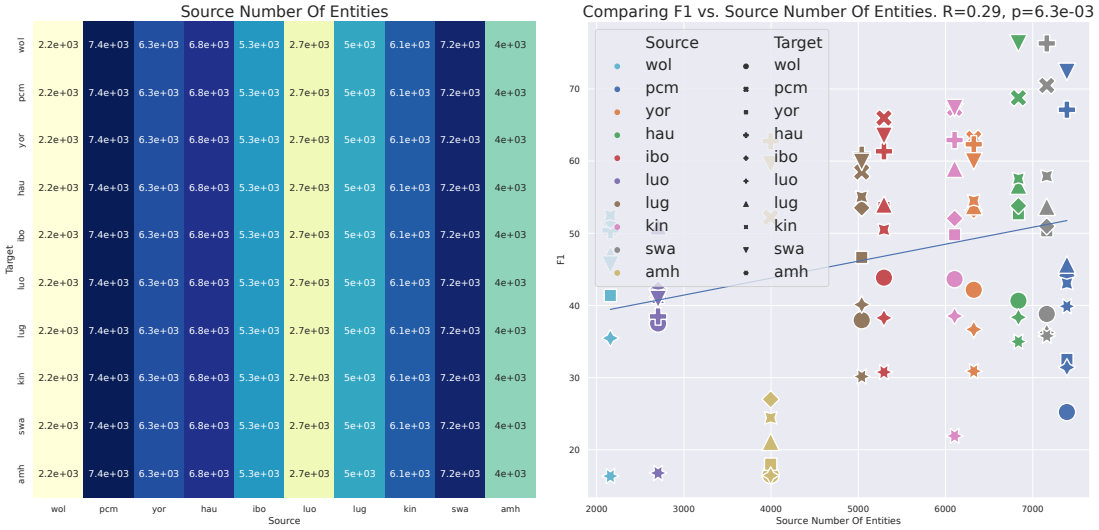


Figure 15: Number of entities in the source language's dataset

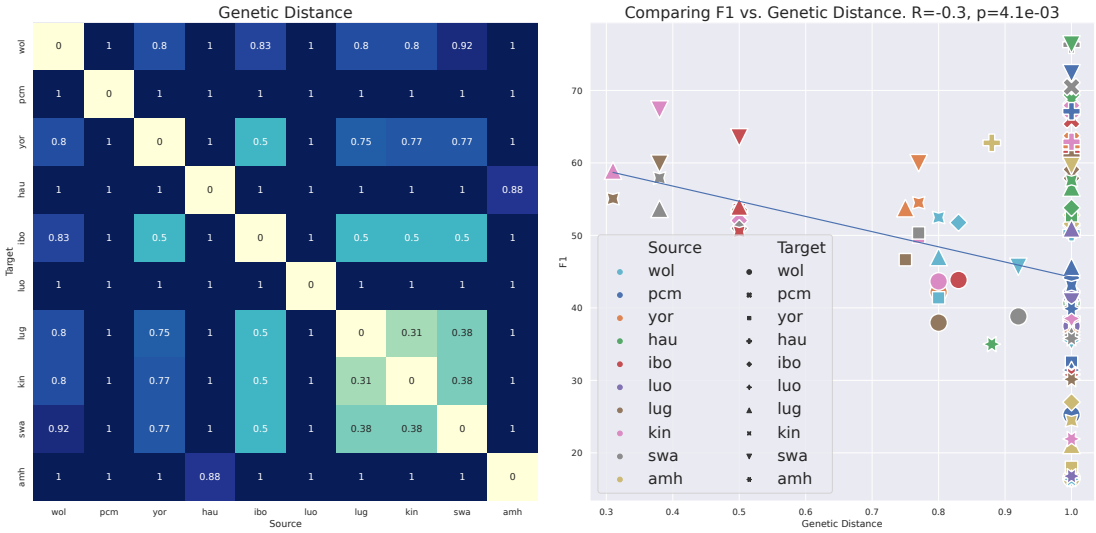


Figure 16: Genetic Distance

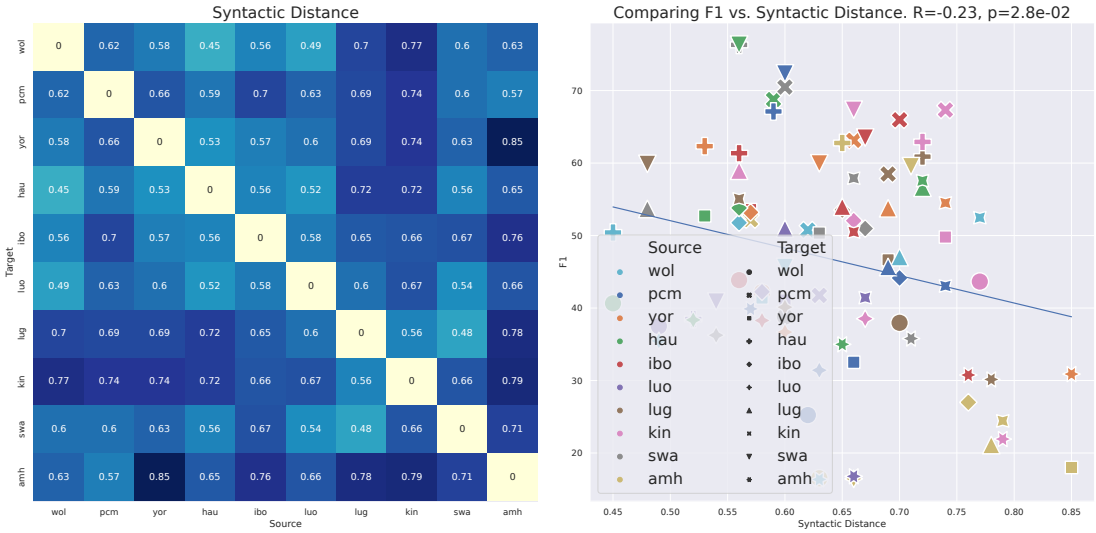


Figure 17: Syntactic Distance

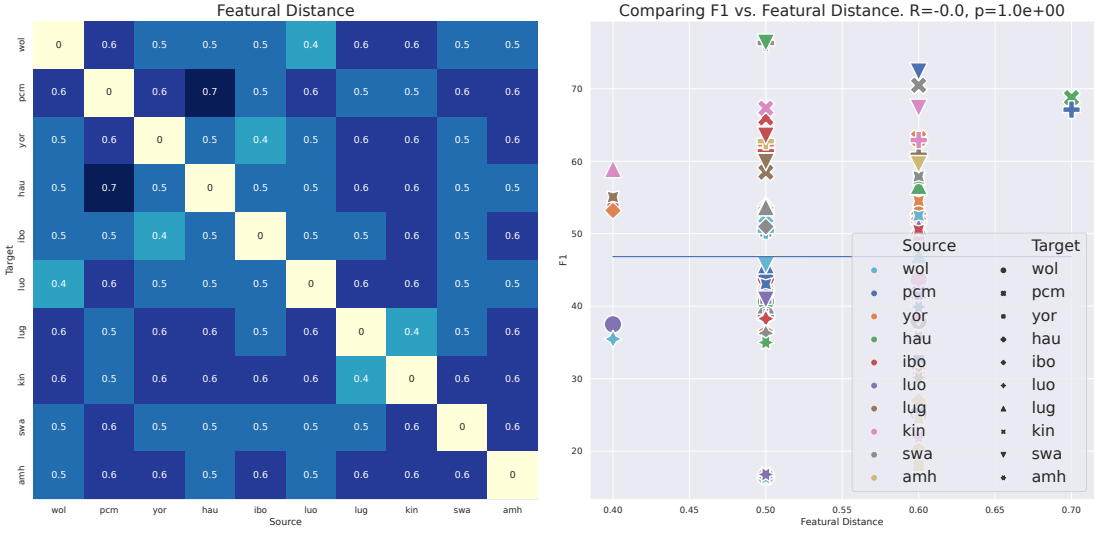


Figure 18: Featural Distance

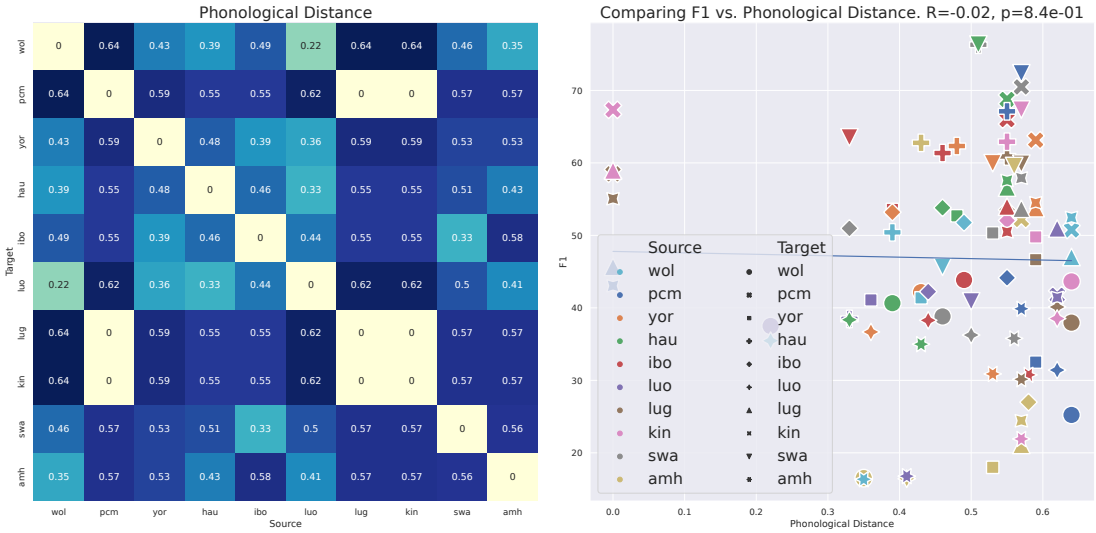


Figure 19: Phonological Distance

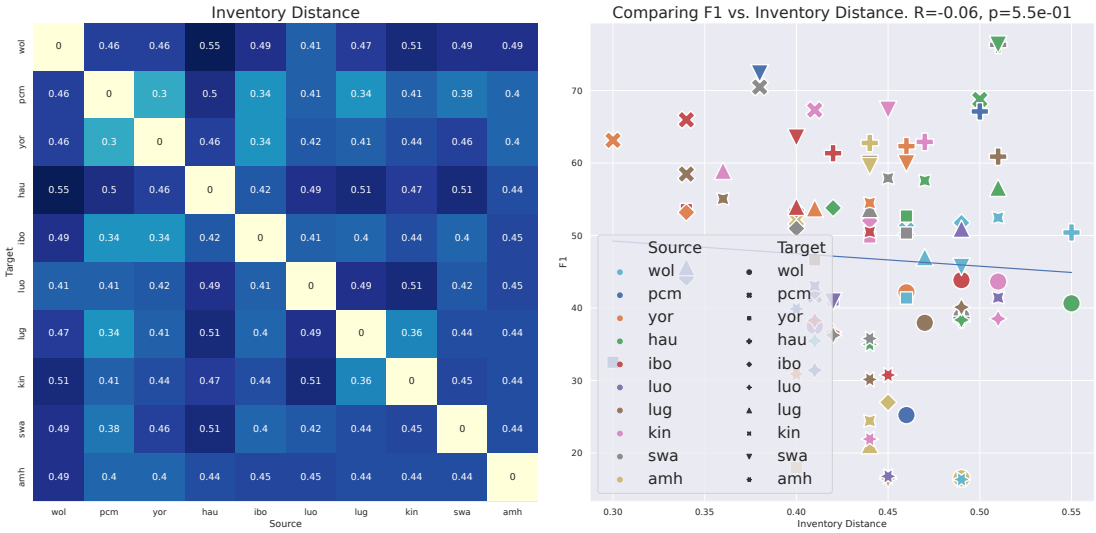


Figure 20: Inventory Distance

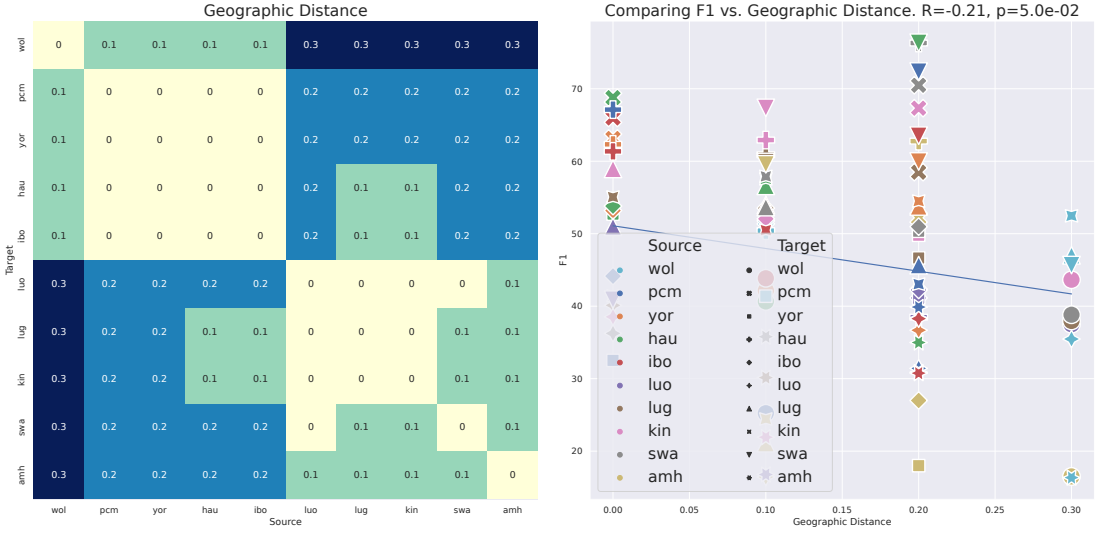


Figure 21: Geographic Distance