MasakhaNER 2.0: Africa-centric Transfer Learning for Named Entity Recognition

David Ifeoluwa Adelani[1,2,*], Graham Neubig[3], Sebastian Ruder[4], Shruti Rijhwani[3], Michael Beukman[5*], Chester Palen-Michel[6*], Constantine Lignos[6*], Jesujoba O. Alabi[1*], Shamsuddeen H. Muhammad[7*], Peter Nabende[8*], Cheikh M. Bamba Dione[9*], Andiswa Bukula[10], Rooweither Mabuya[10], Bonaventure F. P. Dossou[11*], Blessing Sibanda[*], Happy Buzaaba[12*], Jonathan Mukiibi[8*], Godson Kalipe[*], Derguene Mbaye[13*], Amelia Taylor[14*], Fatoumata Kabore[15*], Chris Chinenye Emezue[16*], Anuoluwapo Aremu[*], Perez Ogayo[3*], Catherine Gitau[*], Edwin Munkoh-Buabeng[17*], Victoire M. Koagne[*], Allahsera Auguste Tapo[18*], Tebogo Macucwa[19*], Vukosi Marivate[19*], Elvis Mboning[*], Tajuddeen Gwadabe[*], Tosin Adewumi[20*], Orevaoghene Ahia[21*], Joyce Nakatumba-Nabende[8*], Neo L. Mokono[19*], Ignatius Ezeani[22*], Chiamaka Chukwuneke[22*], Mofetoluwa Adeyemi[23*], Gilles Q. Hacheme[24*], Idris Abdulmumin[25*], Odunayo Ogundepo[23*], Oreen Yousuf[15*], Tatiana Moteu Ngoli[*], Dietrich Klakow[1]

[*]Masakhane NLP, [1]Saarland University, Germany, [2]University College London, UK, [3]Carnegie Mellon University, USA, [4]Google Research, [5]University of the Witwatersrand, South Africa, [6] Brandeis University, USA, [7]LIAAD-INESC TEC, Portugal, [8]Makerere University, Uganda [9]University of Bergen, Norway, [10]SADiLaR, South Africa, [11]Mila Quebec AI Institute, Canada, [12]RIKEN Center for AI Project, Japan, [13]Baamtu, Senegal, [14]Malawi University of Business and Applied Science, Malawi, [15]Uppsala University, Sweden, [16]TU Munich, Germany, [17]TU Clausthal, Germany, [18]Rochester Institute of Technology, USA, [19]University of Pretoria, South Africa, [20]Luleå University of Technology, Sweden, [21]University of Washington, USA, [22]Lancaster University, UK, [23]University of Waterloo, Canada, [24]Ai4innov, France, [25]Ahmadu Bello University, Nigeria.

Abstract

African languages are spoken by over a billion people, but are underrepresented in NLP research and development. The challenges impeding progress include the limited availability of annotated datasets, as well as a lack of understanding of the settings where current methods are effective. In this paper, we make progress towards solutions for these challenges, focusing on the task of named entity recognition (NER). We create the largest human-annotated NER dataset for 20 African languages, and we study the behavior of state-of-the-art cross-lingual transfer methods in an Africa-centric setting, demonstrating that the choice of source language significantly affects performance. We show that choosing the best transfer language improves zero-shot F1 scores by an average of 14 points across 20 languages compared to using English. Our results highlight the need for benchmark datasets and models that cover typologically-diverse African languages.

1 Introduction

Many African languages are spoken by millions or tens of millions of speakers. However, these languages are poorly represented in NLP research, and the development of NLP systems for African languages is often limited by the lack of datasets for training and evaluation (Adelani et al., 2021b).

Additionally, while there has been much recent work in using zero-shot cross-lingual transfer (Ponti et al., 2020; Pfeiffer et al., 2020; Ebrahimi et al., 2022) to improve performance on tasks for low-resource languages with multilingual pretrained language models (PLMs) (Devlin et al., 2019a; Conneau et al., 2020), the settings under which contemporary transfer learning methods work best are still unclear (Pruksachatkun et al., 2020; Lauscher et al., 2020; Xia et al., 2020). For example, several methods use English as the source language because of the availability of training data across many tasks (Hu et al., 2020; Ruder et al., 2021), but there is evidence that English is often not the best transfer language (Lin et al., 2019; de Vries et al., 2022; Oladipo et al., 2022), and the process of choosing the best source language to transfer from remains an open question.

There has been recent progress in creating benchmark datasets for training and evaluating models in African languages for several tasks such as machine translation (∀ et al., 2020; Reid et al., 2021; Adelani et al., 2021a, 2022; Abdulmumin et al., 2022), and sentiment analysis (Yimam et al., 2020; Muhammad et al., 2022). In this paper, we focus on the standard NLP task of named entity recognition (NER) because of its utility in downstream applications such as question answering and information extraction. For NER, annotated datasets exist only in a few African languages (Adelani et al., 2021b; Yohannes and Amagasa, 2022), the largest of which is the MasakhaNER dataset (Adelani et al., 2021b) (which we call MasakhaNER 1.0 in the remainder of the paper). While MasakhaNER 1.0 covers 10 African languages spoken mostly in West and East Africa, it does not include any languages spoken in Southern Africa, which have distinct syntactic and morphological characteristics and are spoken by 40 million people.

In this paper, we tackle two current challenges in developing NER models for African languages: (1) the lack of typologically- and geographically-diverse evaluation datasets for African languages; and (2) choosing the best transfer language for NER in an Africa-centric setting, which has not been previously explored in the literature.

To address the first challenge, we create the MasakhaNER 2.0 corpus, the largest human-annotated NER dataset for African languages. MasakhaNER 2.0 contains annotated text data from 20 languages widely spoken in Sub-Saharan Africa and is complementary to the languages present in previously existing datasets (e.g., Adelani et al., 2021b). We discuss our annotation methodology as well as perform benchmarking experiments on our dataset with state-of-the-art NER models based on multilingual PLMs.

In addition, to better understand the effect of source language on transfer learning, we extensively analyze different features that contribute to cross-lingual transfer, including linguistic characteristics of the languages (i.e., typological, geographical, and phylogenetic features) as well as data-dependent features such as entity overlap across source and target languages (Lin et al., 2019). We demonstrate that choosing the best transfer language(s) in both single-source and co-training setups leads to large improvements in NER performance in zero-shot settings; our experiments show an average of a 14 point increase in F1 score as compared to using English as source language across 20 target African languages. We release the data, code, and models on Github[1]

2 Related Work

African NER Datasets There are some human-annotated NER datasets for African languages such as the SaDiLAR NER corpus (Eiselen, 2016) covering 10 South African languages, LORELEI (Strassel and Tracey, 2016), which covers nine African languages but is not open-sourced, and some individual language efforts for Amharic (Jibril and Tantug, 2022), Yorùbá (Alabi et al., 2020), Hausa (Hedderich et al., 2020), and Tigrinya (Yohannes and Amagasa, 2022). Closest to our work is the MasakhaNER 1.0 corpus (Adelani et al., 2021b), which covers 10 widely spoken languages in the news domain, but excludes languages from the southern region of Africa like isiZulu, isiXhosa, and chiShona with distinct syntactic features (e.g., noun prefixes and capitalization in between words) which limits transfer learning from other languages. We include five languages from Southern Africa in our new corpus.

Cross-lingual Transfer Leveraging cross-lingual transfer has the potential to drastically improve model performance without requiring large amounts of data in the target language (Conneau et al., 2020) but it is not always clear from which language we must transfer from (Lin et al., 2019; de Vries et al., 2022). To this end, recent work investigates methods for selecting good transfer languages and informative features. For instance, token overlap between the source and target language is a useful predictor of transfer performance for some tasks (Lin et al., 2019; Wu and Dredze, 2019). Linguistic distance (Lin et al., 2019; de Vries et al., 2022), word order (K et al., 2020; Pires et al., 2019) and script differences (de Vries et al., 2022), and syntactic similarity (Karamolegkou and Stymne, 2021) have also been shown to impact performance. Another research direction attempts to build models of transfer performance that predicts the best transfer language for a target language by using some linguistic and data-dependent features (Lin et al., 2019; Ahuja et al., 2022).

3 Languages and Their Characteristics

3.1 Focus Languages

Table 1 provides an overview of the languages in our MasakhaNER 2.0 corpus. We focus on 20 Sub-Saharan African languages[2] with varying numbers of speakers (between 1M–100M) that are spoken by over 500M people in around 27 countries in

[1] https://github.com/masakhane-io/masakhane-ner/tree/main/MasakhaNER2.0

[2] Our selection was also constrained by the availability of volunteers that speak the languages in different NLP/AI communities in Africa.

Language	Family	African Region	No. of Speakers	Source	Train / dev / test	% Entities in Tokens	# Tokens
Bambara (bam)	NC / Mande	West	14M	MAFAND-MT (Adelani et al., 2022)	4462/ 638/ 1274	6.5	155,552
Ghomálá' (bbj)	NC / Grassfields	Central	1M	MAFAND-MT (Adelani et al., 2022)	3384/ 483/ 966	11.3	69,474
Éwé (ewe)	NC / Kwa	West	7M	MAFAND-MT (Adelani et al., 2022)	3505/ 501/ 1001	15.3	90420
Fon (fon)	NC / Volta-Niger	West	2M	MAFAND-MT (Adelani et al., 2022)	4343/ 621/ 1240	8.3	173,099
Hausa (hau)	Afro-Asiatic / Chadic	West	63M	Kano Focus and Freedom Radio	5716/ 816/ 1633	14.0	221,086
Igbo (ibo)	NC / Volta-Niger	West	27M	IgboRadio and Ka ỌdỊ Taa	7634/ 1090/ 2181	7.5	344,095
Kinyarwanda (kin)	NC / Bantu	East	10M	IGIHE, Rwanda	7825/ 1118/ 2235	12.6	245,933
Luganda (lug)	NC / Bantu	East	7M	MAFAND-MT (Adelani et al., 2022)	4942/ 706/ 1412	15.6	120,119
Luo (luo)	Nilo-Saharan	East	4M	MAFAND-MT (Adelani et al., 2022)	5161/ 737/ 1474	11.7	229,927
Mossi (mos)	NC / Gur	West	8M	MAFAND-MT (Adelani et al., 2022)	4532/ 648/ 1294	9.2	168,141
Naija (pcm)	English-Creole	West	75M	MAFAND-MT (Adelani et al., 2022)	5646/ 806/ 1613	9.4	206,404
Chichewa (nya)	NC / Bantu	South-East	14M	Nation Online Malawi	6250/ 893/ 1785	9.3	263,622
chiShona (sna)	NC / Bantu	South	12M	VOA Shona	6207/ 887/ 1773	16.2	195,834
Kiswahili (swa)	NC / Bantu	East & Central	98M	VOA Swahili	6593/ 942/ 1883	12.7	251,678
Setswana (tsn)	NC / Bantu	South	14M	MAFAND-MT (Adelani et al., 2022)	3489/ 499/ 996	8.8	141,069
Akan/Twi (twi)	NC / Kwa	West	9M	MAFAND-MT (Adelani et al., 2022)	4240/ 605/ 1211	6.3	155,985
Wolof (wol)	NC / Senegambia	West	5M	MAFAND-MT (Adelani et al., 2022)	4593/ 656/ 1312	7.4	181,048
isiXhosa (xho)	NC / Bantu	South	9M	Isolezwe Newspaper	5718/ 817/ 1633	15.1	127,222
Yorùbá (yor)	NC / Volta-Niger	West	42M	Voice of Nigeria and Asejere	6877/ 983/ 1964	11.4	244,144
isiZulu (zul)	NC / Bantu	South	27M	Isolezwe Newspaper	5848/ 836/ 1670	11.0	128,658

Table 1: **Languages and Data Splits for MasakhaNER 2.0 Corpus**. Language, family (NC: Niger-Congo), number of speakers, news source, and data split in number of sentences

the Western, Eastern, Central and Southern regions of Africa. The selected languages cover four language families. 17 languages belong to the Niger-Congo language family, and one language belongs to each of the Afro-Asiatic (Hausa), Nilo-Saharan (Luo), and English Creole (Naija) families. Although many languages belong to the Niger-Congo language family, they have different linguistic characteristics. For instance, Bantu languages (eight in our selection) make extensive use of affixes, unlike many languages of non-Bantu subgroups such as Gur, Kwa, and Volta-Niger.

3.2 Language Characteristics

Script and Word Order African languages mainly employ four major writing scripts: Latin, Arabic, N'ko and Ge'ez. Our focus languages mostly make use of the Latin script. While N'ko is still actively used by the Mande languages like Bambara, the most widely used writing script for the language is Latin. However, some languages use additional letters that go beyond the standard Latin script, e.g., "ɛ", "ɔ", "ŋ", "ẹ", and more than one character letters like "bv", "gb", "mpf", "ntsh". 17 of the languages are tonal except for Naija, Kiswahili and Wolof. Nine of the languages make use of diacritics (e.g., é, ë, ñ). All languages use the SVO word order, while Bambara additionally uses the SOV word order.

Morphology and Noun classes Many African languages are morphologically rich. According to the World Atlas of Language Structures (WALS; Nichols and Bickel, 2013), 16 of our languages employ strong prefixing or suffixing inflections. Niger-Congo languages are known for their system of noun classification. 12 of the languages *actively* make use of between 6–20 noun classes, including all Bantu languages, Ghomálá', Mossi, Akan and Wolof (Nurse and Philippson, 2006; Payne et al., 2017; Bodomo and Marfo, 2002; Babou and Loporcaro, 2016). While noun classes are often marked using affixes on the head word in Bantu languages, some non-Bantu languages, e.g., Wolof make use of a dependent such as a determiner that is not attached to the head word. For the other Niger-Congo languages such as Fon, Ewe, Igbo and Yorùbá, the use of noun classes is merely *vestigial* (Konoshenko and Shavarina, 2019). Three of our languages from the Southern Bantu family (chiShona, isiXhosa and isiZulu) capitalize proper names after the noun class prefix as in the language names themselves. This characteristic may limit transfer from languages without this feature as NER models overfit on capitalization (Mayhew et al., 2019). Appendix B provides more details regarding the languages' linguistic characteristics.

4 MasakhaNER 2.0 Corpus

4.1 Data source and collection

We annotate news articles from local sources. The choice of the news domain is based on the availability of data for many African languages and the variety of named entities types (e.g., person names and locations) as illustrated by popular datasets such as CoNLL-03 (Tjong Kim Sang and De Meulder, 2003).[3] Table 1 shows the sources and sizes

[3] We also considered using Wikipedia as our data source, but did not due to quality issues (Alabi et al., 2020).

of the data we use for annotation. Overall, we collected between 4.8K–11K sentences per language from either a monolingual or a translation corpus.

Monolingual corpus We collect a large monolingual corpus for nine languages, mostly from local news articles except for chiShona and Kiswahili texts, which were crawled from Voice of America (VOA) websites.[4]As Yorùbá text was missing diacritics, we asked native speakers to manually add diacritics before annotation. During data collection, we ensured that the articles are from a variety of topics e.g. politics, sports, culture, technology, society, and education. In total, we collected between 8K–11K sentences per language.

Translation corpus For the remaining languages for which we were unable to obtain sufficient amounts of monolingual data, we use a translation corpus, MAFAND-MT (Adelani et al., 2022), which consists of French and English news articles translated into 11 languages. We note that translationese may lead to undesired properties, e.g., unnaturalness. However, we did not observe serious issues during the annotation. The number of sentences is constrained by the size of the MAFAND-MT corpus, which is between 4,800–8,000.

4.2 NER Annotation Methodology

We annotated the collected monolingual texts with the ELISA annotation tool (Lin et al., 2018) with four entity types: Personal name (PER), Location (LOC), Organization (ORG), and date and time (DATE), similar to MasakhaNER 1.0 (Adelani et al., 2021b). We made use of the MUC-6 annotation guide.[5] The annotation was carried out by three native speakers per language recruited from AI/NLP communities in Africa. To ensure high-quality annotation, we recruited a language coordinator to supervise annotation in each language. We organized two online workshops to train language coordinators on the NER annotation. As part of the training, each coordinator annotated 100 English sentences, which were verified. Each coordinator then trained three annotators in their team using both English and African language texts with the support of the workshop organizers. All annotators and language coordinators received appropriate remuneration.[6]

At the end of annotation, language coordinators worked with their team to resolve disagreements using the adjudication function of ELISA, which ensures a high inter-annotator agreement score.

Lang.	Fleiss' Kappa	QC flags fixed?	Lang.	Fleiss' Kappa	QC flags fixed?
bam	0.980	✗	pcm	0.966	✗
bbj	1.000	✓	nya	0.988	✓
ewe	0.991	✓	sna	0.957	✓
fon	0.941	✗	swa	0.974	✓
hau	0.950	✗	tsn	0.962	✗
ibo	0.965	✗	twi	0.932	✗
kin	0.943	✗	wol	0.979	✓
lug	0.950	✓	xho	0.945	✓
luo	0.907	✗	yor	0.950	✓
mos	0.927	✗	zul	0.953	✓

Table 2: Inter-annotator agreement for our datasets calculated using Fleiss' kappa κ at the entity level before adjudication. QC flags (✓) are the languages that fixed the annotations for all **Q**uality **C**ontrol flagged tokens.

4.3 Quality Control

As discussed in subsection 4.2, language coordinators helped resolve several disagreements in annotation prior to quality control. Table 2 reports the Fleiss Kappa score after the intervention of language coordinators (i.e. post-intervention score). The pre-intervention Fleiss Kappa score was much lower. For example, for pcm, the pre-intervention Fleiss Kappa score was 0.648 and improved to 0.966 after the language coordinator discussed the disagreements with the annotators.

For the quality control, annotations were automatically adjudicated when there was agreement, but were flagged for further review when annotators disagreed on mention spans or types. The process for reviewing and fixing quality control issues was voluntary and so not all languages were further reviewed (see Table 2).

We automatically identified positions in the annotation that were more likely to be annotation errors and flagged them for further review and correction. The automatic process flags tokens that are commonly annotated as a named entity but were not marked as a named entity in a specific position. For example, the token *Province* may appear commonly as part of a named entity and infrequently not as a named entity, so when it is seen as not marked it was flagged. Similarly, we flagged tokens that had near-zero entropy with regard to a certain entity type, for example a token almost always annotated as ORG but very rarely annotated as PER. We also flagged potential sentence boundary errors by identifying sentences with few tokens

[4] www.voashona.com/ and www.voaswahili.com/

[5] https://cs.nyu.edu/~grishman/muc6.html

[6] $10 per hour, annotating about 200 sentences per hour.

PLM	# Lang.	Languages in MasakhaNER 2.0
mBERT-cased (110M)	104	**swa**, **yor**
XLM-R-base/large (270M / 550M)	100	**hau**, **swa**, **xho**
mDeBERTaV3 (276M)	100	**hau**, **swa**, **xho**
RemBERT (575M)	110	**hau**, **ibo**, **nya**, **sna**, **swa**, **xho**, **yor**, **zul**
AfriBERTa (126M)	11	**hau**, **ibo**, **kin**, **pcm**, **swa**, **yor**
AfroXLMR-base/large (270M/550M)	20	**hau**, **ibo**, **kin**, **nya**, **pcm**, **sna**, **swa**, **xho**, **yor**, **zul**

Table 3: Language coverage and size for PLMs.

or sentences which end in a token that appears to be an abbreviation or acronym. As shown in Table 2, before further adjudication and correction there was already relatively high inter-annotator agreement measured by Fleiss' Kappa at the mention level.

After quality control, we divided the annotation into training, development, and test splits consisting of 70%, 10%, and 20% of the data respectively. Appendix A provide details on the number of tokens per entity (PER, LOC, ORG, and DATE) and the fraction of entities in the tokens.

5 Baseline Experiments

5.1 Baseline Models

As baselines, we fine-tune several multilingual PLMs including mBERT (Devlin et al., 2019b), XLM-R (base & large; Conneau et al., 2020), mDeBERTaV3 (He et al., 2021), AfriBERTa (Ogueji et al., 2021), RemBERT (Chung et al., 2021), and AfroXLM-R (base & large; Alabi et al., 2022). We fine-tune the PLMs on each language's training data and evaluate performance on the test set using HuggingFace Transformers (Wolf et al., 2020).

Massively multilingual PLMs Table 3 shows the language coverage and size of different massively multilingual PLMs trained on 100–110 languages. mBERT was pre-trained using masked language modeling (MLM) and next-sentence prediction on 104 languages, including `swa` and `yor`. RemBERT was trained with a similar objective, but makes use of a larger output embedding size during pre-training and covers more African languages. XLM-R was trained only with MLM on 100 languages and on a larger pre-training corpus. mDeBERTaV3 makes use of ELECTRA-style (Clark et al., 2020) pre-training, i.e., a replaced token detection (RTD) objective instead of MLM.

Africa-centric multilingual PLMs We also obtained NER models by fine-tuning two PLMs that are pre-trained on African languages. AfriBERTa (Ogueji et al., 2021) was pre-trained on less than 1 GB of text covering 11 African languages, including six of our focus languages, and has shown impressive performance on NER and sentiment classification for languages in its pre-training data (Adelani et al., 2021b; Muhammad et al., 2022). AfroXLM-R (Alabi et al., 2022) is a language-adapted (Pfeiffer et al., 2020) version of XLM-R that was fine-tuned on 17 African languages and three high-resource languages widely spoken in Africa ("eng", "fra", and "ara"). Appendix J provides the model hyper-parameters for fine-tuning the PLMs.

5.2 Baseline Results

Table 4 shows the results of training NER models on each language using the eight multilingual and Africa-centric PLMs. All PLMs provided good performance in general. However, we observed worse results for mBERT and AfriBERTa especially for languages they were not pre-trained on. For instance, both models performed between 6–12 F1 worse for `bbj`, `wol` or `zul` compared to XLM-R-base. We hypothesize that the performance drop is largely due to the small number of African languages covered by mBERT as well as AfriBERTa's comparatively small model capacity. XLM-R-base gave much better performance ($>$ 1.0 F1) on average compared to mBERT and AfriBERTa. We found the larger variants of mBERT and XLM-R, i.e., RemBERT and XLM-R-large to give much better performance ($>$ 2.0 F1) than the smaller models. Their larger capacity facilitates positive transfer, yielding better performance for unseen languages. Surprisingly, mDeBERTaV3 provided slightly better results than XLM-R-large and RemBERT despite its smaller size, demonstrating the benefits of the RTD pre-training (Clark et al., 2020).

The best PLM is AfroXLM-R-large, which outperforms mDeBERTaV3, RemBERT and AfriBERTa by +1.3 F1, +2.0 F1 and +4.0 F1 respectively. Even the performance of its smaller variant, AfroXLM-R-base is comparable to mDeBERTaV3. Overall, our baseline results highlight that large PLMs, PLM with improved pre-training objectives, and PLMs pre-trained on the target African languages are able to achieve reasonable baseline performance. Combining these criteria provides improved performance, such as AfroXLM-R-large, a

Model	bam	bbj	ewe	fon	hau	ibo	kin	lug	luo	mos	nya	pcm	sna	swa	tsn	twi	wol	xho	yor	zul	AVG
PLM pre-trained on 100+ world languages																					
mBERT	78.9	60.6	86.9	79.9	85.2	87.3	83.2	85.5	80.3	71.4	88.6	87.1	92.4	92.1	86.4	75.7	79.9	85.0	87.7	81.7	$82.8_{\pm0.2}$
XLM-R-base	78.7	72.3	88.5	81.9	83.8	87.8	82.5	86.7	79.3	72.7	89.9	88.5	93.6	92.2	86.1	78.7	82.3	87.0	85.8	84.6	$84.1_{\pm0.1}$
XLM-R-large	79.4	**75.2**	89.1	81.6	86.3	87.2	84.3	88.1	80.8	74.9	90.5	89.2	94.2	92.6	85.9	79.8	82.0	88.1	86.6	86.7	$85.1_{\pm0.5}$
RemBERT	80.1	74.2	89.2	82.2	84.7	86.4	85.2	87.1	80.4	72.7	91.4	89.5	94.8	92.0	87.0	78.5	83.6	88.3	87.2	85.5	$85.0_{\pm0.2}$
mDeBERTaV3	80.2	73.5	89.8	81.8	85.4	88.8	86.4	88.7	80.3	**76.4**	92.0	**90.1**	95.5	92.5	86.5	79.4	83.6	88.1	86.7	88.3	$85.7_{\pm0.2}$
PLM pre-trained on African languages																					
AfriBERTa	78.6	71.0	86.9	79.9	85.2	87.3	83.2	85.5	78.4	71.4	88.6	87.1	92.4	92.1	83.2	75.7	79.9	85.0	87.7	81.7	$83.0_{\pm0.2}$
AfroXLMR-base	79.6	73.3	89.2	82.3	86.6	88.5	86.1	88.1	80.8	74.4	91.9	89.3	95.7	92.3	87.7	78.9	84.9	88.6	88.3	88.4	$85.7_{\pm0.1}$
AfroXLMR-large	**82.2**	74.8	**90.3**	**82.7**	**87.4**	**89.6**	**87.5**	**89.6**	**82.2**	**76.4**	**92.4**	89.7	**96.2**	**92.7**	**89.4**	**81.1**	**86.8**	**89.9**	**89.3**	**90.6**	$\mathbf{87.0}_{\pm0.2}$

Table 4: **NER Baselines on MasakhaNER 2.0**. We compare several multilingual PLMs including the ones trained on African languages. Average is over 5 runs.

Train Lang.	Data	bam	bbj	ewe	fon	hau	ibo	kin	lug	luo	mos	nya	pcm	sna	swa	tsn	twi	wol	xho	yor	zul	AVG
Language in MasakhaNER 1.0?		✗	✗	✗	✗	✓	✓	✓	✓	✓	✗	✗	✓	✗	✓	✗	✗	✓	✗	✓	✗	-
Evaluation on MasakhaNER 2.0 test set																						
(a) MasakhaNER 1.0	MasakhaNER 1.0	52.2	48.4	78.3	52.9	76.9	86.0	77.6	83.2	68.6	55.0	82.1	86.7	49.6	89.4	80.0	56.6	73.6	56.9	69.4	69.9	$69.7_{\pm0.6}$
(b) MasakhaNER 1.0	MasakhaNER 2.0	50.9	49.8	76.2	57.1	88.7	90.1	87.6	90.0	82.7	49.6	80.4	90.2	42.5	93.1	79.4	57.3	87.0	47.4	89.7	64.3	$72.7_{\pm0.6}$
(c) MasakhaNER 2.0	MasakhaNER 2.0	**82.3**	**75.5**	**89.5**	**83.2**	**87.7**	**92.3**	**87.2**	**89.1**	**81.8**	**75.3**	**92.2**	**89.9**	**95.9**	**93.1**	**89.5**	**78.8**	**86.4**	**89.7**	**89.1**	**90.7**	$\mathbf{87.0}_{\pm1.2}$
Evaluation on MasakhaNER 1.0 test set																						
(a) MasakhaNER 1.0	MasakhaNER 1.0	–	–	–	–	92.1	89.2	79.1	86.0	80.0	–	–	91.2	–	89.5	–	–	70.8	–	85.0	–	$84.8_{\pm0.3}$
(b) MasakhaNER 1.0	MasakhaNER 2.0	–	–	–	–	80.8	84.6	77.7	79.0	67.0	–	–	88.0	–	86.3	–	–	71.6	–	85.0	–	$80.0_{\pm0.3}$
(c) MasakhaNER 2.0	MasakhaNER 2.0	–	–	–	–	80.4	84.3	77.0	79.8	67.6	–	–	87.9	–	86.5	–	–	72.1	–	84.8	–	$80.1_{\pm0.8}$

Table 5: **Multilingual evaluation on African NER datasets**. We compare the performance of AfroXLM-R-large trained on languages of MasakhaNER 2.0 and MasakhaNER 1.0 and evaluated both on the same and on the other dataset. The first column indicate the languages used for training (the 10 languages from MasakhaNER or the 20 languages from MasakhaNER 2.0). The second column indicates the training data. Average is over 5 runs.

large PLM trained on several African languages.

5.3 Entity-level Analysis of MasakhaNER 2.0

5.3.1 Error Analysis with ExplainaBoard

Furthermore, using ExplainaBoard (Liu et al., 2021), we analysed the best three baseline NER models: AfroXLM-R-large, mDeBERTaV3, and XLM-R-large. We discovered that 2-token entities were easier to predict accurately than lengthier entities (4 or more words). Moreover, the result shows that all the models have difficulty predicting zero-frequency entities effectively (entities with no occurrences in the training set). Interestingly, AfroXLMR-large is significantly better than other models for zero-frequency entities, suggesting that training PLMs on African languages promotes generalization to unseen entities. Finally, we observed that the three models perform better when predicting PER and LOC entities compared to ORG and DATE entities by up to (+5%). Appendix D provides more details on the error analysis.

5.3.2 Dataset Geography of Entities

Next, we analyse the geographical representativeness of the entities in our dataset, specifically, we measure the count of entities based on the countries they originate from. Following the approach of Faisal et al. (2022), we first performed entity linking of named entities present in our dataset to Wikidata IDs using mGenre (De Cao et al., 2022), followed by mapping Wikidata IDs to countries.

Figure 1 shows the result of number of entities per continent and the top-10 countries with the largest representation of entities. Over 50% of the entities are from Africa, followed by Europe. This shows that the entities of MasakhaNER 2.0 properly represent the African continent. Seven out of the top-10 countries are from Africa, but also includes USA, United Kingdom and France.

5.4 Transfer Between African NER Datasets

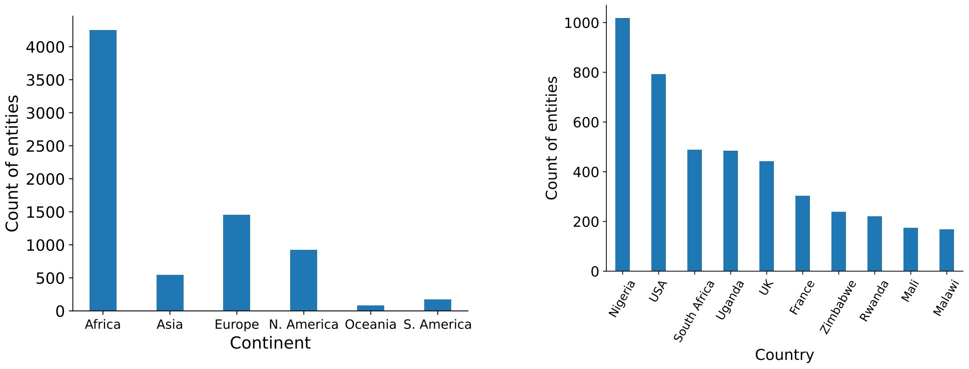
African languages have a diverse set of linguistic characteristics. To demonstrate this heterogeneity, we perform a transfer learning experiment where we compare the performance of multilingual NER models jointly trained on the languages of MasakhaNER 1.0 or MasakhaNER 2.0 and perform zero-shot evaluation on both test sets. We consider three experimental settings:

(a) Train on all languages in MasakhaNER 1.0 using MasakhaNER 1.0 training data.

(b) Train on the languages in MasakhaNER 1.0 (excl. "amh") using the MasakhaNER 2.0 training data.

(c) Train on all languages in MasakhaNER 2.0 using MasakhaNER 2.0 training data.

Table 5 shows the result of the three settings. When evaluating on the MasakhaNER 2.0 test set in set-



(a) Number of entities per continent

(b) Top-10 countries

Figure 1: **Number of entities per continent and the top-10 countries with the largest number of entities**

ting (a), the performance is mostly high (> 65 F1) for languages in MasakhaNER 1.0. Most of the languages that are not in MasakhaNER 1.0 have worse zero-shot performance, typically between 48 − 60 F1 except for ewe, nya, tsn, and zul with over 69 F1. Making use of a larger dataset, i.e., setting (b) from MasakhaNER 2.0 only provides a small improvement (+3 F1). The evaluation on setting (c) shows a large gap of about 15 F1 and 17 F1 compared to settings (b) and (a) on the MasakhaNER 2.0 test set respectively, especially for Southern Bantu languages like sna and xho. On the MasakhaNER 1.0 test set, training on the in-distribution MasakhaNER 1.0 languages and training set achieves the best performance. However, the performance gap compared to training on the MasakhaNER 2.0 data is much smaller. Overall, these results demonstrate the need to create large benchmark datasets (like MasakhaNER 2.0) covering diverse languages with different linguistic characteristics, particularly for the Africa.

6 Cross-Lingual Transfer

The success of cross-lingual transfer either in zero or few-shot settings depends on several factors, including an appropriate selection of the best source language. Several attempts at cross-lingual transfer make use of English as the source language due to its availability of training data. However, English is unrepresentative of African languages and transfer performance is often lower for distant languages (Adelani et al., 2021b).

6.1 Choosing Transfer Languages for NER

Here, we follow the approach of Lin et al. (2019), LangRank, that uses source-target transfer evaluation scores and data-dependent features such as dataset size and entity overlap, and six different linguistic distance measures based on lang2vec (Littell et al., 2017) such as geographic distance (d_{geo}), genetic distance (d_{gen}), inventory distance (d_{inv}), syntactic distance (d_{syn}), phonological distance (d_{pho}), and featural distance (d_{fea}). We provide definitions of the features in Appendix E. LangRank is trained using these features to determine the best transfer language in a leave-one-out setting where, for each target language, we train on all other languages except the target language. We compute transfer F1 scores from a set of N transfer (source) languages and evaluate on N target languages, yielding $N \times N$ transfer scores.

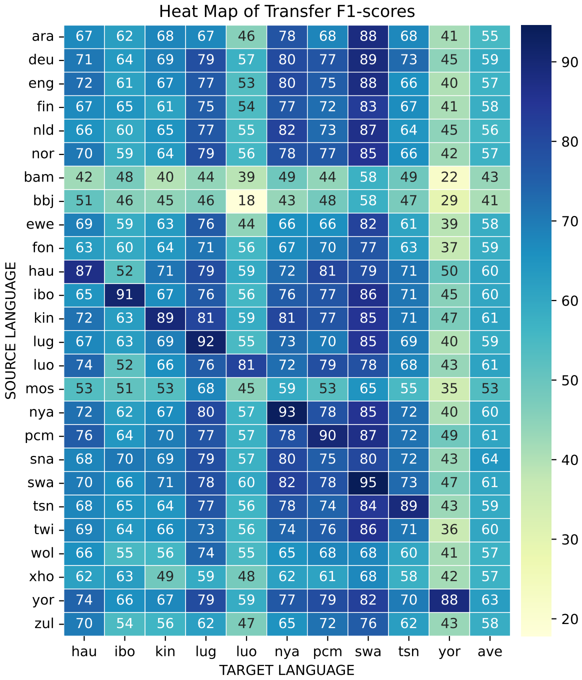


Figure 2: **Zero-shot Transfer** from several source languages to African languages for 10 languages in MasakhaNER 2.0 and the average (ave) over all 20 languages. Appendix G shows results for each of the 20 languages.

Target Lang.	Top-2 Transf. Lang	Top-2 LangRank Model	Top-3 features selected by LangRank model Lang 1; Lang 2	Target Lang. F1	Top-1 LangRank Lang. F1	Top-2 LangRank Lang. F1	Top-2 Transf. Lang. F1	Best Transf. F1	Second Best Transf. F1	eng Tranf. F1
bam	twi, fon	wol, fon	(d_{geo}, d_{inv}, sr); (d_{geo}, sr, d_{pho})	80.4	47.1	52.8	**55.1**	54.3	53.0	38.4
bbj	fon, ewe	twi, ewe	$(s_{tf}, d_{syn}, d_{geo})$; (s_{tf}, d_{geo}, sr)	72.9	53.9	58.8	**60.1**	59.8	58.4	45.8
ewe	swa, twi	pcm, swa	(d_{geo}, s_{tf}, sr); (eo, d_{geo}, s_{tf})	91.7	78.1	81.1	**83.9**	81.6	81.5	76.4
fon	mos, bbj	yor, ewe	(d_{geo}, d_{syn}, sr); $(s_{tf}, d_{geo}, d_{gen})$	84.9	58.4	64.9	**69.9**	65.4	62.0	50.6
hau	pcm, yor	yor, swa	(d_{geo}, sr, eo); (eo, sr, s_{tf})	86.9	74.3	74.8	**77.4**	75.9	74.3	72.4
ibo	sna, yor	pcm, kin	(eo, d_{geo}, s_{tf}); (d_{geo}, sr, eo)	91.0	64.2	63.9	**77.1**	70.4	66.0	61.4
kin	hau, swa	sna, yor	(eo, d_{geo}, s_{tf}); (eo, s_{tf}, sr)	89.5	69.2	71.8	**74.0**	71.1	70.6	67.4
lug	kin, nya	luo, zul	(d_{geo}, sr, eo); (d_{syn}, d_{geo}, sr)	91.5	75.9	78.1	**82.1**	81.1	80.0	76.5
luo	swa, hau	lug, sna	(d_{geo}, sr, eo); (d_{geo}, eo, sr)	81.2	54.9	61.6	**61.1**	60.4	59.5	53.4
mos	fon, ewe	yor, fon	(d_{geo}, d_{inv}, sr); (d_{geo}, s_{tf}, sr)	78.9	50.8	62.5	**65.6**	64.2	60.4	45.4
nya	swa, nld	zul, sna	(eo, d_{geo}, sr); (d_{geo}, eo, d_{syn})	93.5	65.5	81.5	**81.8**	81.8	81.7	80.1
pcm	hau, yor	eng, yor	(eo, d_{gen}, d_{syn}); (eo, d_{geo}, sr)	89.9	75.5	79.9	**81.8**	80.5	79.1	75.5
sna	zul, xho	swa, zul	(eo, sr, s_{tf}); (d_{geo}, sr, eo)	96.0	32.4	80.0	**80.0**	77.5	74.5	37.1
swa	deu, ara	ita, nld	(sr, d_{inv}, eo); (eo, s_{tf}, sr)	94.6	84.5	86.0	**89.6**	88.7	88.1	87.9
tsn	deu, swa	swa, nya	(eo, d_{inv}, s_{tf}); $(d_{inv}, d_{geo}, d_{gen})$	88.7	73.1	73.4	**74.0**	73.3	73.1	65.8
twi	swa, nya	swa, ewe	(eo, s_{tf}, d_{geo}); (d_{geo}, s_{tf}, sr)	82.0	61.9	57.2	**64.3**	61.0	61.9	49.5
wol	fon, mos	fon, yor	(d_{geo}, sr, s_{tf}); (sr, d_{geo}, d_{syn})	85.2	62.0	59.4	**63.0**	62.0	58.9	44.8
xho	zul, sna	zul, pcm	(eo, d_{geo}, d_{gen}); (eo, s_{tf}, d_{inv})	90.8	83.7	83.0	**84.3**	83.7	74.0	24.5
yor	hau, pcm	fon, pcm	$(d_{geo}, d_{inv}, d_{syn})$; (eo, d_{geo}, d_{inv})	88.3	37.3	43.2	**50.3**	50.3	48.8	40.4
zul	xho, sna	xho, sna	(eo, d_{gen}, d_{geo}); (d_{syn}, sr, d_{geo})	88.6	82.1	85.5	**85.5**	82.1	69.4	44.7
AVG	–			87.3	64.2	69.8	**73.1**	71.3	68.8	56.9

Table 6: **Best Transfer Languages for NER.** The best zero-shot result is **bolded**, numbers that are not significantly different are underlined. The ranking model features are based on the definitions in (Lin et al., 2019) like: geographic distance (d_{geo}), genetic distance (d_{gen}), inventory distance (d_{inv}), syntactic distance (d_{syn}), phonological distance (d_{pho}), transfer language dataset size (s_{tf}), transfer over target size ratio (sr), and entity overlap (eo). The languages highlighted in gray have very good transfer performance ($> 70\%$) using the best transfer language.

Choice of Transfer Languages We selected 22 human-annotated NER datasets of diverse languages by searching the web and HuggingFace Dataset Hub (Lhoest et al., 2021). We required each dataset to contain at least the PER, ORG, and LOC types, and we limit our analysis to these types. We also added our MasakhaNER 2.0 dataset with 20 languages. In total, the datasets cover 42 languages (21 African). Each language is associated with a single dataset. Appendix C provides details about the languages, datasets, and data splits. To compute zero-shot transfer scores, we fine-tune mDeBERTaV3 on the NER dataset of a source language and perform zero-shot transfer to the target languages. We choose mDeBERTaV3 because it supports 100 languages and has the best performance among the PLMs trained on a similar number of languages.

6.2 Single-source Transfer Results

Figure 2 shows the zero-shot evaluation of training on 42 NER datasets and evaluation on the test sets of the 20 MasakhaNER 2.0 languages. On average, we find the transfer from non-African languages to be slightly worse (51.7 F1) than transfer from African languages (57.3 F1). The worst transfer result is using bbj as source language (41.0 F1) while the best is using sna (64 F1), followed by yor (63 F1).

We identify German (deu) and Finnish (fin) as the top-2 transfer languages among the non-African languages. In most cases, languages that are geographically and syntactically close tend to benefit most from each other. For example, sna, xho, and zul have very good transfer among themselves due to both syntactic and geographical closeness. Similarly, for Nigerian languages (hau, ibo, pcm, yor) and East African languages (kin, lug, luo, swa), geographical proximity plays an important role. While most African languages prefer transfer from another African language, there are few exceptions, like swa preferring transfer from deu or ara. The latter can be explained by the presence of Arabic loanwords in Swahili (Versteegh, 2001). Similarly, nya and tsn also prefer deu. Appendix G provides results for transfer to non-African languages.

6.3 LangRank and Co-training Results

We also investigate the benefit of training on the second-best language in addition to the languages selected by LangRank. We jointly train on the combined data of the top-2 transfer languages or the top-2 languages predicted by LangRank and evaluate their zero-shot performance on the target language. Table 6 shows the result for the top-2 transfer languages using the best from 42 × 42 transfer F1-scores and LangRank model predictions. LangRank predicted the right language as one of the top-2 best transfer language in 13 target languages. The target languages with incorrect predictions are fon, ibo, kin, lug, luo, nya, and swa. The transfer languages predicted as alternative are often in the top-5 transfer languages or are less than (-5 F1) worse than the best transfer language. For example, the best transfer language for lug is kin (81 F1) but LangRank predicted luo (76 F1). Appendix H gives results for non-African languages.

Features that are important for transfer The most important features for the selection of best language by LangRank are geographic distance (d_{geo}) and entity overlap (eo). The d_{geo} is influential because named entities (e.g. name of a politician or a city) are often similar from languages spoken in the same country (e.g. Nigeria with 4 languages in MasakhaNER 2.0) or region (e.g. East African languages). Similarly, we find entity overlap to have a positive Spearman correlation ($R = 0.6$) to transfer F1-score. Appendix F provides more details on the correlation results. d_{geo} occurred as part of the top-3 features for 15 best transfer language and 16 second-best languages. Similarly, for eo, it appeared 11–13 times for the top-2 transfer languages. Interestingly, dataset size was not among the most important features, highlighting the need for typologically diverse training data.

Best Transfer Language Outperforms English We compare the zero-shot transfer performance of the top-2 transfer languages to using eng as the transfer language. They significantly outperform the eng average of 56.9 by $+14$ and $+12$ F1 for the first and second-best source language, respectively.

Co-training of Top-2 Transfer Languages Improves Performance We find that co-training the top-2 transfer languages further improves zero-shot performance over the best transfer by around $+3$ F1. It is most significant for fon, ibo, kin and twi with 3–7 F1 improvement. Co-training the top-2 transfer languages predicted by LangRank is better than using the second-best transfer language, but often performs worse than the best transfer language.

6.4 Sample Efficiency Results

Figure 3 shows the performance when the model is trained on a few target language samples compared to when the best transfer language is used prior to fine-tuning on the same number of target language samples. We show the results for four languages (which reflect common patterns across all languages) and an average (ave) over the 20 languages. As seen in the figure, models achieve less than 50 F1 when we train on 100 sentences and over 75 F1 when training on 500 sentences. In practice, annotating 100 sentences takes about 30 minutes while annotating 500 sentences takes around 2 hours and 30 minutes; therefore, slightly more annotation effort can yield a substantial quality improvement. We also find that using the best transfer language in zero-shot settings gives a performance very close to annotating 500 samples in most cases, showing the importance of transfer language selection. By additionally fine-tuning the model on 100 or 500 target language samples, we can further improve the NER performance. Appendix I provides the sample efficiency results for individual languages.

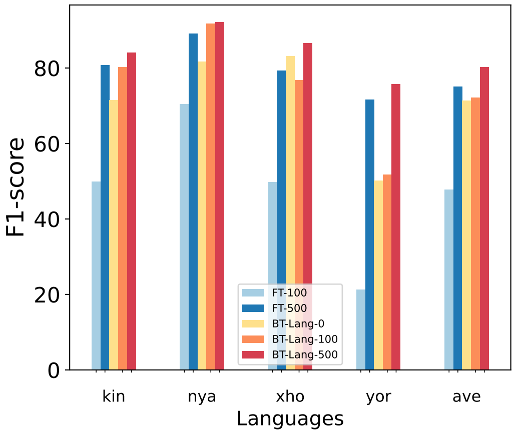


Figure 3: **Sample Efficiency Results** for 100 and 500 samples in the target language, model fine-tuned from a PLM (e.g. FT-100 – trained on 100 samples from the target language) or fine-tuned from the best transfer language NER model (e.g BT-Lang-0 – trained on 0 samples from the target language or zero-shot)

7 Conclusion

In this paper, we present the creation of MasakhaNER 2.0, the largest NER dataset for 20 diverse African languages and provide strong baseline results on the corpus by fine-tuning multilingual PLMs on in-language NER and multilingual datasets. Additionally, we analyze cross-lingual transfer in an Africa-centric setting, showing the importance of choosing the best transfer language in both zero-shot and few-shot scenarios. Using English as the default transfer language can have detrimental effects, and choosing a more appropriate language substantially improves fine-tuned NER models. By analyzing data-dependent, geographical, and typological features for transfer in NER, we conclude that geographical distance and entity overlap contribute most effectively to transfer performance.

Acknowledgements

This work was carried out with support from Lacuna Fund, an initiative co-founded by The Rockefeller Foundation, Google.org, and Canada's International Development Research Centre. David Adelani acknowledges the EU-funded Horizon 2020 projects: COMPRISE under grant agreement No. 3081705 and ROXANNE under grant number 833635. Vukosi Marivate acknowledges funding from ABSA covering the ABSA Data Science chair as well as funding from the Google Research Scholar program. We thank Mengzhou Xia, Antonis Anastasopoulos, and Fahim Faisal for their help with the LangRank code and data geography analysis. We thank Haneul Yoo for her comment on the initial Korean NER result. We thank Heng Ji and Ying Lin for providing the ELISA NER tool used for annotation. We thank Daniel D'souza for helping to set-up ELISA NER annotation tool. We thank Kelechi Ogueji for providing monolingual corpus for isiZulu. We thank Google for providing GCP credits to run some of the experiments. We thank the ML Group, Luleå University of Technology, for the compute resources for running some of the experiments. Finally, we thank the Masakhane leadership, Melissa Omino, Davor Orlič and Knowledge4All for their administrative support throughout the project.

Limitations

Some Language families not covered While we try to cover 20 topologically diverse languages and language families, a few locations in Africa and smaller language family groups were not covered. For example, languages from the Khoisan and Austronesian (like Malagasy) family were not covered. Also, languages spoken in the central Africa like South Sudan, Chad, and DRC were not covered.

News Domain Data As the data we annotated belonged to the news domain, models trained from this data may not generalize well to other domains. In particular, the models may not perform well on more casual text that may use different vocabulary, discuss different entities, and contain more orthographic variation. This limitation also applies for the English NER Corpus.

Generalizability of Transfer Learning Findings As we only experimented with one task (NER), our findings regarding effective approaches to transfer learning for African languages and PLMs may not generalize to other tasks (e.g. machine translation, part of speech tagging); other features of language similarity may be more important for other tasks.

Explaining Transfer Learning Findings We found that the `LangRank` model could not predict the top transfer languages with 100% accuracy. This suggests that there are other, unknown factors that could affect transfer performance, which we did not explore. For example, there is still work to be done to understand the sociolinguistic connections and language contact conditions that may correlate with effective transfer.

Ethics Statement

Our research process has been deeply rooted in the principles of participatory AI research (∀ et al., 2020), where the populations most affected by the research—the native speakers of the languages in this case—are involved throughout the project as stakeholders.

We believe our work will be of benefit to the speakers of the included languages by enabling better language technology for their languages. By keeping them engaged throughout the process and as collaborators in this work, we have been able to become aware of any potential harms. As the data we use for annotation is news data that was already publicly available, the release of our annotation is unlikely to cause unintended harm.

However, there are always potential unintended consequences when creating NER data and models. The data selection, annotation, adjudication, and model training process can all introduce biases that may have negative effects. Specifically, within each language, the models trained may perform better when processing names that commonly appear in newswire, and worse when processing names belonging to entities less well-represented in the news domain, propagating biases to downstream tasks.

References

Idris Abdulmumin, Satya Ranjan Dash, Musa Abdullahi Dawud, Shantipriya Parida, Shamsuddeen Muhammad, Ibrahim Sa'id Ahmad, Subhadarshi Panda, Ondřej Bojar, Bashir Shehu Galadanci, and Bello Shehu Bello. 2022. Hausa visual genome: A dataset for multi-modal English to Hausa machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6471–6479, Marseille, France. European Language Resources Association.

David Adelani, Dana Ruiter, Jesujoba Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Esther Awokoya, and Cristina España-Bonet. 2021a. The effect of domain and diacritics in Yoruba–English neural machine translation. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 61–75, Virtual. Association for Machine Translation in the Americas.

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 H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu 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. 2021b. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

David Ifeoluwa Adelani, Jesujoba Oluwadara 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 Chinenye 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 Ayoade Ajibade, Oluwaseyi Ajayi Ajayi, Yvonne Wambui Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Koffi 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. 2022. A few thousand translations go a long way! leveraging pre-trained models for african news translation. In *NAACL-HLT*.

Kabir Ahuja, Shanu Kumar, Sandipan Dandapat, and Monojit Choudhury. 2022. Multi task learning for zero shot performance prediction of multilingual models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5454–5467, Dublin, Ireland. Association for Computational Linguistics.

Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David Adelani, and Cristina España-Bonet. 2020. Massive vs. curated embeddings for low-resourced languages: the case of Yorùbá and Twi. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2754–2762, Marseille, France. European Language Resources Association.

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*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Cheikh Anta Babou and Michele Loporcaro. 2016. Noun classes and grammatical gender in wolof. *Journal of African Languages and Linguistics*, 37(1):1–57.

Yassine Benajiba, Paolo Rosso, and José Miguel BenedíRuiz. 2007. Anersys: An arabic named entity recognition system based on maximum entropy. In *Computational Linguistics and Intelligent Text Processing*, pages 143–153, Berlin, Heidelberg. Springer Berlin Heidelberg.

Michael Beukman. 2022. Analysing the effects of transfer learning on low-resourced named entity recognition performance. In *3rd Workshop on African Natural Language Processing*.

Adams Bodomo and Charles Marfo. 2002. The morphophonology of noun classes in dagaare and akan.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *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*, pages 8440–8451, Online. Association for Computational Linguistics.

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2022. Multilingual autoregressive entity linking. *Transactions of the Association for Computational Linguistics*, 10:274–290.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Dublin, Ireland. Association for Computational Linguistics.

A J De Waal, A L Louis, and J P Venter. 2006. Named entity recognition in a south african context.

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

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

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

Stefan Daniel Dumitrescu and Andrei-Marius Avram. 2020. Introducing RONEC - the Romanian named entity corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4436–4443, Marseille, France. European Language Resources Association.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, 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)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Roald Eiselen. 2016. Government domain named entity recognition for South African languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3344–3348, Portorož, Slovenia. European Language Resources Association (ELRA).

Fahim Faisal, Yinkai Wang, and Antonios Anastasopoulos. 2022. Dataset geography: Mapping language data to language users. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3381–3411, Dublin, Ireland. 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, 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. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online.

Cláudia Freitas, Cristina Mota, Diana Santos, Hugo Gonçalo Oliveira, and Paula Carvalho. 2010. Second HAREM: Advancing the state of the art of named entity recognition in Portuguese. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Normunds Gruzitis, Lauma Pretkalnina, Baiba Saulite, Laura Rituma, Gunta Nespore-Berzkalne, Arturs Znotins, and Peteris Paikens. 2018. Creation of a balanced state-of-the-art multilayer corpus for NLU. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *ArXiv*, abs/2111.09543.

Michael A. Hedderich, David Adelani, Dawei Zhu, Jesujoba Alabi, Udia Markus, and Dietrich Klakow. 2020. Transfer learning and distant supervision for multilingual transformer models: A study on African languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2580–2591, Online. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Rasmus Hvingelby, Amalie Brogaard Pauli, Maria Barrett, Christina Rosted, Lasse Malm Lidegaard, and Anders Søgaard. 2020. DaNE: A named entity resource for Danish. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4597–4604, Marseille, France. European Language Resources Association.

Ebrahim Chekol Jibril and A. Cüneyd Tantug. 2022. Anec: An amharic named entity corpus and transformer based recognizer. *ArXiv*, abs/2207.00785.

Bjarte Johansen. 2019. Named-entity recognition for norwegian. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics, NoDaLiDa*.

C. Junior, H. Macedo, T. Bispo, F. Oliveira, N. Silva, and L. Barbosa. 2015. Paramopama: a brazilian-portuguese corpus for named entity recognition. In *12th National Meeting on Artificial and Computational Intelligence (ENIAC)*.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Antonia Karamolegkou and Sara Stymne. 2021. Investigation of transfer languages for parsing Latin: Italic branch vs. Hellenic branch. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 315–320, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Siti Oryza Khairunnisa, Aizhan Imankulova, and Mamoru Komachi. 2020. Towards a standardized dataset on Indonesian named entity recognition. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 64–71, Suzhou, China. Association for Computational Linguistics.

Maria Yu Konoshenko and Dasha Shavarina. 2019. A microtypological survey of noun classes in kwa. *Journal of African Languages and Linguistics*, 40:114 – 75.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

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

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ying Lin, Cash Costello, Boliang Zhang, Di Lu, Heng Ji, James Mayfield, and Paul McNamee. 2018. Platforms for non-speakers annotating names in any language. In *Proceedings of ACL 2018, System Demonstrations*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

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 Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, 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.

Pengfei Liu, Jinlan Fu, Yang Xiao, Weizhe Yuan, Shuaichen Chang, Junqi Dai, Yixin Liu, Zihuiwen Ye, and Graham Neubig. 2021. ExplainaBoard: An explainable leaderboard for NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 280–289, Online. Association for Computational Linguistics.

Bernardo Magnini, Amedeo Cappelli, Fabio Tamburini, Cristina Bosco, Alessandro Mazzei, Vincenzo Lombardo, Francesca Bertagna, Nicoletta Calzolari, Antonio Toral, Valentina Bartalesi Lenzi, Rachele Sprugnoli, and Manuela Speranza. 2008. Evaluation of natural language tools for Italian: EVALITA 2007. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Stephen Mayhew, Tatiana Tsygankova, and Dan Roth. 2019. ner and pos when nothing is capitalized. 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 6256–6261, Hong Kong, China. Association for Computational Linguistics.

Hans J. Melzian. 1933. Introduction to the phonology of the bantu languages. *Bulletin of the School of Oriental and African Studies*, 7(1):246–247.

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

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Sa'id Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alípio Jorge, and Pavel Brazdil. 2022. NaijaSenti: A nigerian Twitter sentiment corpus for multilingual sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 590–602, Marseille, France. European Language Resources Association.

Clemens Neudecker. 2016. An open corpus for named entity recognition in historic newspapers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4348–4352, Portorož, Slovenia. European Language Resources Association (ELRA).

Johanna Nichols and Balthasar Bickel. 2013. Possessive classification. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Derek Nurse and Gerard Philippson, editors. 2006. *The Bantu Languages*. Routledge Language Family Series. Routledge, London, England.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

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*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Akintunde Oladipo, Odunayo Ogundepo, Kelechi Ogueji, and Jimmy Lin. 2022. An exploration of vocabulary size and transfer effects in multilingual language models for african languages. In *3rd Workshop on African Natural Language Processing*.

JC Oosthuysen. 2016. *The Grammar of isiXhosa*, 1 edition. African Sun Media.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jungwoo Ha, and Kyunghyun Cho. 2021. Klue: Korean language understanding evaluation.

Doris L. Payne, Sara Pacchiarotti, and Mokaya Bosire, editors. 2017. *Diversity in African languages*. Number 1 in Contemporary African Linguistics. Language Science Press, Berlin.

Jonas Pfeiffer, Ivan Vulić, 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)*, pages 7654–7673, Online. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Hanieh Poostchi, Ehsan Zare Borzeshi, Mohammad Abdous, and Massimo Piccardi. 2016. PersoNER: Persian named-entity recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3381–3389, Osaka, Japan. The COLING 2016 Organizing Committee.

A R Priatama, , and Y Setiawan. 2022. Regression models for estimating aboveground biomass and stand volume using landsat-based indices in post-mining area. *J. Manaj. Hutan Trop. (J. Trop. For. Manag.)*, 28(1):1–14.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics.

Machel Reid, Junjie Hu, Graham Neubig, and Yutaka Matsuo. 2021. AfroMT: Pretraining strategies and reproducible benchmarks for translation of 8 African languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1306–1320, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Teemu Ruokolainen, Pekka Kauppinen, Miikka Silfverberg, and Krister Lindén. 2019. A finnish news corpus for named entity recognition. *Language Resources and Evaluation*, pages 1–26.

O. M. Singh, A. Padia, and A. Joshi. 2019. Named entity recognition for nepali language. In *2019 IEEE 5th International Conference on Collaboration and Internet Computing (CIC)*, pages 184–190.

Stephanie Strassel and Jennifer Tracey. 2016. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280, Portorož, Slovenia. European Language Resources Association (ELRA).

György Szarvas, Richárd Farkas, László Felföldi, András Kocsor, and János Csirik. 2006. A highly accurate named entity corpus for Hungarian. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

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 at HLT-NAACL 2003*, pages 142–147.

Kees Versteegh. 2001. Linguistic contacts between arabic and other languages. *Arabica*, 48:470–508.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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 833–844, Hong Kong, China. Association for Computational Linguistics.

Mengzhou Xia, Antonios Anastasopoulos, Ruochen Xu, Yiming Yang, and Graham Neubig. 2020. Predicting performance for natural language processing tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8625–8646, Online. Association for Computational Linguistics.

Seid Muhie Yimam, Hizkiel Mitiku Alemayehu, Abinew Ayele, and Chris Biemann. 2020. Exploring Amharic sentiment analysis from social media texts: Building annotation tools and classification models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1048–1060, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hailemariam Mehari Yohannes and Toshiyuki Amagasa. 2022. Named-entity recognition for a low-resource language using pre-trained language model. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, SAC '22, page 837–844, New York, NY, USA. Association for Computing Machinery.

A Data Source and Splits

Table 7 shows the MasakhaNER 2.0 language, data source, train/dev/test split, and the number of tokens per entity type.

B Language Characteristics

Table 8 provides the details about the language characteristics.

B.1 Morphology and Noun classes

Many African languages are morphologically rich. According to the World Atlas of Language Structures (WALS; Nichols and Bickel, 2013), 16 of

			# Tokens				% Entities	
Language	**Data Source**	**Train / dev / test**	**PER**	**LOC**	**ORG**	**DATE**	**in Tokens**	**#Tokens**
Bambara (bam)	MAFAND-MT (Adelani et al., 2022)	4462/ 638/ 1274	4281	2557	429	2898	6.5	155,552
Ghomálá' (bbj)	MAFAND-MT (Adelani et al., 2022)	3384/ 483/ 966	2464	1371	1586	2457	11.3	69,474
Éwé (ewe)	MAFAND-MT (Adelani et al., 2022)	3505/ 501/ 1001	3931	5168	2064	2665	15.3	90420
Fon (fon)	MAFAND-MT (Adelani et al., 2022)	4343/ 621/ 1240	3572	2595	3082	5120	8.3	173,099
Hausa (hau)	Kano Focus and Freedom Radio	5716/ 816/ 1633	9853	6759	7089	7251	14.0	221,086
Igbo (ibo)	IgboRadio and Ka ỌdỊ Taa	7634/ 1090/ 2181	8532	7077	5418	4727	7.5	344,095
Kinyarwanda (kin)	IGIHE, Rwanda	7825/ 1118/ 2235	6889	8960	7012	8187	12.6	245,933
Luganda (lug)	MAFAND-MT (Adelani et al., 2022)	4942/ 706/ 1412	6058	3706	5441	3484	15.6	120,119
Luo (luo)	MAFAND-MT (Adelani et al., 2022)	5161/ 737/ 1474	6806	5605	7099	7359	11.7	229,927
Mossi (mos)	MAFAND-MT (Adelani et al., 2022)	4532/ 648/ 1294	2804	3044	3209	6334	9.2	168,141
Naija (pcm)	MAFAND-MT (Adelani et al., 2022)	5646/ 806/ 1613	4711	5077	5940	3654	9.4	206,404
Chichewa (nya)	Nation Online Malawi	6250/ 893/ 1785	9657	4600	5924	4308	9.3	263,622
Shona (sna)	VOA Shona	6207/ 887/ 1773	10667	5289	12418	3423	16.2	195,834
Swahili (swa)	VOA Swahili	6593/ 942/ 1883	9510	10515	6515	5331	12.7	251,678
Setswana (tsn)	MAFAND-MT (Adelani et al., 2022)	3489/ 499/ 996	3991	2285	2905	3190	8.8	141,069
Akan/Twi (twi)	MAFAND-MT (Adelani et al., 2022)	4240/ 605/ 1211	3588	2474	2375	1433	6.3	155,985
Wolof (wol)	MAFAND-MT (Adelani et al., 2022)	4593/ 656/ 1312	3588	2474	2375	1433	7.4	181,048
isiXhosa (xho)	Isolezwe Newspaper	5718/ 817/ 1633	8098	3087	5633	2433	15.1	127,222
Yorùbá (yor)	Voice of Nigeria and Asejere	6877/ 983/ 1964	8537	5819	6998	6372	11.4	244,144
isiZulu (zul)	Isolezwe Newspaper	5848/ 836/ 1670	5050	1900	5229	2012	11.0	128,658

Table 7: **Languages and Data Splits for MasakhaNER 2.0 Corpus**. Distribution of the number of entities

Language	**No. of Letters**	**Latin Letters Omitted**	**Letters added**	**Tonality**	**diacritics**	**Word Order**	**Morphological typology**	**Inflectional Morphology (WALS)**	**Noun Classes**
Bambara (bam)	27	q,v,x	ɛ, ɔ, ɲ, ŋ	yes, 2 tones	yes	SVO & SOV	isolating	strong suffixing	absent
Ghomálá' (bbj)	40	q, w, x, y	bv, dz, ə, aə, ɛ, gh, ny, nt, ŋ, ŋk, ɔ, pf, mpf, sh, ts, ʉ, zh, '	yes, 5 tones	yes	SVO	agglutinative	strong prefixing	active, 6
Éwé (ewe)	35	c, j, q	ɖ, dz, ɛ, ƒ, gb, ɣ, kp, ny, ŋ, ɔ, ts, ʋ	yes, 3 tones	yes	SVO	isolating	equal prefixing and suffixing	vestigial
Fon (fon)	33	q	ɖ, ɛ,gb, hw, kp, ny, ɔ, xw	yes, 3 tones	yes	SVO	isolating	little affixation	vestigial
Hausa (hau)	44	p,q,v,x	ɓ, ɗ, ƙ, ƴ, kw, ƙw, gw, ky, ƙy, gy, sh, ts	yes, 2 tones	no	SVO	agglutinative	little affixation	absent
Igbo (ibo)	34	c, q, x	ch, gb, gh, gw, kp, kw, nw, ny, ọ, ȯ, sh, ụ	yes, 2 tones	yes	SVO	agglutinative	little affixation	vestigial
Kinyarwanda (kin)	30	q, x	cy, jy, nk, nt, ny, sh	yes, 2 tones	no	SVO	agglutinative	strong prefixing	active, 16
Luganda (lug)	25	h, q, x	ŋ, ny	yes, 3 tones	no	SVO	agglutinative	strong prefixing	active, 20
Luo (luo)	31	c, q, x, v, z	ch, dh, mb, nd, ng', ng, ny, nj, th, sh	yes, 4 tones	no	SVO	agglutinative	equal prefixing and suffixing	absent
Mossi (mos)	26	c, j, q, x	', ɛ, ɩ, ʋ	yes, 2 tones	yes	SVO	isolating	strongly suffixing	active, 11
Chichewa (nya)	31	q, x, y	ch, kh, ng, ŋ, ph, tch, th, ŵ	yes, 2 tones	no	SVO	agglutinative	strong prefixing	active, 17
Naija (pcm)	26	–	–	no	no	SVO	mostly analytic	strongly suffixing	absent
Shona (sna)	29	c, l, q, x	bh, ch, dh, nh, sh, vh, zh	yes, 2 tones	no	SVO	agglutinative	strong prefixing	active, 20
Swahili (swa)	33	x, q	ch, dh, gh, kh, ng', ny, sh, th, ts	no	yes	SVO	agglutinative	strong suffixing	active, 18
Setswana (tsn)	36	c, q, v, x, z	ê, kg, kh, ng, ny, ô, ph, š, th, tl, tlh, ts, tsh, tš, tšh	yes, 2 tones	no	SVO	agglutinative	strong prefixing	active, 18
Akan/Twi (twi)	22	c,j,q,v,x,z	ɛ, ɔ	yes, 5 tones	no	SVO	isolating	strong prefixing	active, 6
Wolof (wol)	29	h,v,z	ŋ, à, é, ë, ó, ñ	no	yes	SVO	agglutinative	strong suffixing	active, 10
isiXhosa (xho)	68	–	bh, ch, dl, dy, dz, gc, gq, gr, gx, hh, hl, kh, kr, lh, mh, ng, ngc, ngh, ngq, ngx, nkq, nkx, nh, nkc, nx, ny, nyh, ph, qh, rh, sh, th, ths, thsh, ts, tsh, ty, tyh, wh, xh, yh, zh	yes, 2 tones	no	SVO	agglutinative	strong prefixing	active, 17
Yorùbá (yor)	25	c, q, v, x, z	ẹ, gb, ṣ , ọ	yes, 3 tones	yes	SVO	isolating	little affixation	vestigial, 2
isiZulu (zul)	55	–	nx, ts, nq, ph, hh, ny, gq, hl, bh, nj, ch, ngc, ngq, th, ngx, kl, ntsh, sh, kh, tsh, ng, nk, gx, xh, gc, mb, dl, nc, qh	yes, 3 tones	no	SVO	agglutinative	strong prefixing	active, 17

Table 8: Linguistic Characteristics of the Languages

our languages employ strong prefixing or suffixing inflections. Niger-Congo languages are known for their system of noun classification. 12 of the languages *actively* make use of between 6–20 noun classes, including all Bantu languages and Ghomálá', Mossi, Akan and Wolof (Nurse and Philippson, 2006; Payne et al., 2017; Bodomo and Marfo, 2002; Babou and Loporcaro, 2016). While noun classes are often marked using affixes on the head word in Bantu languages, some non-Bantu languages, e.g., Wolof make use of a dependent such as a determiner that is not attached to the head word. For the other Niger-Congo languages such as Fon, Ewe, Igbo and Yorùbá, the use of noun classes is merely *vestigial* (Konoshenko and Shavarina, 2019). For example, Yorùbá only distinguishes between human and non-human nouns. Bambara is the only Niger-Congo language without noun classes, and some have argued that the Mande family should be regarded as an independent language family. Three of our languages from the Southern Bantu family (chiShona, isiXhosa and isiZulu) capitalize proper names after the noun class prefix as in the language names themselves. This characteristic limits the transfer learning of NER from languages without this feature, since NER models overfit on capitalization (Mayhew et al., 2019).

B.2 IsiXhosa and isiZulu morphological structure

IsiXhosa and isiZulu are agglutinative languages with a complex morphology. Each root or stem can attach a variety of affixes to form new inflections and derivations, with a variety of affixes added to root and stem morphemes to vary their meaning and convey syntactic agreement. The noun class

Language	Data Source	# Train	# dev	# test
Amharic (amh)	MasakhaNER 1.0 (Adelani et al., 2021b)	1,750	250	500
Arabic (ara)	ANERcorp (Benajiba et al., 2007; Obeid et al., 2020)	3,472	500	924
Danish (dan)	DANE (Hvingelby et al., 2020)	4,383	564	565
German (deu)	CoNLL03 (Tjong Kim Sang and De Meulder, 2003)	12,152	2,867	3,005
English (eng)	CoNLL03 (Tjong Kim Sang and De Meulder, 2003)	14,041	3,250	3,453
Spanish (spa)	CoNLL02 (Tjong Kim Sang, 2002)	8,322	1,914	1,516
Farsi (fas)	PersoNER (Poostchi et al., 2016)	4,121	1,000	2,560
Finnish (fin)	FINER (Ruokolainen et al., 2019)	13,497	986	3,512
French (fra)	Europeana (Neudecker, 2016)	9,546	2,045	2,047
Hungarian (hun)	Hungarian MTI (Szarvas et al., 2006)	4,532	648	1,294
Indonesia (ind)	(Khairunnisa et al., 2020)	6,707	1,437	1,438
Italian (ita)	I-CAB EVALITA 2007 & 2009 (Magnini et al., 2008)	11,227	4,136	2,068
Korean (kor)	KLUE (Park et al., 2021)	20,008	1,000	5,000
Latvian (lav)	(Gruzitis et al., 2018)	7,997	1,713	1,715
Nepali (nep)	(Singh et al., 2019)	2,301	328	659
Dutch (nld)	CoNLL02 (Tjong Kim Sang, 2002)	15,806	2,895	5,195
Norwegian (nor)	(Johansen, 2019)	15,696	2,410	1,939
Portuguese (por)	Second HAREM (Freitas et al., 2010) & Paramopama (Junior et al., 2015)	11,258	2,412	2,414
Romanian (ron)	RONEC (Dumitrescu and Avram, 2020)	5,886	1,000	2,453
Swedish (swe)	"swedish_ner_corpus" on HuggingFace Datasets (Lhoest et al., 2021)	9,000	1,330	2,000
Ukrainian (ukr)	"benjamin/ner-uk" on HuggingFace Datasets (Lhoest et al., 2021)	10,833	1,307	668
Chinese (zho)	"msra_ner" on HuggingFace Datasets (Lhoest et al., 2021)	45,057	3,442	1,721

Table 9: **Languages and Data Splits for Other NER Datasets**.

system and the concord agreement system are the foundations of isiXhosa and isiZulu noun grammar. This section offers an overview of these two principles and their applicability to the realization of NEs. First, we briefly describe the noun class system, after which we discuss prefixing and capitalization work for both languages.

According to the Meinhoff system (Melzian, 1933), nouns in African languages are classified into one of 18 numbered classes based on their prefix. As shown in the following example, singular nouns in class 1 take the prefix um-, while associated plural nouns in class 2 take the prefix aba-.

B.2.1 Prefix

Even though all named entities are nouns since they designate a distinct entity, noun class designations are critical in identifying NEs. According to Oosthuysen (2016), the purpose of the noun class prefix is to distinguish the class to which it belongs. It shows whether the noun is singular or plural. The derivation of all significant prefixes and cordial agreements is based on this.

In isiXhosa, named entities referring to personal nouns with the prefix um- belongs to noun class 1 with noun class 2 being its plural form. Named entities such as jobs, objects and concepts belong to noun class 3, e.g. umpheki (cook) and umthwalo (burden). Lastly in isiXhosa, borrowed words from English and Afrikaans such as ibhanka (bank) and ihamire (hammer), belong to class 9. In isiZulu, noun class 1 is a singular class which uses the prefix umu-/um-. The allomorph umu- occurs when the noun stem consists of one syllable, e.g. umuntu (person) and the allomorph um- occurs when the noun stem has more than one syllable, e.g. umfana (boy). The noun class 2 is a plural class, with its singular in class 1. Noun class 2 uses the prefix aba-/ab-, e.g. abantu (people), abafana (boys). Noun classes 1 and 2 are a personal class only containing personal nouns.

Noun class 1a is a subclass of noun class 1. This class contains personal nouns referring to family relationships, professions, proper names and personalized nouns. This class uses the prefix u- with no allomorphs, e.g. ugogo (grandmother), unesi (nurse) or uSipho (personal name). The noun class 2a is a regular plural of class 1a which uses the prefix o-, e.g. ogogo (grandmothers), onesi (nurses) or oSipho (Sipho and company).

B.2.2 Capitalization

Capitalization is a very common feature for a number of natural language processing tools, such as named entity recognition systems that identify people's names, and locations (De Waal et al., 2006). Following are the four different types of the usage of capitalization in isiXhosa and isiZulu (Priatama et al., 2022):

1. Initial capitalization of words in which only the initial letter is capitalized;
2. Mixed capitalization of words in which the

Language	XLM-R-large					mDeBERTaV3-base					AfroXLMR-large				
	all	0-freq	Δ 0-freq	long	Δ long	all	0-freq	Δ 0-freq	long	Δ long	all	0-freq	Δ 0-freq	long	Δ long
bam	79.4	62.3	-17.1	74.7	-4.7	81.3	66.3	-15.0	78.6	-2.7	82.1	67.2	-14.9	81.1	-1.0
bbj	74.8	66.1	-8.7	87.4	12.6	75.0	65.8	-9.2	63.9	-11.1	76.5	65.8	-10.7	80.0	3.5
ewe	89.5	75.6	-13.9	70.6	-18.9	90.0	76.9	-13.1	70	-20.0	91.0	79.7	-11.3	74.2	-16.8
fon	81.5	71.2	-10.3	69.6	-11.9	83.3	74.5	-8.8	68.1	-15.2	82.8	73.6	-9.2	68.7	-14.1
hau	87.4	83.8	-3.6	77.6	-9.8	84.8	80.0	-4.8	72.2	-12.6	87.8	84.6	-3.2	78.1	-9.7
ibo	87.0	77.4	-9.6	75.6	-11.4	89.7	82.6	-7.1	71.8	-17.9	89.1	80.9	-8.2	64.0	-25.1
kin	84.1	74.9	-9.2	75.3	-8.8	86.2	79.0	-7.2	75.3	-10.9	87.8	81.7	-6.1	77.1	-10.7
lug	87.3	75.3	-12.0	74.1	-13.2	88.7	77.4	-11.3	78.6	-10.1	89.4	79.7	-9.7	74.7	-14.7
mos	77.1	69.5	-7.6	55.8	-21.3	78.0	71.2	-6.8	58.9	-19.1	77.5	70.2	-7.3	60.1	-17.4
nya	89.7	82.0	-7.7	81.6	-8.1	91.9	86.5	-5.4	86.7	-5.2	92.2	87.3	-4.9	87.1	-5.1
pcm	89.8	84.5	-5.3	76.8	-13.0	90.2	84.9	-5.3	79.7	-10.5	90.4	86.1	-4.3	79.1	-11.3
sna	94.9	89.9	-5.0	93.3	-1.6	95.3	91.4	-3.9	92.4	-2.9	96.3	93.9	-2.4	93.9	-2.4
swa	92.8	84.1	-8.7	73.0	-19.8	92.4	82.8	-9.6	65.1	-27.3	92.3	83.0	-9.3	65.9	-26.4
tsn	86.4	74.9	-11.5	34.5	-51.9	87.0	75.8	-11.2	45.7	-41.3	89.8	80.9	-8.9	42.9	-46.9
twi	77.9	65.5	-12.4	52.2	-25.7	80.4	70.9	-9.5	62.3	-18.1	81.4	72.3	-9.1	63.2	-18.2
wol	83.3	65.9	-17.4	59.1	-24.2	83.3	67.2	-16.1	58.6	-24.7	86.2	72.0	-14.2	62.2	-24.0
xho	88.0	83.2	-4.8	76.7	-11.3	88.0	83.8	-4.2	76.2	-11.8	90.1	86.5	-3.6	78.5	-11.6
yor	86.4	78.2	-8.2	67.0	-19.4	86.8	79.2	-7.6	74.4	-12.4	90.2	85.0	-5.2	74.0	-16.2
zul	86.4	83.2	-3.2	69.5	-16.9	89.4	86.1	-3.3	68.8	-20.6	90.1	87.5	-2.6	67.1	-23.0
avg	85.5	76.2	-9.3	70.8	-14.7	86.4	78.0	-8.4	70.9	-15.5	87.5	79.9	-7.6	72.2	-15.3

Table 10: F1 score for two varieties of hard-to-identify entities: zero-frequency entities that do not appear in the training corpus, and longer entities of four or more words.

Language	XLM-R-large				mDeBERTaV3-base				AfroXLMR-large			
	DATE	LOC	ORG	PER	DATE	LOC	ORG	PER	DATE	LOC	ORG	PER
bam	90.3	83.2	80.7	87.1	90.1	86.4	79.2	88.4	92.6	87.7	82.4	86.1
bbj	87.6	82.9	79.4	83.6	79.9	86.4	72.5	87.2	85.7	87.0	75.2	84.7
ewe	91.8	96.8	85.5	95.9	91.8	96.4	88.6	97.1	92.0	97.8	85.6	98.6
fon	85.4	89.2	86.9	94.6	86.8	93.3	89.3	94.3	85.9	91.9	86.4	94.6
hau	86.8	90.0	92.5	98.0	86.4	89.2	89.1	98.0	87.4	91	92.2	98.2
ibo	84.5	91.6	83.5	97.7	85.4	95.6	82.5	99.1	87.2	96.5	73.4	98.8
kin	88.4	92.7	84.0	94.8	87.4	95.0	87.8	97.7	88.1	95.6	89.1	99.1
lug	78.2	93.1	94.2	95.8	80.2	95.1	94.3	96.0	81.7	93.1	95.1	97.3
mos	80.3	92.7	74.4	93.1	81.6	92.1	78.9	88.3	83.2	93.7	75.4	88.9
pcm	96.6	91.1	89.7	96.9	96.1	93.1	90.9	97.3	95.6	92.4	90.9	97.1
nya	89.1	94.1	94.2	94.4	89.6	96.7	96.0	94.9	89.1	96.2	94.8	95.6
sna	95.6	95.6	96.1	98.1	96.0	95.1	96.5	98.7	96.6	95.4	97.4	99.3
swa	92.2	97.0	95.2	98.8	91.5	96.9	94.6	98.8	91.5	97.4	93.7	98.2
tsn	88.1	88.3	89.1	97.1	87.8	90.0	89.0	97.6	90.5	94.8	92.2	98.6
twi	66.7	89.3	79.4	96.1	76.5	90.4	82.9	97.5	75.7	91.4	85.1	97.7
wol	80.6	84.9	87.0	95.9	80.8	88.2	88.4	95.0	82.6	91.9	88.0	97.0
xho	90.7	91.6	93.1	96.9	89.7	92.0	93.4	98.1	91.1	93.5	95.0	98.3
yor	89.6	94.0	90.3	93.6	89.6	92.1	91.4	94.6	91.3	95.8	92.5	96.4
zul	85.0	90.1	87.8	97.1	92.2	95.5	88.1	97.1	90.8	96.2	91.8	97.2
avg	86.7	91.0	87.5	95.0	87.3	92.6	88.1	95.6	88.4	93.7	88.2	95.9

Table 11: F1 score for the different entity types.

initial letter of the prefix is capitalized as well as the initial letter of the main word;

3. Internal capitalization in words which are found in the middle of a sentence where the prefix remains lower case and the first letter of the main word is capitalized.

4. All CAPS in words that are fully capitalized. These are usually abbreviations or acronyms;

C Other NER Corpus

Table 9 provides the NER corpus found online that we make use for determining the best transfer languages

D Error Analysis of NER

Table 10 and Table 11 provides error analysis of MasakhaNER 2.0 based on performance on zero-frequency entities, long entities and distribution by named entity tags.

E LangRank Feature Descriptions

The following definitions are listed here, originally from Lin et al. (2019).

Geographic distance (d_{geo}) based on the orthodromic distance between language locations obtained from Glottolog (Hammarström et al., 2018).

Genetic distance (d_{gen}) based on the genealogical distance of Glottolog language tree.

Inventory distance (d_{inv}) based on the cosine distance between phonological feature vectors obtained from PHOIBLE database (Moran et al., 2014).

Syntactic distance (d_{syn}) based on cosine distance between feature vectors obtained from syntactic structures derived from WALS database (Dryer and Haspelmath, 2013).

Phonological distance (d_{pho}) based on the cosine distance between phonological feature vectors obtained from WALS and Ethnologue databases (Lewis, 2009).

Featural distance (d_{fea}) based on the cosine distance between feature vectors combining all 5 features mentioned above.

Transfer language dataset size (s_{tf}) The size of the transfer language's dataset.

Target language dataset size (s_{tg}) The size of the target language's dataset.

Transfer over target size ratio (sr) The size of the transfer language's dataset divided by the size of the target language's dataset.

Entity Overlap (eo) The number of unique words that overlap between the source and target languages' training datasets.

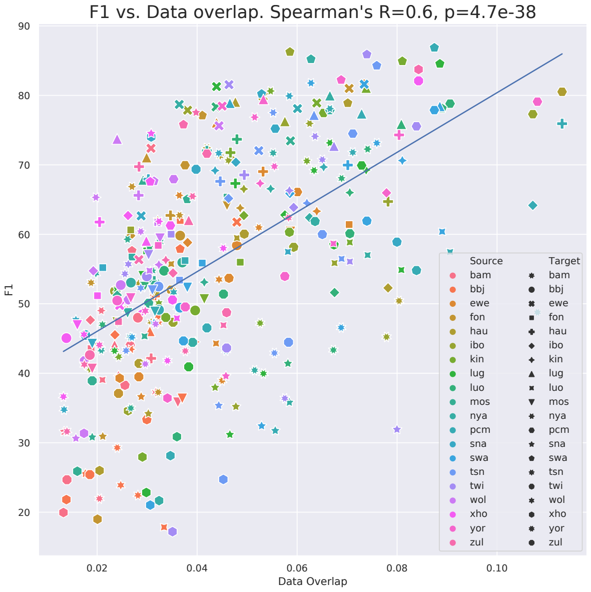


Figure 4: The correlation between the data overlap and F1 transfer performance. For source language X and target language Y, denote the set of unique named entities (PER, ORG, LOC, DATE) by T_X and T_Y respectively. The overlap here was calculated as $\frac{|T_X \cap T_Y|}{|T_X|+|T_Y|}$, as in Lin et al. (2019).

F Overlap Results

In Figure 4, we examine the word overlap between different languages, and how this correlates with the transfer performance. In general, these two quantities are strongly correlated (Spearman's $R = 0.6, p < 0.05$), echoing a similar result described by Beukman (2022). Note that the entity overlap feature used by the ranking model in the main text was calculated in a slightly different way; namely, considering *all* tokens instead of just the 4 named entities and not normalizing the overlap. This case still shows a positive correlation, although it is slightly smaller with Spearman's $R = 0.49$.

G Zero-shot Transfer

Figure 5 shows $N \times N$ transfer results to languages in MasakhaNER 2.0. We see that English is not the best transfer language in general. It is better to choose a more geographically close African language.

Figure 6 shows $N \times N$ transfer results to languages not in MasakhaNER 2.0. We see that English appears to be the best transfer on average, which is not the case for African languages. The reason for this is that many of the non-African languages we evaluated on are from the Indo-European, similar to English.

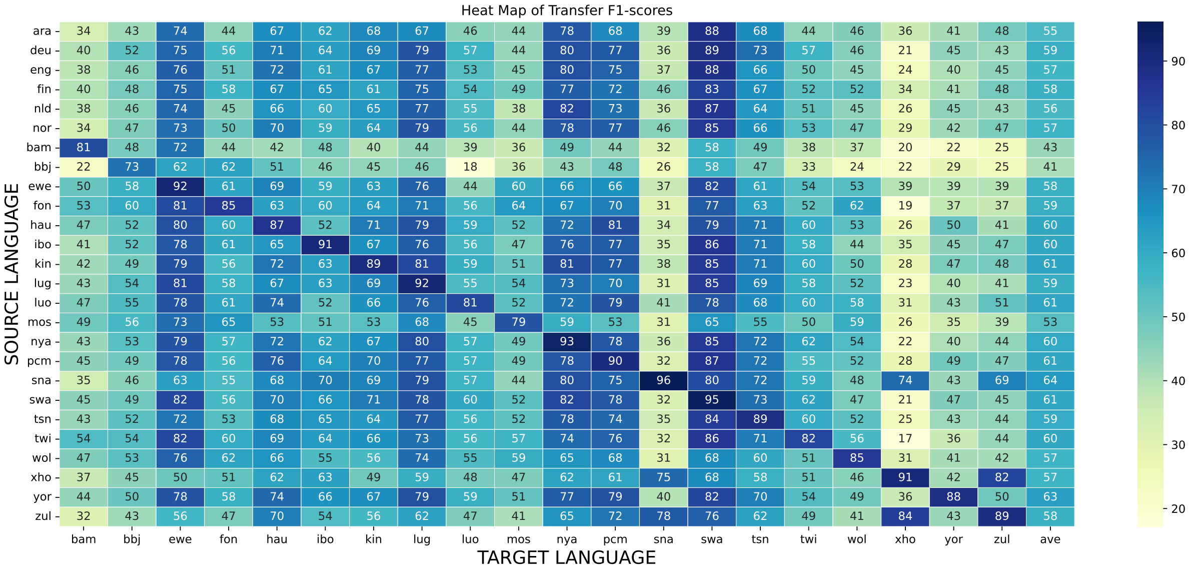


Figure 5: **Zero-shot Transfer** from several source languages to African languages in MasakhaNER 2.0.

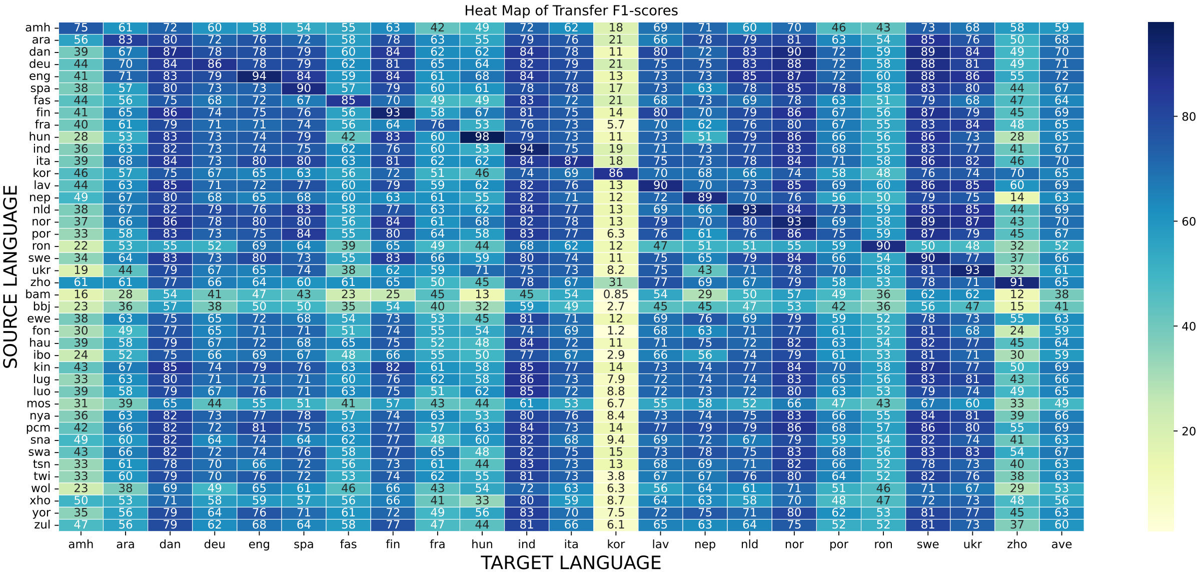


Figure 6: **Zero-shot Transfer** from several source languages to other languages not in MasakhaNER 2.0

H Best Transfer Language for Other Languages

Table 12 provides the result of the best transfer language for other languages not in MasakhaNER 2.0.

I Sample Efficiency Results

Figure 7 shows the result of training NER models using 100 and 500 samples for each language.

J Model Hyper-parameters for Reproducibility

For training NER models, we *fine-tune* PLM, we make use of a maximum sequence length of 200, batch size of 16, gradient accumulation of 2, learning rate of 5e-5, and number of epochs 50. The experiments of the large PLMs were performed on using Nvidia V100 GPU. For AfriBERTa and mBERT, we make use of Nvidia GeForce RTX-2080Ti. For evaluation, we make use of the micro-averaged F1 score.

Target Lang.	Top-2 Transf. Lang	Top-2 LangRank Model	Top-3 features selected by the LangRank Model Lang 1; Lang 2	Target Lang. F1	Best Transf. F1	Second Best Transf. F1	eng Tranf. F1	LangRank First Lang F1	LangRank Second Lang F1
African languages									
amh	zho, ara	pcm, luo	(s_{tf}, s_{tg}, sr); (s_{tf}, d_{geo}, sr)	75.0	**61.0**	55.9	40.6	42.5	38.6
bam	twi, fon	wol, fon	(d_{geo}, d_{inv}, sr); (d_{geo}, sr, d_{pho})	80.4	**54.3**	53.0	38.4	47.1	53.0
bbj	fon, ewe	twi, ewe	$(s_{tf}, d_{syn}, d_{geo})$; (s_{tf}, d_{geo}, sr)	72.9	**59.8**	58.4	45.8	53.9	58.4
ewe	swa, twi	pcm, swa	(d_{geo}, s_{tf}, sr); (eo, d_{geo}, s_{tf})	91.7	**81.6**	81.5	76.4	78.1	**81.6**
fon	mos, bbj	yor, ewe	(d_{geo}, d_{syn}, sr); $(s_{tf}, d_{geo}, d_{gen})$	84.9	**65.4**	62.0	50.6	58.4	61.4
hau	pcm, yor	yor, swa	(d_{geo}, sr, eo); (eo, sr, s_{tf})	86.9	75.9	**74.3**	72.4	74.3	70.0
ibo	sna, yor	pcm, kin	(eo, d_{geo}, s_{tf}); (d_{geo}, sr, eo)	91.0	**70.4**	66.0	61.4	64.2	62.7
kin	hau, swa	sna, yor	(eo, d_{geo}, s_{tf}); (eo, s_{tf}, sr)	89.5	**71.1**	70.6	67.4	69.2	67.3
lug	kin, nya	luo, zul	(d_{geo}, sr, eo); (d_{syn}, d_{geo}, sr)	91.5	**81.1**	80.0	76.5	75.9	62.0
luo	swa, hau	lug, sna	(d_{geo}, sr, eo); (d_{geo}, eo, sr)	81.2	**60.4**	59.5	53.4	54.9	57.5
mos	fon, ewe	yor, fon	(d_{geo}, d_{inv}, sr); (d_{geo}, s_{tf}, sr)	78.9	**64.2**	60.4	45.4	50.8	**64.2**
nya	swa, nld	zul, sna	(eo, d_{geo}, sr); (d_{geo}, eo, d_{syn})	93.5	**81.8**	81.7	80.1	65.5	79.9
pcm	hau, yor	eng, yor	(eo, d_{gen}, d_{syn}); (eo, d_{geo}, sr)	89.9	**80.5**	79.1	75.5	75.5	79.1
sna	zul, xho	swa, zul	(eo, sr, s_{tf}); (d_{geo}, sr, eo)	96.0	**77.5**	74.5	37.1	32.4	77.5
swa	deu, ara	ita, nld	(sr, d_{inv}, eo); (eo, s_{tf}, sr)	94.6	**88.7**	88.1	87.9	84.5	86.6
tsn	deu, swa	swa, nya	(eo, d_{inv}, s_{tf}); $(d_{inv}, d_{geo}, d_{gen})$	88.7	**73.3**	73.1	65.8	73.1	71.7
twi	swa, nya	swa, ewe	(eo, s_{tf}, d_{geo}); (d_{geo}, s_{tf}, sr)	82.0	61.0	**61.9**	49.5	**61.9**	53.7
wol	fon, mos	fon, yor	(d_{geo}, sr, s_{tf}); (sr, d_{geo}, d_{syn})	85.2	**62.0**	58.9	44.8	**62.0**	49.0
xho	zul, sna	zul, pcm	(eo, d_{geo}, d_{gen}); (eo, s_{tf}, d_{inv})	90.8	**83.7**	74.0	24.5	83.7	28.1
yor	hau, pcm	fon, pcm	$(d_{geo}, d_{inv}, d_{syn})$; (eo, d_{geo}, d_{inv})	88.3	**50.3**	48.8	40.1	37.3	48.8
zul	xho, sna	xho, sna	(eo, d_{gen}, d_{geo}); (d_{syn}, sr, d_{geo})	88.6	**82.1**	69.4	44.7	**82.1**	69.4
Non-African languages									
ara	eng, deu	fas, pcm	(eo, d_{inv}, d_{syn}); (d_{syn}, sr, d_{inv})	82.8	**71.5**	69.9	**71.5**	55.7	57.9
dan	nor, fin	swe, nor	(eo, d_{gen}, d_{geo}); (eo, d_{geo}, d_{syn})	87.1	**86.3**	85.6	83.1	82.8	86.3
deu	nld, eng	dan, nld	$(d_{geo}, eo, s_{tf}, d_{syn})$; (eo, d_{syn}, d_{geo})	86.5	**79.3**	78.8	78.8	79.3	79.3
eng	pcm, swe	nld, pcm	(eo, d_{geo}, d_{syn}); $(eo, d_{gen} d_{pho})$	93.5	81.3	79.7	**93.5**	76.0	81.3
fas	hau, pcm	ara, eng	(d_{syn}, d_{inv}, eo); $(d_{syn}, d_{geo}, s_{tf})$	84.8	**64.8**	63.4	59.3	57.9	59.2
fin	dan, eng	deu, eng	(eo, s_{tf}, d_{geo}); (d_{syn}, d_{geo}, eo)	93.4	**83.7**	83.6	83.6	80.8	83.6
fra	swe, swa	nld, deu	(eo, d_{syn}, d_{geo}); (d_{geo}, eo, sr)	75.5	**66.3**	65.4	60.6	63.3	64.9
hun	ukr, eng	deu, ron	(d_{geo}, d_{syn}, eo); (d_{geo}, eo, d_{syn})	98.0	**70.7**	68.4	68.4	63.6	43.8
ind	lug, luo	zho, nld	(s_{tg}, s_{tf}, sr); (d_{syn}, s_{tf}, eo)	93.7	**85.9**	85.2	83.9	78.6	84.1
ita	deu, spa	nld, eng	(d_{syn}, eo, d_{geo}); (eo, d_{syn}, d_{geo})	86.7	**79.1**	78.2	77.0	77.1	77.1
kor	zho, ind	ara, nep	(sr, s_{tf}, d_{syn}); $(d_{inv}, d_{syn}, s_{tf})$	85.7	**31.1**	21.5	12.7	21.3	11.9
lav	fin, dan	eng, nld	(s_{tf}, d_{syn}, sr); $(s_{tf}, d_{syn}, d_{geo})$	89.7	**80.4**	80.1	73.5	73.5	69.5
nep	pcm, swa	kor, zho	$(d_{syn}, s_{tf}, d_{pho})$; (s_{tf}, sr, d_{geo})	89.5	**79.0**	77.7	73.4	68.2	68.5
nld	eng, deu	eng, nor	(eo, d_{geo}, d_{syn}); (eo, d_{geo}, s_{tf})	93.4	**85.4**	83.7	85.4	**85.4**	79.9
nor	dan, deu	dan, eng	(eo, d_{geo}, s_{tf}); (eo, d_{geo}, sr)	92.5	**89.8**	87.8	87.3	89.8	87.2
por	es, nld	spa, eng	(eo, d_{syn}, d_{gen}); (eo, d_{syn}, d_{geo})	75.0	**77.8**	73.5	72.0	77.8	72.0
ron	lav, eng	eng, ita	(eo, d_{syn}, d_{geo}); (eo, d_{geo}, d_{syn})	89.6	**59.6**	59.5	59.5	59.5	57.8
spa	eng, por	por, lav	(eo, d_{geo}, d_{syn}); (d_{syn}, eo, d_{geo})	89.6	**83.9**	83.6	**83.9**	83.6	77.3
swe	dan, nor	nor, nld	(eo, d_{syn}, d_{geo}); (d_{syn}, d_{geo}, eo)	90.3	**89.4**	89.1	88.1	89.3	85.2
ukr	nor, eng	deu, eng	(d_{geo}, d_{syn}, sr); $(d_{syn}, d_{geo}, s_{tf})$	92.6	**87.2**	85.6	85.6	81.5	85.6
zho	lav, amh	pcm, deu	$(d_{syn}, s_{tf}, s_{geo})$; $(d_{syn}, s_{tf}, d_{pho})$	91.4	**60.2**	58.3	54.7	54.7	48.9
AVG	–			87.7	73.3	71.2	64.6	67.3	66.2

Table 12: **Best Transfer Language for NER.** The ranking model features are based on the definitions in (Lin et al., 2019) like: geographic distance (d_{geo}), genetic distance (d_{gen}), inventory distance (d_{inv}), syntactic distance (d_{syn}), phonological distance (d_{pho}), transfer language dataset size (s_{tf}), target language dataset size(s_{tg}), transfer over target size ratio (sr), and entity overlap (eo).

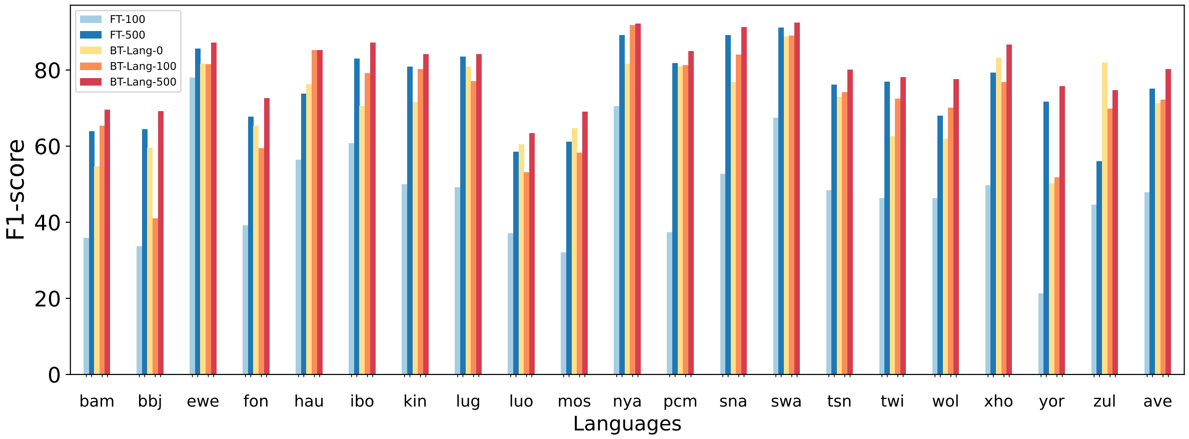


Figure 7: **Sample Efficiency Results** for 100 and 500 samples in the target language, model fine-tuned on a PLM (e.g. FT-100 – trained on 100 samples from the target language) or fine-tuned on the best transfer language NER model (e.g. BT-Lang-0 – trained on 0 samples from the target language or zero-shot)