Deep learning and low-resource languages: How much data is enough? A case study of three linguistically distinct South African languages

Roald Eiselen and **Tanja Gaustad**
Centre for Text Technology (CTexT)
North-West University
Potchefstroom, South Africa
roald.eiselen@nwu.ac.za, tanja.gaustad@nwu.ac.za

Abstract

In this paper we present a case study for three under-resourced linguistically distinct South African languages (Afrikaans, isiZulu, and Sesotho sa Leboa) to investigate the influence of data size and linguistic nature of a language on the performance of different embedding types. Our experimental setup consists of training embeddings on increasing amounts of data and then evaluating the impact of data size for the downstream task of part of speech tagging. We find that relatively little data can produce useful representations for this specific task for all three languages. Our analysis also shows that the influence of linguistic and orthographic differences between languages should not be underestimated: morphologically complex, conjunctively written languages (isiZulu in our case) need substantially more data to achieve good results, while disjunctively written languages require substantially less data. This is not only the case with regard to the data for training the embedding model, but also annotated training material for the task at hand. It is therefore imperative to know the characteristics of the language you are working on to make linguistically informed choices about the amount of data and the type of embeddings to use.

1 Introduction

Over the last decade vectorised word representations and the use of deep learning have become de facto standards in Natural Language Processing (NLP) (Alzubaidi et al., 2021; Khurana et al., 2023). There has also been a push to broaden the linguistic diversity in NLP research (Joshi et al., 2020). Both learning vectorised representations, a.k.a. embeddings, and deep learning are inherently data-driven procedures where models are trained from vast amounts of data to either represent language numerically or learn some downstream task. Including a bigger variety of languages than mainstream languages, such as English, Spanish, German, Japanese, etc., to achieve more linguistic diversity typically means studying low-resource or under-resourced languages.

This, however, leads to a dichotomy: High-performing deep learning models, like BERT, have been trained on billions of words. When developing models for languages other than English, lesser resourced languages (such as the South African languages) get left behind because there is very little available data. Also, evaluation of existing techniques is only partially applied to under-resourced languages and researchers typically assume that the generalisations achieved with training on a lot of data will mostly hold true with less data.

More recently there have been efforts to extend the usefulness of embeddings trained on well-resourced languages with languages that have substantially less data in so-called multi-lingual models, such as XLM-R (Conneau et al., 2020) and mBERT (Devlin et al., 2019). These models generate representations that are a combination of language specific information as well as information learned across all of the languages included in the model. Typically they are trained exclusively on web data like Common Crawl or Wikipedia, which have some inherent limitations (as discussed later in this paper), but also have limited availability for South African languages. For instance, isiNdebele has no Wikipedia data and is therefore not even present in Common Crawl. Consequently, most of the South African languages are not included in these multilingual models.

Doing NLP research on South African languages in this day and age then leads to the question of what the implications of working with very little data is on current standard techniques like embeddings and neural language models. Or in other words: how much data is needed for learning useful vector representations? The underlying assumption is that learning from less data will yield less representative and thus less useful models. It re-

Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023), pages 42–53

mains to be seen, however, if this is truly the case. From a linguistic diversity point of view, it is also relevant to know how the embedding models vary from each other for structurally different languages and how the amount of available data influences the learned representations for typographically different languages.

In this paper, we present a case study attempting to answer these questions. Our setup consists of training embeddings for three linguistically distinct South African languages (Afrikaans, isiZulu, and Sesotho sa Leboa) to evaluate the impact of embeddings trained on increasing amounts of data for a part of speech (POS) tagging downstream task. The goal is to determine the influence of data size on the performance of different embedding types and to describe the effects observed for different languages. The results of our experiments show that even relatively little data can be useful in some scenarios and that morphologically complex and conjunctively written languages require substantially more data, both for training the embeddings and the downstream task, especially when using full/sub word representations.

2 Background

2.1 South African linguistic context

South Africa's eleven official languages include nine Niger-Congo-B (NCB) languages and two Germanic languages. The NCB languages (van der Velde et al., 2022) have a number of linguistic characteristics that make them substantially different from most Indo-European languages: all of them are tone languages; they use an elaborate system of noun classes with up to 21 classes; and their nominal and verbal morphology is highly agglutinative and very productive, which can result in a large vocabulary for those languages that follow the conjunctive writing system.

For historic reasons, the South African NCB languages adopted two different writing systems, either conjunctive or disjunctive, where a distinction is generally made between linguistic words and orthographic words. For conjunctively written languages one orthographic word (token) corresponds to one or more linguistic words, whereas for the disjunctively written languages several orthographic words can correspond to one linguistic word (Louwrens and Poulos, 2006). The four Nguni languages, isiNdebele, isiXhosa, isiZulu, and Siswati, are written conjunctively, while the three Sotho languages, Sesotho, Sesotho sa Leboa (also known as Sepedi), and Setswana, Tshivenḓa, a Venda language, as well as Xitsonga, a Tswa-Ronga language, are disjunctively written. This is a marked difference from the two Germanic languages present in South Africa, Afrikaans and English, where mostly a linguistic word and an orthographic word coincide.

The implication of these different writing systems is that multiple tokens in disjunctively written languages can correspond to a single token in the conjunctively written languages. This leads to sparse token frequency for the conjunctively written languages, while the opposite is true for the disjunctive languages. As an illustration, the parallel equivalents of a 50,000 word English corpus will have approximately 43,000 words for the conjunctively written languages, while the disjunctive languages will have approximately 60,000 tokens. This difference and its implications are discussed in more detail in Section 3.

One of the objectives of this paper is to investigate the influence of linguistic and orthographic differences in South African languages on using embedding models in NLP tasks, specifically POS tagging. To that purpose we have chosen one language from each family for our experiments: Afrikaans (Germanic), isiZulu (conjunctively written Nguni) and Sesotho sa Leboa (disjunctively written Sotho).

2.2 Embedding models

Since the introduction of the word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) vectorised word representations, most (if not all) NLP tasks make use of learned vector representations, referred to as embeddings, to model the occurrences of words and their context. With these algorithms, embedding models are trained efficiently on large amounts of data and the learned representations, in combination with deep learning techniques, generally improve the results of downstream NLP tasks. In this paper we apply three embedding architectures, namely fastText (an extension of word2vec) and GloVe, two classical embeddings, and FLAIR embeddings, a character-based recurrent neural network.

GloVe embeddings (Pennington et al., 2014) learn representations of words using global co-occurrences to train a log-bilinear regression model. For each word in the vocabulary of the GloVe model a single n-dimensional vector is learned,

while all unseen words generate the same vector representation. fastText embeddings (Bojanowski et al., 2017), an extension of word2vec (Mikolov et al., 2013), are based on local co-occurrences of words. In addition to the full word, the generation of the vector representations includes character n-grams, or "subwords", allowing these embeddings to also generate distinct representations for previously unseen words by combining the n-grams from the unseen word. fastText embeddings come in two variants, continuous bag-of-words (CBoW) and Skipgram models. Both GloVe and fastText learn a single vector representation for each word in the vocabulary. When retrieving this vector, the word will always receive the same vector, irrespective of the context in which the word appears.

In contrast, FLAIR embeddings (Akbik et al., 2018, 2019) learn representations for character sequences by training a long-short-term-memory (LSTM) recurrent neural network. This means that a distinct word occurring in different contexts can have a different vector representation depending on the character sequences (context) around the word. Furthermore, all character sequences receive a representation whether a sequence has been seen during training or not. This can help with the representation of rare or misspelled words as well as with individual morphemes or morphologically complex words. FLAIR allows for two variants, namely Forward and Backward depending on the direction in which the text is processed – either from the start (forward) or from the end (backward).

With the availability of these improved deep learning frameworks and an increased focus on linguistic diversity in the deep learning community, there has been a substantial rise in research on African languages. The focus of this work has been broad: From applications of embeddings and deep learning on individual languages and individual applications (Dlamini et al., 2021; Heyns and Barnard, 2020; Loubser and Puttkammer, 2020; Marivate et al., 2020; Ralethe, 2020), to investigations of multilingual embedding architectures for African languages (Alabi et al., 2022; Hanslo, 2021; Moeng et al., 2022) and transfer learning from well-resourced languages (Hedderich et al., 2020). The outcomes of these investigations have had mixed results, in some cases substantially improving technologies over previous best results, while other approaches show how the nature and quality of the data have a significant impact on the quality of the trained models. As far as we are aware, none of these studies have explicitly investigated the quality and nature of the embeddings when considering data size and morphosyntactic attributes of African languages.

3 Data

The major prerequisite for training embeddings and language models for any language is the availability of large amounts of text data. Although there have been several efforts to create such corpora for the South African languages (Eiselen and Puttkammer, 2014; Goldhahn et al., 2012; Marivate et al., 2020), there is still relatively little data available for most of them. The data collected for this study is a combination of various open data sets (mostly CC-BY and CC-NC licenses), as well as some data only available to the authors with copyright restrictions prohibiting the distribution of the full corpora. The data included in the training corpora for the isiZulu and Sesotho sa Leboa embeddings are primarily from the NCHLT Text Corpora (Eiselen and Puttkammer, 2014; Puttkammer et al., 2014c,d,e), Autshumato Corpora (McKellar, 2022a,b,c), Leipzig Corpus Collection (Goldhahn et al., 2012)[1], and Common Crawl corpus[2]. All of these sources are also used in the Afrikaans training corpus, along with additional data from publishers and private sources, i.e. the NWU/Lapa Corpus, NWU/Protea Boekhuis Corpus, and NWU/ATKV-Taalgenoot Corpus.

Although the data in both the Leipzig and Common Crawl corpora are language identified, an initial investigation showed that a substantial amount of the data is incorrectly attributed to one of the languages. This is primarily due to the fact that all three languages in this study have related languages that share similar orthographic features which leads to misclassification of the language data, specifically:

- Afrikaans ⇔ Dutch;
- isiZulu ⇔ isiNdebele, isiXhosa, and Siswati;
- Sesotho sa Leboa ⇔ Setswana and Sesotho.

Consequently, all of the data from the Leipzig and Common Crawl corpora were further cleaned with the NCHLT Language Identifier (Hocking,

[1] https://corpora.uni-leipzig.de/en
[2] https://commoncrawl.org

Language	Embeddings			POS tagging				
	Tokens	Vocab	Token:Vocab ratio	Train	Dev	Test	Orig. tags	Red. tags
Afrikaans	40,610,635[a]	311,719	0.0077	50,034	5,451	5,835	97	12
isiZulu	16,271,123	488,822	0.0300	39,768	4,376	4,955	97	17
Sesotho sa Leboa	8,909,133	80,919	0.0091	53,745	5,556	7,127	138	14

[a]Please note that this data was sampled from a larger 430 million token corpus.

Table 1: Summary of data available for training embeddings and POS tagging

2014; Puttkammer et al., 2018) at 80% confidence level. Since most of the data in the respective corpora originate from the web, all duplicates on paragraph level in the combined data are removed prior to training.

A summary of the data available for training embeddings is presented in Table 1. As was discussed in Section 2.1, there is a marked difference in the number of tokens in the vocabulary for each of the three languages for the same corpus sizes. One way of representing the combined effects of these morphosyntactic and writing system differences is by adding the token-vocabulary ratio to the reported token counts: a text in the conjunctively written, morphologically complex language of isiZulu typically displays a higher token-vocabulary ratio than a text in Sesotho sa Leboa, where a number of morphemes are written separately and therefore count as multiple tokens. In Afrikaans, where one token typically corresponds to one orthographic word, the token-vocabulary ratio is somewhere between the two extremes of the disjunctive and conjunctive languages. The vocabulary for Afrikaans is still more sparse than is typical in English since compounding is very common in Afrikaans and leads to a larger number of unique tokens, although it is not nearly as productive as the conjunctively written isiZulu.

The POS data used in this study is the NCHLT Annotated Corpora for Afrikaans and Sesotho sa Leboa (Puttkammer et al., 2014a,b), and the Linguistically enriched corpora for conjunctively written South African languages for isiZulu (Gaustad and Puttkammer, 2022; Puttkammer and Gaustad, 2021). Each annotated corpus consists of approximately 50,000 tokens for the training set, and a separate test set of approximately 5,000 tokens. Although the data is annotated on very fine-grained POS tag sets (typically consisting of 90+ tags), for this investigation we reduced the tag sets to between 12 and 17 tags by e.g. excluding class information and using only main POS classes. This makes the results between languages more comparable, but does not obscure the functional differences a conversion to UPOS[3] would. An overview of the POS data and tags is presented in Table 1.

4 Experimental Design

In order to determine the impact of different embedding architectures and morphosyntactic attributes on the usefulness of embeddings in low-resource environments, we perform a set of experiments to establish how these attributes in combination with data size affect the quality of a single downstream task, namely POS tagging.

The first step in the process is generating embeddings in each of the chosen architectures – fastText, GloVe, and FLAIR – with different data set sizes. For each language a random selection of paragraphs from the available corpus is made in iteratively larger sizes, starting with 10,000 paragraphs and doubling the amount of data randomly for each iteration. For isiZulu and Sesotho sa Leboa this process is repeated up to the full available corpus (292,600 and 838,000 paragraphs respectively), while for Afrikaans we only select data up to one increment above the largest of the other two languages (1,280,000 paragraphs).[4] Based on each data iteration, embeddings for all three architectures, including their different flavours, are trained.

To make the comparison of models as consequent as possible, the hyperparameters for each of the architectures are kept the same (typically the default settings, see Table 2) and no hyperparameter tuning is performed. Consequently, there may be certain hyperparameter selections for the different data set sizes and languages, that could lead to slight improvements in the results presented in this work, but different hyperparameters would make the comparison and resultant conclusions less generally applicable. Furthermore, this would also substantially increase the number of experiments that need to be trained (probably into the thousands) and cannot be ethically and environmentally

[3]https://universaldependencies.org/u/pos/

[4]See Table 3 in appendix for details.

Embedding	Embedding type	Hyperparameters
GloVe	Static word	Dimensions: 300 Epochs: 50 Min. occurences: 2 Window size: 20
fastText	Static word and subword	Dimensions: 300 Learning rate: 0.05 Epochs: 15 Min. occurences: 2 Minimum n: 3 Maximum n: 6
FLAIR	Contextual character	Dimensions: 2048 Learning rate: 10.0 Epochs: 15 Sequence length: 250 Layers: 1 Batch size: 64

Table 2: Hyperparameter settings for embedding training

justified due to additional power consumption. In total 110 embedding models are trained on an Intel i79700 CPU and four NVIDIA GeForce RTX 2060 6Gb GPUs, totalling 30 CPU hours and 252 GPU hours.

The quality of the embeddings trained on the different corpus sizes for the three languages is evaluated using a vanilla bidirectional LSTM-CRF POS tagger implemented as part of the FLAIR framework[5] (Akbik et al., 2019). A separate POS tagger is trained for each of the embedding models, on the same GPU hardware used to train the embeddings. The main reason for using the FLAIR framework is the fact that it has built-in support for all of the different embedding architectures, thus ensuring that the results of the respective taggers can reasonably be compared. As with the embedding models, the hyperparameters are kept to the default settings (hidden size: 256, learning rate: 0.1, drop out: 0.05, epochs: 40). It should also be noted that because the POS training sets are relatively small, fine-tuning of the embeddings during the POS tagger training is not carried out. All taggers are evaluated using the Accuracy metric, and compared to the NCHLT Web Services[6] POS taggers (Puttkammer et al., 2018) as a baseline for each language.

With such a large set of taggers (24 in total), the POS taggers were not trained multiple times with averaged scores. Doing so would once again substantially increase the required GPU hours (currently 82) required for the experiment, and cannot be justified, since the aim of the paper is not to create the best possible tagger, but rather to establish how the different embeddings influence the downstream results.

[5] https://github.com/flairNLP/flair

[6] https://hlt.nwu.ac.za/

5 Results and discussion

We will now discuss the performance of the various embedding models with different data sizes for Afrikaans, isiZulu and Sesotho sa Leboa. A graphic representation of the performances on the POS tagging task can be found in Figures 1 (for Afrikaans), 2 (for isiZulu) and 3 (for Sesotho sa Leboa), with the full numerical results available in Table 3 in the appendix.

The first notable conclusion that can be drawn from inspecting the results is that even embeddings trained on very small corpora can benefit the quality of relatively simple downstream tasks, such as POS tagging, when compared to baseline systems. Specifically, the FLAIR contextual character embeddings produce downstream results that are surprisingly good, even for the conjunctively written isiZulu. It was expected that the character-based models would perform best with very small amounts of data, but the models trained on the smallest data sets actually perform comparably to the best results for any of the other embeddings. Conversely, although the FLAIR models perform best when trained with the largest data sets, there is not nearly the same level of improvement as there is for the other two architectures. One implication of these findings is that much smaller and faster models may perform well enough for certain purposes, if not necessarily attaining state-of-the-art results. This allows researchers and developers with limited hardware capacity to also benefit from using these types of embeddings.

As expected, GloVe embeddings consistently perform the worst of all the embedding types, especially so with the first couple of iterations of very small corpora, for two main reasons. Firstly, as these embeddings only generate representations for words in the vocabulary, all words in the tagging task that are not part of the vocabulary are represented by the same vector, and therefore do not have any distinctive representations. Secondly, since many words will only appear a small number of times in the training data, learning complex representations is difficult when a word is only seen in a small number of contexts. These problems are exacerbated for isiZulu where the conjunctive

writing style causes a large number of distinctive co-occurrences for all words, especially less frequent words, and learning meaningful representations is almost impossible. For the disjunctively written Sesotho sa Leboa, however, the vocabulary is relatively representative even with a small corpus, and GloVe embeddings perform only slightly worse than the other embedding types.

The fastText models perform substantially better than the GloVe models with very small data sets, while still performing worse than the FLAIR embeddings. With the largest data sets, however, the fastText CBoW models perform either very similarly or better than the FLAIR models. Interestingly, with very small data sets the Skipgram models outperform the CBoW models for the first two or three data iterations, after which the CBoW models consistently perform better across all languages. It is not immediately obvious why this would be the case, but the fact that this occurs across all three languages definitively shows that with very small data sets Skipgrams are preferable over CBoW, whereas for any data set with more than 500,000 or a million tokens (corresponding to about 40,000 paragraphs in our data), the CBoW models generate better representations measured on the POS tagging task. This contradicts the initial findings of Mikolov et al. (2013), but is in line with the latest released fastText models[7] which have also switched to CBoW models by default, as opposed to previously released models (Bojanowski et al., 2017).

Our results also clearly show that the writing system of the language plays a major part in how much data is required to train embeddings that are useful to any degree. For conjunctively written languages, as the token-level morphological complexity of the language increases, so does the amount of data required to create meaningful representations. In the two extreme cases of isiZulu and Sesotho sa Leboa, even the embeddings from the largest available corpora for isiZulu perform substantially worse than the embeddings based on the smallest Sesotho sa Leboa corpus, with accuracies between 4.28% and 15.19% lower depending on the model. Afrikaans, which is slightly more morphologically complex on token level than Sesotho sa Leboa, but not nearly as complex as isiZulu, also performs somewhere between the two languages when considering the different data sizes and embeddings.

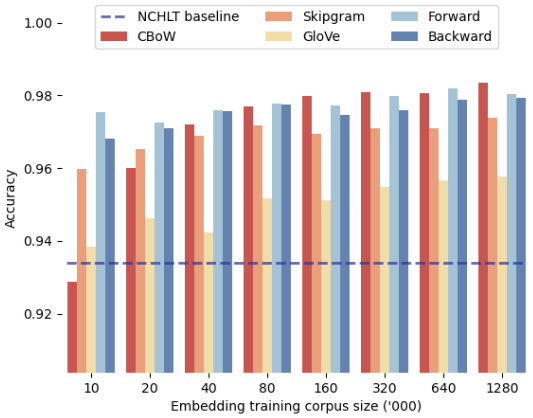

Figure 1: Accuracy of Afrikaans POS tagging using different embedding models with increasing data sizes

Apart from the general findings presented in the previous paragraphs, there are also certain language specific aspects of the results that warrant discussion. We include some broad linguistic error analysis for each of the languages to determine where the main focus of errors are for the best models for each language.[8]

For Afrikaans, the fastText CBoW model performs the worst of all models on the smallest data set, but shows the largest degree of improvement as the data size increases, to the point where it is the best performing of all models on the largest data set. Also, the Afrikaans FLAIR Forward model performs at almost an identical level to the fastText CBoW model, while the FLAIR Backward model is slightly worse. When considering the tag error differences between the embeddings trained on the smallest and largest corpora, it becomes clear that the main source of improvements for both fastText and GloVe embeddings is the size of the data. As more words are included in the vocabulary, the relative error rates for nouns, verbs, adjectives, and adverbs are reduced by between 35% and 45%. The relative error rate reductions for the FLAIR models are not as uniform across all open word classes. As an example, the FLAIR Forward model reduces the percentage of errors for adjective and verbs by 43% and 72% respectively, while increasing the number of errors for nouns or adverbs. The FLAIR Backward model shows improvements for adjectives (25%) and verbs (61%) as well, but also substantial improvements for nouns (64%) and adverbs (48%).

[7] https://fasttext.cc/docs/en/crawl-vectors.html

[8] Detailed information on the linguistic error analysis can be found in Tables 4, 5 and 6 in the appendix.

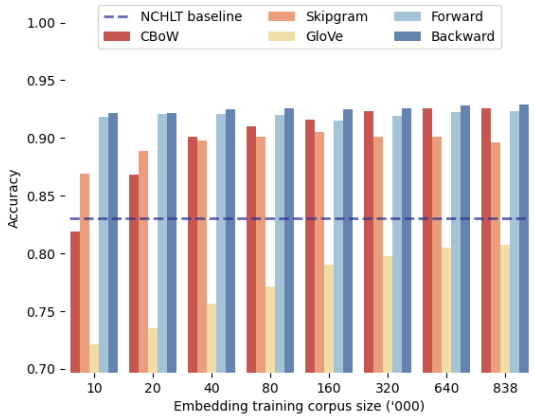

Figure 2: Accuracy of isiZulu POS tagging using different embedding models with increasing data sizes

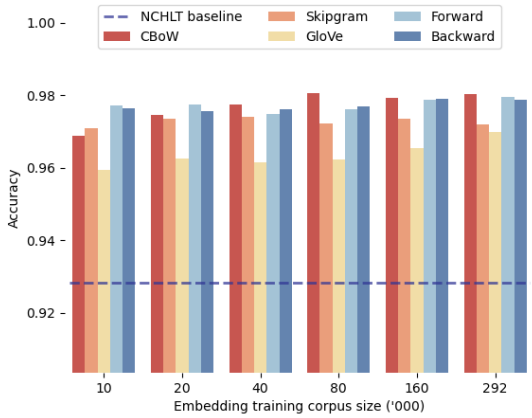

Figure 3: Accuracy of Sesotho sa Leboa POS tagging using different embedding models with increasing data sizes

In the case of isiZulu, the FLAIR Backward model performs best overall, although the FLAIR Forward performs comparably. This differentiation with Afrikaans is likely due to the fact that isiZulu uses prefixation more productively than suffixation, and processing data from the end of the text to the beginning leads to a slightly more informative model. The GloVe model for isiZulu is significantly worse than any of the other models trained in this investigation and is definitely a consequence of data sparsity during training as well as previously unseen words in the tagging task. This problem is less prevalent for the fastText models: Since the n-grams of previously unseen words can still generate a representation, and although not as informative to the task, these "subword" representations obviously have a substantial impact on the quality of the results. The CBoW and GloVe models show the largest error rate reductions across the major POS classes of between 17% and 77%, particularly for Possessives and Adverbs. The FLAIR models on the other hand do not show large improvements for any of the categories, and the improvements are counteracted by regressions in other classes, to which end the results remain relatively stable between the models trained on the smallest and largest corpora.

Even though the FLAIR embeddings perform best for isiZulu, there is very little improvement for these models as the size of the data increases. There are two possible, and related, reasons why this may be the case. Firstly, since most of the affixation in isiZulu is fairly regular, most of the morphological structure of the language may be encoded well with small amounts of data. This is supported by the fact that the most productive word classes (i.e. nouns, verbs, adverbs, possessives, and relatives) are tagged more accurately with a FLAIR model trained on the smallest amount of data, than for any of the other embedding models. The fastText CBoW model does perform similarly on these classes with the largest training set, and may possibly outperform the FLAIR models if more data is made available. The second possible reason is that the annotated POS training data is just too small for further improvements to be possible, and a substantially larger set is required to attain results comparable to those of Afrikaans or Sesotho sa Leboa.

All of the embedding models for Sesotho sa Leboa, with its disjunctive writing style, clearly already perform well with very small data sets. The improvements on the POS tagging task with larger data sets is also not nearly as large as for the other two languages. As with the other languages, the FLAIR models perform the best with very little data, while the fastText CBoW models generate the best overall results with more data. Surprisingly, the fastText Skipgram models do not show much improvement between the smallest and largest data sets, and there does not seem to be an easily identifiable reason for this result. As with both other languages, the GloVe embeddings generally show improvements with each iteration of larger data, and are likely to keep improving if more data were available to be included. For the GloVe and FLAIR models, the improvements in tag classes are much more moderate, between 8% and 46% for the noun, verb, concord, and adjective classes. Both of the

FLAIR models also regress on the adverb class. The fastText models do show more substantial improvements for some of the classes, but for the Skipgram model the error reduction in one class is counteracted by an increase in errors in another class. For example, the noun class errors are reduced by 40%, but the adjective and concordial classes increase their errors by more than 40%, resulting in Accuracies that are very similar to the model trained on the smallest data sets.

Generalizing our findings for the type of embeddings to use with little data, the takeaway is that FLAIR models will produce decent results, especially with very little data. With slightly more data, fastText CBoW embeddings will also perform adequately. GloVe, however, needs large amounts of data to reach enough generalization power to be applied successfully to a morpho-syntactic downstream task.

Our analysis also shows that the influence of linguistic and orthographic differences between languages should not be underestimated. A language such as isiZulu with a complex morphology and large vocabulary (and consequently more data sparseness) will need more data to train representative language models. But a better language model alone is not sufficient. More task related annotated data is also needed to substantially increase the POS accuracy – again an effect of trying to learn from sparse data. It is important to acknowledge the influence of data sparseness in both the learned representations and the actual task to be learned on the final tagging results.

6 Conclusion and Future Work

In this paper, we investigated how the amount of available training data and the linguistic attributes of a language influence the quality of learned embeddings. Our case study consisted of training three different embedding architectures on varying amounts of data, and evaluating the embeddings extrinsically on the downstream task of POS tagging for three linguistically distinct South African languages (Afrikaans, isiZulu and Sesotho sa Leboa).

Our results indicate that under certain conditions even relatively little data can produce useful representations for a specific task. We explicitly show that with very little data (approximately 300,000 tokens) FLAIR embeddings generate representations that perform comparably to any of the other architectures trained on the largest data sets, irrespective of the morphological complexity of the language. The FLAIR models do not generally show the same level of improvements as the other embedding types when larger data sets are available, and in some cases are out-preformed by the fastText CBoW embeddings with the largest available training sets.

The results further reinforce the knowledge that for morphologically complex, conjunctively written languages, substantially more data is needed to achieve good results, not only unannotated text for training the language model, but also annotated training material for the task at hand. Overall we conclude that it is imperative to know the characteristics of the language you are working on to make linguistically informed choices about the amount of data and the type of embeddings to use.

Although these results are encouraging for the relatively simple task of POS tagging, the same may not be true for other, more complex tasks, especially where semantic attributes are of interest. We do however expect that the shortcomings apparent in the fastText Skipgram and GloVe models will remain in under-resourced settings regardless of the task they are applied to. With this in mind there are two areas for future investigation. Firstly, these embedding models should be applied to different tasks that may require different linguistic attributes. Secondly, a comparable experimental design should be applied to transformer models, such as RoBERTa, and to fine-tuning multi-lingual language models (e.g. mBERT, XLM-R) to determine whether similar encouraging results are possible with these more complex architectures.

7 Limitations

There are several limitations of the research reported on in this submission, some of which are explicitly stated in the paper, and others that are expressed in this section as they do not fit well within the discussions of the paper.

The first major limitation of the work relates to the fact that the reported results for the downstream task are not averages of multiple runs, and that there was no hyperparameter tuning performed. Due to the nature of the investigation, i.e. not attempting to achieve state-of-the-art results, and the number of separate runs required to address this limitation, the authors do not expect the results to change such that the conclusions would be significantly affected. For these reasons and the ethical implications of

performing unnecessary runs the authors decided not to perform these additional experiments.

Secondly, even though the submission reports on three linguistically distinct languages, care should still be taken when interpreting and applying these findings to other languages, especially where those languages differ significantly from those described in the paper, such as for instance Dravidian and Sino-Tibetan languages.

Lastly, the results of this submission specifically target the relatively simple POS tagging task, and further investigations on the findings of this paper in more complex NLP tasks is necessary to support these findings.

Acknowledgements

This work was made possible with the financial support of the National Centre for Human Language Technology, an initiative of the South African Department of Sports, Arts and Culture. The authors would also like to thank Martin Puttkammer for feedback during the writing and revision of this work.

References

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, Minnesota. Association for Computational Linguistics.

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

Jesujoba O. Alabi, David I. 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.

Laith Alzubaidi, Jinglan Zhang, Amjad J. Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, J. Santamaría, Mohammed A. Fadhel, Muthana Al-Amidie, and Laith Farhan. 2021. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8(53).

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.

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

Sibonelo Dlamini, Edgar Jembere, Anban Pillay, and Brett van Niekerk. 2021. isiZulu word embeddings. In *2021 Conference on Information Communications Technology and Society (ICTAS)*, pages 121–126. IEEE.

Roald Eiselen and Martin Puttkammer. 2014. Developing text resources for ten South African languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3698–3703, Reykjavik, Iceland. European Language Resources Association (ELRA).

Tanja Gaustad and Martin 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 Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Ridewaan Hanslo. 2021. Evaluation of neural network transformer models for named-entity recognition on low-resourced languages. In *16th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pages 115–119. IEEE.

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.

Nuette Heyns and Etienne Barnard. 2020. Optimising word embeddings for recognised multilingual speech. In *1st Southern African Conference for Artificial Intelligence Research*, pages 102–116. Southern African Conference for Artificial Intelligence Research.

Justin Hocking. 2014. Language identification for South African languages. In *Annual Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*. PRASA.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2023. Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, 82:3713–3744.

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

Louis Jacobus Louwrens and George Poulos. 2006. The status of the word in selected conventional writing systems - the case of disjunctive writing. *Southern African Linguistics and Applied Language Studies*, 24(3):389–401.

Vukosi Marivate, Tshephisho Sefara, Vongani Chabalala, Keamogetswe Makhaya, Tumisho Mokgonyane, Rethabile Mokoena, and Abiodun Modupe. 2020. Investigating an approach for low resource language dataset creation, curation and classification: Setswana and Sepedi. In *Proceedings of the first workshop on Resources for African Indigenous Languages (RAIL)*, pages 15–20, Marseille, France. European Language Resources Association (ELRA).

Cindy McKellar. 2022a. Autshumato monolingual Afrikaans corpus. https://hdl.handle.net/20.500.12185/580. South African Centre for Digital Language Resources (SADiLaR).

Cindy McKellar. 2022b. Autshumato monolingual isiZulu corpus. https://hdl.handle.net/20.500.12185/581. South African Centre for Digital Language Resources (SADiLaR).

Cindy McKellar. 2022c. Autshumato monolingual Sepedi corpus. https://hdl.handle.net/20.500.12185/582. South African Centre for Digital Language Resources (SADiLaR).

Tomas Mikolov, Ken Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations ICLR 2013, Workshop Track Proceedings*, Scottsdale, Arizona, USA.

Tumi Moeng, Sheldon Reay, Aaron Daniels, and Jan Buys. 2022. Canonical and surface morphological segmentation for Nguni languages. In *Artificial Intelligence Research. Second Southern African Conference, SACAIR 2021*, pages 125–139, Durban, South Africa. Springer.

Jeffrey Pennington, Richard Socher, and Christopher 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. Association for Computational Linguistics.

Martin Puttkammer, Roald Eiselen, Justin Hocking, and Frederik Koen. 2018. NLP web services for resource-scarce languages. In *Proceedings of ACL 2018, System Demonstrations*, pages 43–49, Melbourne, Australia. Association for Computational Linguistics.

Martin Puttkammer and Tanja Gaustad. 2021. Linguistically enriched corpora for conjunctively written South African languages. https://hdl.handle.net/20.500.12185/546. South African Centre for Digital Language Resources (SADiLaR).

Martin Puttkammer, Martin Schlemmer, and Ruan Bekker. 2014a. NCHLT Afrikaans annotated text corpora. https://hdl.handle.net/20.500.12185/296. South African Centre for Digital Language Resources (SADiLaR).

Martin Puttkammer, Martin Schlemmer, and Ruan Bekker. 2014b. NCHLT Sepedi annotated text corpora. https://hdl.handle.net/20.500.12185/325. South African Centre for Digital Language Resources (SADiLaR).

Martin Puttkammer, Martin Schlemmer, Wikus Pienaar, and Ruan Bekker. 2014c. NCHLT Afrikaans text corpora. https://hdl.handle.net/20.500.12185/293. South African Centre for Digital Language Resources (SADiLaR).

Martin Puttkammer, Martin Schlemmer, Wikus Pienaar, and Ruan Bekker. 2014d. NCHLT isiZulu text corpora. https://hdl.handle.net/20.500.12185/321. South African Centre for Digital Language Resources (SADiLaR).

Martin Puttkammer, Martin Schlemmer, Wikus Pienaar, and Ruan Bekker. 2014e. NCHLT Sepedi text corpora. https://hdl.handle.net/20.500.12185/330. South African Centre for Digital Language Resources (SADiLaR).

Sello Ralethe. 2020. Adaptation of deep bidirectional transformers for Afrikaans language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2475–2478, Marseille, France. European Language Resources Association.

Mark van der Velde, Koen Bostoen, Derek Nurse, and Gérard Philippson, editors. 2022. *The Bantu Languages*, 2nd edition. Routledge.

A Appendix

A.1 Full Results Table

Paragraph count	Token Count	Vocabulary	Token:Vocab ratio	fastText CBoW	fastText Skipgram	GloVe	FLAIR Forward	FLAIR Backward
Afrikaans								
10,000	316,704	13,152	0.0415	0.9287	0.9597	0.9385	0.9755	0.9680
20,000	628,359	21,032	0.0335	0.9601	0.9652	0.9462	0.9726	0.9710
40,000	1,253,597	33,657	0.0268	0.9719	0.9688	0.9424	0.9758	0.9757
80,000	2,527,103	52,844	0.0209	0.9769	0.9717	0.9518	0.9777	0.9775
160,000	5,067,551	82,314	0.0162	0.9798	0.9693	0.9513	0.9772	0.9746
320,000	10,172,939	128,172	0.0126	0.9808	0.9710	0.9548	0.9798	0.9760
640,000	20,303,831	199,335	0.0098	0.9806	0.9710	0.9566	0.9820	0.9787
1,280,000	40,610,635	311,719	0.0077	0.9834	0.9738	0.9578	0.9803	0.9794
isiZulu								
10,000	193,814	16,207	0.0836	0.8188	0.8694	0.7215	0.9181	0.9219
20,000	394,523	29,120	0.0738	0.8686	0.8892	0.7354	0.9205	0.9217
40,000	783,393	50,153	0.0640	0.9009	0.8977	0.7562	0.9209	0.9249
80,000	1,561,536	50,914	0.0326	0.9100	0.9015	0.7711	0.9197	0.9255
160,000	3,115,721	143,262	0.0460	0.9160	0.9051	0.7903	0.9154	0.9251
320,000	6,232,015	240,454	0.0386	0.9233	0.9011	0.7980	0.9189	0.9261
640,000	12,438,302	401,423	0.0323	0.9257	0.9015	0.8054	0.9227	0.9280
838,000	16,271,123	488,822	0.0300	0.9261	0.8965	0.8075	0.9233	0.9290
Sesotho sa Leboa								
10,000	302,923	10,464	0.0345	0.9689	0.9708	0.9594	0.9771	0.9763
20,000	605,780	16,197	0.0267	0.9745	0.9736	0.9624	0.9773	0.9756
40,000	1,214 472	25,243	0.0208	0.9774	0.9739	0.9616	0.9747	0.9761
80,000	2,435,686	38,401	0.0158	0.9806	0.9721	0.9623	0.9760	0.9770
160,000	4,878,117	58,097	0.0119	0.9792	0.9736	0.9655	0.9788	0.9791
292,600	8,909,133	80,919	0.0091	0.9802	0.9719	0.9698	0.9794	0.9788

Table 3: Full results for Afrikaans, Sesotho sa Leboa and isiZulu on all types of embeddings with different input data sizes using a reduced POS tagset

A.2 POS Linguistic Analysis Tables

	# POS errors	N	ADJ	V	ADV	Other
fastText CBoW	in 10,000 paragr.	109	68	60	55	123
	in 1,280,000 paragr.	22	11	3	11	49
	% Improvement	79.82%	83.82%	95.00%	80.00%	60.16%
fastText Skipgram	in 10,000 paragr.	72	30	18	31	83
	in 1,280,000 paragr.	23	19	11	15	84
	% Improvement	68.06%	36.67%	38.89%	51.61%	-1.20%
GloVe	in 10,000 paragr.	104	59	37	35	123
	in 1,280,000 paragr.	66	36	31	19	92
	% Improvement	36.54%	38.98%	16.22%	45.71%	25.20%
FLAIR Forward	in 10,000 paragr.	22	23	25	15	58
	in 1,280,000 paragr.	25	13	7	20	50
	% Improvement	-13.64%	43.48%	72.00%	-33.33%	13.79%
FLAIR Backward	in 10,000 paragr.	53	24	13	31	65
	in 1,280,000 paragr.	19	18	5	16	61
	% Improvement	64.15%	25.00%	61.54%	48.39%	6.15%

Table 4: Linguistic error analysis for Afrikaans POS

	# POS errors	N	POSS	REL	ADV	V	ADJ	Other
fastText CBoW	in 10,000 paragr.	179	158	150	145	111	23	132
	in 838,000 paragr.	106	36	51	38	72	11	52
	% Improvement	40.78%	77.22%	66.00%	73.79%	35.14%	52.17%	60.61%
fastText Skipgram	in 10,000 paragr.	140	82	122	80	104	18	101
	in 838,000 paragr.	120	60	82	51	98	23	79
	% Improvement	14.29%	26.83%	32.79%	36.25%	5.77%	-27.78%	21.78%
GloVe	in 10,000 paragr.	272	217	284	222	229	29	127
	in 838,000 paragr.	184	179	196	132	146	20	97
	% Improvement	32.35%	17.51%	30.99%	40.54%	36.24%	31.03%	23.62%
FLAIR Forward	in 10,000 paragr.	113	33	56	35	86	15	68
	in 838,000 paragr.	102	41	49	36	80	12	60
	% Improvement	9.73%	-24.24%	12.50%	-2.86%	6.98%	20.00%	11.76%
FLAIR Backward	in 10,000 paragr.	123	26	47	29	83	11	68
	in 838,000 paragr.	93	37	45	35	60	9	73
	% Improvement	24.39%	-42.31%	4.26%	-20.69%	27.71%	18.18%	-7.35%

Table 5: Linguistic error analysis for isiZulu POS

	# POS errors	N	V	CONC	ADV	ADJ	Other
fastText CBoW	in 10,000 paragr.	48	31	17	13	10	90
	in 292,600 paragr.	12	25	17	12	4	63
	% Improvement	75.00%	19.35%	0.00%	7.69%	60.00%	30.00%
fastText Skipgram	in 10,000 paragr.	32	27	21	15	9	93
	in 292,600 paragr.	19	29	30	13	13	85
	% Improvement	40.63%	-7.41%	-42.86%	13.33%	-44.44%	8.60%
GloVe	in 10,000 paragr.	53	34	29	23	14	123
	in 292,600 paragr.	43	30	23	21	9	80
	% Improvement	18.87%	11.76%	20.69%	8.70%	35.71%	34.96%
FLAIR Forward	in 10,000 paragr.	23	20	17	12	6	81
	in 292,600 paragr.	19	16	13	13	6	71
	% Improvement	17.39%	20.00%	23.53%	-8.33%	0.00%	12.35%
FLAIR Backward	in 10,000 paragr.	25	30	21	12	15	57
	in 292,600 paragr.	21	25	17	14	8	59
	% Improvement	16.00%	16.67%	19.05%	-16.67%	46.67%	-3.51%

Table 6: Linguistic error analysis for Sesotho sa Leboa POS