Benchmarking Text Embedding Models for South African Languages

Ockert de Villiers, Roald Eiselen
Centre for Text Technology (CTexT)
North-West University, Potchefstroom, South Africa
{Almaro.DeVilliers|Roald.Eiselen}@nwu.ac.za

Abstract

In this work we introduce a collection of monolingual embedding models for ten South African languages in four different architectures. To determine the quality of the embedding models we evaluate the embeddings on two sequence-labelling tasks, namely Part-of-Speech (POS) tagging and Named Entity Recognition (NER). Languages are grouped into conjunctive (isiNdebele, isiXhosa, isiZulu, and Siswati), disjunctive (Sepedi, Sesotho, Setswana, Tshivenḓa, and Xitsonga), and Afrikaans to establish the influence of training data set size and typology on the quality of the different embeddings. To isolate representation effects we train BiLSTM-CRF taggers, while keeping the architecture, data splits, and training budget fixed, varying only the input imbedding representations, namely GloVe, fastText, Flair, and RoBERTa. In our experiments, GloVe lags behind fastText, Flair, and the transformer-based models, confirming that static word-level vectors are less suited to morphologically complex, low-resource languages. Subword-aware embeddings such as fastText remain a reliable and computationally efficient baseline, while Flair is the most competitive overall across both POS tagging and NER tasks.

Keywords: South African languages, embeddings, POS tagging, named entity recognition, low-resource NLP

1. Introduction

Vectorised representations of words in the form of embeddings signalled the beginning of a major change in Natural Language Processing (NLP). The models based on these embeddings have led to many state-of-the-art improvements in a variety of computational linguistic methods and applications. Embeddings also underscored the initial improvements that were made possible with deep learning architectures, eventually leading to the emergence of transformers and the subsequent 'AI revolution' that is currently underway. Although the initial embedding models, such as Word2Vec and GloVe, have been superseded by contextualised representations such as BERT, these contextual models generally require much more training data to build good representations than the original embeddings. In resource-constrained environments, more static embeddings may still have a role to play in NLP applications and computational linguistic research. In this paper, we introduce a set of embedding models for ten South African languages covering the major embedding architectures and ascertain their relative quality across two linguistic annotation tasks, part-of-speech (POS) tagging and named entity recognition (NER).

Although different embedding models have been around for more than a decade (Mikolov et al., 2013a; Pennington et al., 2014; Bojanowski et al., 2017; Akbik et al., 2018; Liu et al., 2019), there are still a limited number of monolingual embeddings available for under-resourced languages, and some of the available models (Grave et al., 2018) are trained only on data from Wikipedia, which may be limited in their scope. Most South African languages are considered agglutinative, with a rich morphosyntactic structure, but there is a distinction in the orthographies of the languages. The Sotho and Tswa-Ronga languages adhere to a disjunctive orthography where morphemes are written as separate tokens, e.g., Sesotho *ke a mo rata*, while others are conjunctive (morphemes fused into a single word), e.g., isiZulu *ngiyamthanda* - both meaning "I love him/her". These characteristics, when coupled with rich morphology and small training corpora, have been shown in the past to negatively impact the quality of sequence labelling models (Loubser and Puttkammer, 2020). There is also a practical question that remains under-documented: *Which embeddings are the best choice for languages under typical low-resource constraints?* Because orthography, morphology, and data availability differ markedly for the South African languages, this provides a good test bed for ascertaining how different embeddings behave between different typologies and training data availability.

This study has three aims: **(i)** introduce six different embedding models for ten South African languages from four embedding architectures; **(ii)** provide a controlled comparison of these embedding models for POS tagging and NER; and **(iii)** deter-

Proceedings of Resources for African Indigenous Languages (RAIL) 2026 @ LREC 2026, pages 41–51
12 May 2026.

mine how data availability and orthographic typology (conjunctive vs. disjunctive) systematically influence the quality of different embedding types.

By keeping the model, data splits, and training budget the same, this work offers an embedding benchmark that isolates representation effects. We evaluate GloVe (Pennington et al., 2014), fastText (Bojanowski et al., 2017), Flair (Akbik et al., 2018, 2019), RoBERTa (Liu et al., 2019), and XLM-R (Conneau et al., 2020) embedding models on POS tagging and NER for ten South African languages, grouping languages by conjunctive vs. disjunctive orthographies, and additionally Afrikaans.

Contributions

- Introduce embedding models for ten South African languages in four different architectures.
- Provide a controlled benchmark comparing the different embedding architectures across POS and NER tasks for ten South African languages.
- Provide a typology analysis contrasting results for conjunctive and disjunctive writing systems.

Key findings

For sequence labelling tasks like POS tagging and NER, Flair character-based LSTM language models show the highest and most consistent quality, particularly for conjunctive languages. fastText remains a close competitor, offering strong POS tagging performance at reduced computational cost.

2. Background and Related Work

2.1. Word Embeddings

Word embeddings map tokens to dense vectors that can serve as inputs to various linguistic processing applications, but also as an investigative tool for linguistic analysis. Although vectorised representations of words were first introduced in 2003 (Bengio et al., 2003), these representations became more widely used after the introduction of the Word2Vec (Mikolov et al., 2013b,a) and GloVe (Pennington et al., 2014) embedding models. These embeddings allowed for efficient training of representations that included both semantic and morphosyntactic information. This in turn allowed more complex neural architectures to accurately model various NLP tasks. Subsequently, alternative embedding models have been introduced, initially by including subwords in the calculation of embeddings (Bojanowski et al., 2017), followed by the introduction of contextualised embeddings on character (Akbik et al., 2018) and word level (Devlin et al., 2019).

Training embedding models requires large amounts of text data to learn informative representations, with the original models trained on corpora of several billion tokens (Pennington et al., 2014; Mikolov et al., 2013b; Bojanowski et al., 2017), while the latest transformer models for well-resourced and multilingual models are trained on hundreds of billions of tokens. Availability of data at this scale remains a significant challenge for most languages in the global South, and even when substantial amounts of data are available, the quality of the data is often questionable (Kreutzer et al., 2022).

We consider three families of embeddings with different trade-offs relevant to South African languages: (i) static word-level embeddings that assign one vector per token (GloVe); (ii) subword-aware static embeddings that compose embeddings from a combination of word and character n-grams (fastText); and (iii) contextual embeddings whose token representations depend on sentence context (Flair and RoBERTa). We emphasise out-of-vocabulary handling, morphological robustness, and data efficiency.

Static word vectors *GloVe* (Pennington et al., 2014) learns static word vectors by fitting a weighted least-squares model to global co-occurrence statistics, training the dot product of word and context vectors to approximate the logarithm of co-occurrence counts. GloVe represents each word type with a single vector and cannot produce unique representations for unseen words (OOVs).

Subword-aware static vectors *fastText* (Bojanowski et al., 2017) yields static yet subword-aware embeddings that are more robust to morphological complexity and can compose vectors for unseen (OOV) words. Each word is represented as the combination of the token and character n-gram vectors, trained with either a skip-gram or continuous bag-of-words (CBOW) objective.

Contextual embeddings *Flair* (Akbik et al., 2018, 2019) provides contextual string embeddings by training forward and backward character-level LSTM language models, making each token's representation a function of the characters comprising the token, as well as surrounding context. Transformer models (e.g., *BERT* (Devlin et al., 2019), *RoBERTa* (Liu et al., 2019)) produce subword-level contextual embeddings via self-attention, generally delivering strong performance at higher compute cost.

2.2. South African NLP Context

Over the last two decades, South African NLP has grown substantially with various research groups focussing on developing resources and NLP technologies for both the South African languages and African languages more generally, with a relatively substantial ecosystem of corpora and datasets (Eiselen and Puttkammer, 2014; Barnard et al., 2014; Orife et al., 2020; Adelani et al., 2021; Dione et al., 2023), yet several challenges remain: uneven data availability, orthographic differences (Loubser and Puttkammer, 2020), and frequent code-switching observed in multilingual communication (Moodley, 2007; Biswas et al., 2022). The Masakhane community (Orife et al., 2020) has advanced collaborative dataset creation and evaluation and have released the MasakhaNER (Adelani et al., 2021) and MasakhaPOS (Dione et al., 2023) datasets and baseline sequence labellers (CNN-BiLSTM-CRF; fine-tuned mBERT/XLM-R) for both South African and African languages. Local institutions have contributed parallel corpora (e.g. Autshumato (Groenewald and Fourie, 2009; Gaustad and McKellar, 2024)) and morphological analysers (du Toit and Puttkammer, 2021); recent work adds morphologically annotated corpora for nine South African languages (Gaustad and McKellar, 2024).

More recently, there have been several studies that evaluated the viability of neural approaches for South African languages. Loubser and Puttkammer (2020) reported strong gains on compound analysis and modestly lower averages for conjunctive vs. disjunctive languages on POS tagging and NER. du Toit and Puttkammer (2021) highlighted how typology impacts model design for Nguni languages. AfriBERTa (Ogueji et al., 2021) and AfroXLMR (Alabi et al., 2022) demonstrated that competitive multilingual transformers can be trained on relatively small African language corpora, underscoring the potential of contextual subword embeddings in low-resource settings.

Despite this progress, data scarcity and imbalance persist, specifically for the South African languages. While Afrikaans, isiXhosa, and isiZulu have larger corpora available, isiNdebele, Siswati, Tshivenḓa, and Xitsonga remain severely under-resourced. There are still several ongoing efforts to improve the availability of data in these languages, but most of these languages will remain under-resourced for the foreseeable future.

3. Embedding Models

Although the various embedding architectures use different strategies for learning vectorised representations, they all require large amounts of text data. In the context of the South African languages, the amount of data available for the different languages vary greatly, from very limited amounts of data available for isiNdebele, to relatively large amounts of data available for Afrikaans. In order to maximise the amount of available training data, several sources were investigated to ascertain their quality and usefulness in training embedding models, including OPUS (Skadiņš et al., 2014), Leipzig Corpora Collection (Goldhahn et al., 2012), CommonCrawl[1], NCHLT (Eiselen and Puttkammer, 2014), and Autshumato (Gaustad et al., 2024a,b; Gaustad and McKellar, 2024; Groenewald and Fourie, 2009).

After removing duplicate items, all data from these publicly available data collections were run through the NCHLT South African Language identifier (Puttkammer et al., 2016) to ensure that all data was in the requisite language. After language identification, the data was cleaned by removing items that were likely to be ill-formed, such as sentences consisting only of numbers, lines containing e-mail addresses and hyperlinks, and malformed UTF-8 characters. No capitalisation removal or normalisation was performed, since these characteristics may be helpful in some contexts such as named entity recognition, where capitalisation remains important. These sources were combined with internal material for which copyright agreements have been signed, but which is not in the public domain, to create monolingual datasets for each of the languages.

Table 1 provides a summary of the final sentence and token counts of the embedding training data available for each language. Apart from the disparity in data availability for the respective languages, the distinction between conjunctively and disjunctively written languages is also apparent from this table. As an example, although Sesotho (sot) and Siswati (ssw) have a similar amount of sentences available, Siswati has less than half the number of tokens. This is especially relevant to word embeddings, as the learned representations are based on word co-occurrence, and consequently, for languages with lower token counts and larger vocabulary sizes, the quality of the embeddings is also likely to be lower. In total, six different embedding models were trained for each language, two flavours for fastText, continuous bag-of-words (CBOW) and Skip-grams, GloVe embeddings, forward and backward Flair embeddings, and a RoBERTa masked language model. Details of the training hyperparameters for each of the models is provided in Table 3 in Appendix A.

[1] https://commoncrawl.org/

Language (ISO code)	Sentence count	Token count
Afrikaans (afr)	12,794,432	381,087,586
isiNdebele (nbl)	247,926	3,633,845
Sepedi (nso)	292,594	8,908,709
Siswati (ssw)	299,112	4,436,576
Sesotho (sot)	535,853	17,425,650
Setswana (tsn)	515,961	14,518,437
Xitsonga (tso)	360,698	7,357,764
Tshivenḓa (ven)	304,248	7,363,713
isiXhosa (xho)	718,751	13,190,962
isiZulu (zul)	816,776	15,801,081

Table 1: Embedding training data sentence and token counts

4. Experimental Setup

4.1. Datasets

In order to establish the quality and usefulness of the embeddings, we evaluate embeddings on two linguistic annotation tasks: POS tagging and NER. For both tasks, we use existing annotated corpora for South African languages from five publicly available data sets:

- MasakhaNER[2] and MasakhaPOS[3] data for tsn, xho, and zul;
- NCHLT annotated POS data (Eiselen and Puttkammer, 2014) for afr (Puttkammer et al., 2014a) and disjunctive languages (nso, sot, tsn, tso, and ven) (Puttkammer et al., 2014b,c,d,e,f);
- Linguistically enriched corpora (LEC) (Gaustad and Puttkammer, 2022) for conjunctively written languages (nbl, ssw, xho, and zul) (Puttkammer and Gaustad, 2021); and
- NCHLT Named entity annotated corpora for all languages (Eiselen, 2016) (Golele et al., 2016; Mahlangu and Eiselen, 2016; Malangwane et al., 2016; Manzini and Eiselen, 2016; Phakedi and Eiselen, 2016; Podile and Eiselen, 2016; Prinsloo and Eiselen, 2016; Setaka and Eiselen, 2016; Tshikota et al., 2016; van Huyssteen et al., 2016).

Given that the respective corpora annotated with POS did not all follow the same annotation schemas, and to make comparisons between the respective data sets possible, all annotations were simplified to the set of universal parts-of-speech as defined as part of the Universal dependency project (De Marneffe et al., 2021). Table 2 provides a summary of the respective data sets used in the benchmarking experiments.

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

[3] https://github.com/masakhane-io/masakhane-pos

Language	Source	Train	Test
	POS data sets		
tsn	Masakhane	26,211	15,520
xho	Masakhane	15,598	9,725
zul	Masakhane	14,802	9,215
afr	NCHLT	55,483	5,834
nbl	LEC	44,663	5,026
nso	NCHLT	65,908	7,153
sot	NCHLT	66,877	6,847
ssw	LEC	42,596	4,789
tsn	NCHLT	65,784	6,803
tso	NCHLT	64,534	6,518
ven	NCHLT	59,818	6,646
xho	LEC	43,825	4,910
zul	LEC	44,098	4,981
	NER data sets		
tsn	Masakhane	30,852	3,779
xho	Masakhane	30,513	3,863
zul	Masakhane	32,356	4,198
afr	NCHLT	205,977	23,841
nbl	NCHLT	161,544	15,006
nso	NCHLT	201,431	20,831
sot	NCHLT	269,624	21,966
ssw	NCHLT	175,378	15,731
tsn	NCHLT	231,390	19,444
tso	NCHLT	268,496	22,071
ven	NCHLT	235,450	17,144
xho	NCHLT	121,166	11,346
zul	NCHLT	201,331	18,593

Table 2: Token counts for POS and NER corpora across South African languages.

4.2. Embedding Models

We compare the six trained embedding models as discussed in Section 3 for each of the languages[4]. In addition to these monolingual models, we also include a multilingual language model, XLM-R-base, to establish whether the monolingual models outperform the multilingual model.

- **fastText** (CBOW, Skip-gram) (Bojanowski et al., 2017)
- **GloVe** (Pennington et al., 2014)
- **Flair** (forward, backward) (Akbik et al., 2018, 2019)
- **RoBERTa** (Liu et al., 2019)
- **XLM-R-base** (Conneau et al., 2020)

4.3. Training Framework and Hyperparameters

For all experiments, we use the Flair BiLSTM-CRF tagger (Akbik et al., 2019) to train and evaluate

[4] All embedding models are available for download from https://repo.sadilar.org/home

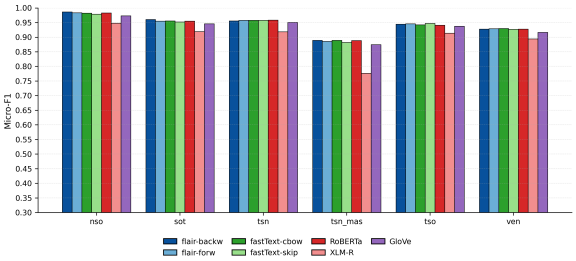

Figure 1: Micro-F1 scores for UPOS per embedding model variant for disjunctive languages

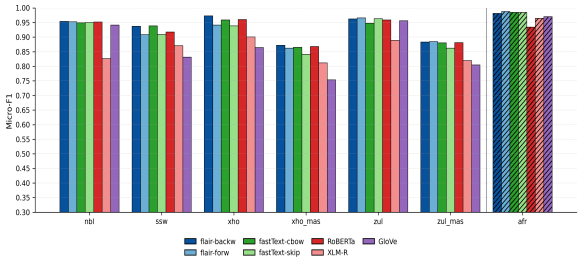

Figure 2: Micro-F1 scores for UPOS per embedding model variant for conjunctive languages and Afrikaans

the respective models. Although more modern architectures, such as transformer-based classifiers, have been shown to improve performance on some sequence labelling tasks, initial experiments with transformers consistently underperformed in the low resource environment.

In order to isolate the embeddings representation effects in the experiments, we keep the model parameters, data splits, and training budget constant across runs; the only factor that changes is the embedding type used. We optimize using stochastic gradient descent (SGD) with an $\mathrm{AnnealOnPlateau}$ learning rate scheduler, configured with a patience of 3 epochs, a reduction factor of 0.5, and a minimum learning rate (*min LR*) of 1×10^{-4}. We use word dropout (0.05) and a locked (variational) dropout (0.5). Training is run on an NVIDIA RTX5000 GPU.

For evaluation purposes, we use the the Micro-F1 on token level for POS tagging (a single token labelling task), and exact entity level match for NER (a multi-token labelling task).

5. Results

5.1. Part-of-Speech (POS) Tagging

Figures 1 and 2 show the results of POS tagging for embedding model variants per language for the disjunctive languages and the conjunctive languages (with afr shown separately) respectively [5]. For the disjunctive languages, the monolingual embedding models all perform very similarly, with GloVe embeddings performing slightly worse for all the disjunctive languages, and the XLM-R model performing substantially worse. However, there is no single embedding model which consistently performs best across all the languages. The two languages with significantly less embedding model training data, Xitsonga and Tshivenḓa, do perform worse than the other three disjunctive languages, but the difference is not as large as would be expected.

[5]Full results are available in Tables 4 and 5 in Appendix B

For the conjunctive languages, there is somewhat more variance for the respective embedding models, especially for the GloVe models that perform significantly worse than the other models, except for isiNdebele and isiZulu, even when compared to the XLM-R embeddings. This relatively poorer performance for GloVe is most likely due to the fact that the conjunctive languages have a much more complex morphology, and by extension a larger vocabulary with sparser representation in the embedding training corpora. Since GloVe does not account for morphological features, such as sub-words, it is not surprising that these embeddings perform the worst. Once again, there is no clear single embedding model that is better in all cases, but the Flair variants are generally one of the two best performing models in this task. As was the case for disjunctive languages, the XLM-R model performs substantially worse, even for languages included in its training regime (Afrikaans and isiXhosa). Rather surprisingly, both conjunctive languages with very little embedding training data, isiNdebele and Siswati, still attain relatively good POS tagging accuracy, indicating that even with little training data, monolingual embedding models do encapsulate enough linguistic information to provide accurate POS tagging.

The relatively lower scores on the Masakhane data sets for both conjunctive and disjunctive languages can mainly be ascribed to the fact that these data sets have substantially less training data, but may also be due to target domains of the respective data sets. Since a substantial part of the embedding training data originates from the government domain, and the NCHLT annotated data is also from the government domain, the learned representations may not be as representative of the domains from which the Masahkane data originates.

5.2. Named Entity Recognition (NER)

The results for NER are provided in Figures 3 and 4 for the disjunctive and conjunctive languages respectively The results reflect a similar pattern

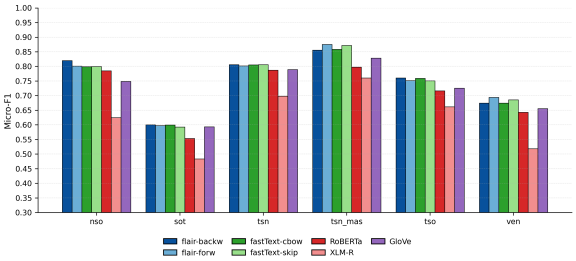

Figure 3: Micro-F1 scores for NER per embedding variant for disjunctive languages.

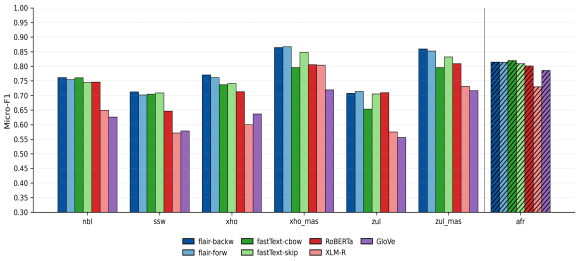

Figure 4: Micro-F1 scores for NER per embedding for conjunctive languages, including Afrikaans

to those found for POS tagging, with relatively little variance for the disjunctive languages, somewhat more variance for the conjunctive languages, and no single embedding model performing best across the board. As with the POS experiments, GloVe embeddings perform substantially worse than the other embeddings, especially for the conjunctive languages. In general though, the best performing embeddings across languages appear to be one of the Flair variants, especially for the conjunctive languages, and can likely be attributed to the fact that the character-level contextual embeddings are better suited to provide the necessary vector information to distinguish named entities in these languages.

The substantially poorer performance of Sesotho and Tshivenḓa ($<$.70) appears to be a function of annotation consistency in the training data, but may also indicate some shortcomings in the embedding models for these languages, especially since Sesotho has a relatively large embedding training corpus, and performance on par with Sesotho sa Leboa and Setswana would be expected. It is also surprising that both isiNdebele and Siswati perform relatively well in this task, given the limited available embedding training data.

With regard to the results on the Masakhane data sets, and specifically the XLM-R results, our results somewhat contradict those presented by (Dione et al., 2023; Adelani et al., 2022). This may be due to the fact that, in their experiments, they fine-tuned the embedding models as part of their training regime. We opted not to fine-tune the embeddings based on the assumption that the training sets under consideration are very small, and adjusting the embeddings may hinder their quality on a larger or more general test set.

6. Discussion

From our experiments with different monolingual embedding models across a range of languages and typologies, Flair embeddings is the most consistent top performer across both POS and NER tasks and different language typologies. fastText embeddings still offer a strong and computationally efficient baseline (within less than $\approx$0.3 point for POS and $\approx$1 point for NER on average). The more complex transformer RoBERTa models match Flair on POS but do not surpass Flair or fastText on NER. The static GloVe embeddings perform notably worse across both tasks and for all languages, especially for the conjunctive languages, where their lack of subword modelling and handling of out-of-vocabulary or morphologically complex tokens limits the quality of the available representations. Furthermore, all of the models outperform the multilingual XLM-R model, even when a language has been included in the XLM-R training regime.

7. Conclusion

In this work we introduced six monolingual embedding models for ten South African languages from four embedding architectures, namely GloVe, fastText, Flair, and RoBERTa. The embedding models were trained on a combination of publicly available corpora and institution-internal copyrighted material, with a large variance between the largest corpus, more than 380 million tokens for Afrikaans, and the smallest, only 3.6 million tokens for isiNdebele. This disparity in training corpus size, and the diverse typographic nature of the South African languages made these models an ideal test bed to evaluate the impact both training data size and language typology have on linguistic annotation tasks.

In our experiments for POS tagging and NER we found that although languages with larger embedding training sets outperformed those with smaller sets, the embedding representations across all of the languages produced relatively acceptable results for all language with one or two exceptions. It was also shown that these monolingual embedding models outperform the multilingual XLM-R model across all experiments, even for languages included in the XLM-R training regime. Although most embedding models performed comparably

within each language, Flair embeddings consistently perform well, irrespective of task or typology. The more complex and computationally expensive transformer RoBERTa models did not perform as well, but may be more suited to tasks other than linguistic sequence labelling in low-resource environments.

Although these models provide a good baseline and starting point for providing vectorised representations, there are several avenues for future research that remain. In addition to fine-tuning existing multilingual models, and specifically Afrocentric models, with additional data for these under-resourced languages, these embedding models should be tested on a wider variety of tasks beyond sequence labelling. Further research into the nature and internal representations of the embedding models may also provide greater insight into the type of information encoded in the models that may be used in further NLP developments and computational linguistic research for the South African languages.

8. Limitations

Our conclusions are limited by dataset size imbalances across languages and by a fixed architecture; results may vary with larger models or alternative taggers, as well as other settings or configurations. Furthermore, much of the data used to train the embeddings and the subsequent taggers originate from government documents, especially for languages with very little data. These models and the reported results may substantially degrade on test sets in other domains. Lastly, the embedding models have only been tested on two sequence labelling tasks, and their utility for more complex tasks and other benchmarks should be verified.

9. Bibliographical References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'Souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. 2021. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

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

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 54–59, Minneapolis, USA. ACL.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 1638–1649, Santa Fe, New Mexico. ACL.

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.

Etienne Barnard, Marelie H. Davel, Charl van Heerden, Febe de Wet, and Jaco Badenhorst. 2014. The NCHLT speech corpus of the south african languages. In *Proceedings of the 4th Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU)*, St. Petersburg, Russia. ISCA.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Astik Biswas, Emre Yılmaz, Ewald van der Westhuizen, Febe de Wet, and Thomas Niesler. 2022. Code-switched automatic speech recognition in five south african languages. *Computer Speech & Language*, 71:101262.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jonathan Mukiibi, Blessing Sibanda, Bonaventure F. P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Fatoumata Ouoba Kabore, Amelia Taylor, Godson Kalipe, Tebogo Macucwa, Vukosi Marivate, Tajuddeen Gwadabe, Mboning Tchiaze Elvis, Ikechukwu Onyenwe, Gratien Atindogbe, Tolulope Adelani, Idris Akinade, Olanrewaju Samuel, Marien Nahimana, Théogène Musabeyezu, Emile Niyomutabazi, Ester Chimhenga, Kudzai Gotosa, Patrick Mizha, Apelete Agbolo, Seydou Traore, Chinedu Uchechukwu, Aliyu Yusuf, Muhammad Abdullahi, and Dietrich Klakow. 2023. Masakha-POS: Part-of-Speech tagging for typologically diverse African languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10883–10900, Toronto, Canada. Association for Computational Linguistics.

J. du Toit and M. Puttkammer. 2021. Neural approaches to core technologies for nguni languages. *Information*, 12(7):276.

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).

Roald Eiselen and Martin J. Puttkammer. 2014. Developing text resources for ten South African languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 3698–3703, Reykjavik, Iceland. European Language Resources Association (ELRA).

Tanja Gaustad and Cindy McKellar. 2024. Updated morphologically annotated corpora for 9 South African languages. *Journal of Open Humanities Data*, 10(38):1–5.

Tanja Gaustad, Cindy McKellar, and Martin Puttkammer. 2024a. Dataset for Siswati: Parallel textual data for English and Siswati and monolingual textual data for Siswati. *Data in Brief*, 54.

Tanja Gaustad, Cindy A. McKellar, and Martin J. Puttkammer. 2024b. Machine translation training data for English–Tshiven a. *Data in Brief*, 57.

Tanja Gaustad and Martin J. Puttkammer. 2022. Linguistically annotated dataset for four official South African languages with a conjunctive orthography: isiNdebele, isiXhosa, isiZulu, and Siswati. *Data in Brief*, 41.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the 8th International Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resource Association (ELRA).

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Hendrik Johannes Groenewald and Wildrich Fourie. 2009. Introducing the autshumato integrated translation environment. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT)*, Barcelona, Spain.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, and Claytone Sikasote. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Melinda Loubser and Martin J. Puttkammer. 2020. Viability of neural networks for core technologies for resource-scarce languages. *Information*, 11(1):41.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NeurIPS 2013)*, volume 26.

Visvaganthie Moodley. 2007. Codeswitching in the multilingual english first language classroom. *International Journal of Bilingual Education and Bilingualism*, 10(6):707–722.

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.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, et al. 2020. Masakhane: Machine translation for africa. *arXiv preprint arXiv:2003.11529*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. ACL.

Raivis Skadiņš, Jörg Tiedemann, Roberts Rozis, and Daiga Deksne. 2014. Billions of parallel words for free: Building and using the eu bookshop corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

10. Language Resource References

N.C.P. Golele and X.E. Mabaso and Roald Eiselen. 2016. *NCHLT Xitsonga Named Entity Annotated Corpus*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/362.

K.S. Mahlangu and Roald Eiselen. 2016. *NCHLT isiNdebele Named Entity Annotated Corpus*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/306.

B.B. Malangwane and M.N. Kekana and S.S. Sedibe and B.C. Ndhlovu and Roald Eiselen. 2016. *NCHLT Siswati Named Entity Annotated Corpus*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/346.

A.N. Manzini and Roald Eiselen. 2016. *NCHLT isiZulu Named Entity Annotated Corpus*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/319.

S.S.B.M. Phakedi and Roald Eiselen. 2016. *NCHLT Setswana Named Entity Annotated Corpus*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/341.

K. Podile and Roald Eiselen. 2016. *NCHLT isiXhosa Named Entity Annotated Corpus*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/312.

D.J. Prinsloo and Roald Eiselen. 2016. *NCHLT Sepedi Named Entity Annotated Corpus*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/328.

Martin Puttkammer and Tanja Gaustad. 2021. *Linguistically enriched corpora for conjunctively written South African languages*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/546.

Martin Puttkammer and Justin Hocking and Roald Eiselen. 2016. *NCHLT South African Language Identifier*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/350.

Martin Puttkammer and Martin Schlemmer and Ruan Bekker. 2014a. *NCHLT Afrikaans Annotated Text Corpora*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/296.

Martin Puttkammer and Martin Schlemmer and Ruan Bekker. 2014b. *NCHLT Sepedi Annotated Text Corpora*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/325.

Martin Puttkammer and Martin Schlemmer and Ruan Bekker. 2014c. *NCHLT Sesotho Annotated Text Corpora*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/332.

Martin Puttkammer and Martin Schlemmer and Ruan Bekker. 2014d. *NCHLT Setswana Annotated Text Corpora*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/337.

Martin Puttkammer and Martin Schlemmer and Ruan Bekker. 2014e. *NCHLT Siswati Annotated Text Corpora*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/344.

Martin Puttkammer and Martin Schlemmer and Ruan Bekker. 2014f. *NCHLT Tshivenda Annotated Text Corpora*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/353.

M. Setaka and Roald Eiselen. 2016. *NCHLT Sesotho Named Entity Annotated Corpus*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/334.

S.L. Tshikota and M.E. Takalani and A. Nyoni and Roald Eiselen. 2016. *NCHLT Tshivenda Named Entity Annotated Corpus*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/355.

Gerhard van Huyssteen and Martin Puttkammer and E.B. Trollip and J.C. Liversage and Roald Eiselen. 2016. *NCHLT Afrikaans Named Entity Annotated Corpus*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/299.

Appendix A: Embedding Model Training Parameters

Table 3 provides details on pertinent hyperparameters used during training of the embedding models. The reported hyperparameters were initially selected after running preliminary tests for both conjunctive and disjunctive languages. Due to the large number of possible hyperparameters, only parameters that were explicitly tested and deviate from default values are reported here.

Parameter	Afrikaans	Conjunctive	Disjunctive
fastText-CBoW			
Dimensions	600	600	600
Epochs	40	40	40
Min occur	5	2	5
Min n	5	2	3
Max n	5	4	4
fastText-Skipgram			
Dimensions	500	600	600
Epochs	40	40	40
Min occur	5	2	5
Min n	2	2	2
Max n	4	6	6
Flair (forward/backward)			
Hidden size	2048	2048	2048
Epochs	20	20	20
Layers	2	2	2
Sequence length	250	250	250
GloVe			
Dimensions	500	300	400
Max iter	50	50	50
Min occur	5	2	5
Window	20	20	20
RoBERTa			
Hidden size	768	768	768
Epochs	40	40	40
Vocabulary	30000	30000	30000
Attn. heads	6	6	6
Layers	6	6	6

Table 3: Training hyperparameters for embedding models across Afrikaans, conjunctive, and disjunctive languages.

Appendix B: Full Evaluation Results

	Embedding model						
Language	fastText-cbow	fastText-skipgram	Flair-backw	Flair-forw	GloVE	RoBERTa	XLM-R
afr	0.9847	0.9851	0.9817	**0.9878**	0.9702	0.9349	0.9647
nbl	0.9491	0.9507	**0.9539**	0.9528	0.9417	0.9524	0.8267
nso	0.9824	0.9789	**0.9863**	0.9835	0.9734	0.9832	0.9480
sot	0.9555	0.9518	**0.9604**	0.9552	0.9457	0.9552	0.9191
ssw	**0.9390**	0.9096	0.9371	0.9090	0.8313	0.9175	0.8710
tsn	0.9578	0.9578	0.9556	0.9578	0.9509	**0.9587**	0.9184
tsn (Mas)	**0.8891**	0.8825	0.8889	0.8855	0.8743	0.8885	0.7769
tso	0.9422	**0.9469**	0.9445	0.9460	0.9373	0.9409	0.9135
ven	**0.9294**	0.9270	0.9273	0.9288	0.9168	0.9276	0.8938
xho	0.9593	0.9391	**0.9733**	0.9415	0.8646	0.9607	0.9008
xho (Mas)	0.8653	0.8416	**0.8724**	0.8624	0.7536	0.8681	0.8123
zu	0.9476	0.9636	0.9627	**0.9643**	0.9561	0.9589	0.8892
zu (Mas)	0.8804	0.8623	0.8837	**0.8846**	0.8047	0.8812	0.8197

Table 4: Micro-F1 scores for UPOS per embedding model variant across ten South African languages

	Embedding model						
Language	fastText-cbow	fastText-skipgram	Flair-backw	Flair-forw	GloVE	RoBERTa	XLM-R
afr	**.8198**	.8094	.8150	.8137	.7863	.8014	.7300
nbl	.7609	.7444	**.7613**	.7547	.6264	.7459	.6492
nso	.7988	.7995	**.8199**	.8007	.7485	.7847	.6250
sot	.5999	.5929	**.6006**	.5973	.5930	.5536	.4831
ssw	.7049	.7088	**.7124**	.7022	.5784	.6465	.5714
tsn	.8052	.8058	**.8059**	.8020	.7891	.7867	.6981
tsn (Mas)	.8586	.8718	.8555	**.8755**	.8281	.7975	.7602
tso	.7594	.7507	**.7608**	.7514	.7255	.7167	.6616
ven	.6741	.6854	.6743	**.6943**	.6557	.6431	.5186
xho	.7372	.7413	**.7708**	.7617	.6369	.7133	.6006
xho (Mas)	.7956	.8475	.8646	**.8670**	.7198	.8054	.8040
zu	.6536	.7054	.7077	**.7137**	.5564	.7099	.5753
zu (Mas)	.7958	.8320	**.8596**	.8529	.7167	.8090	.7314

Table 5: Micro-F1 scores for NER per embedding model variant across ten South African languages