Recipe for Zero-shot POS Tagging: Is It Useful in Realistic Scenarios?

Zeno Vandenbulcke, Lukas Vermeire, Miryam de Lhoneux
Department of Computer Science,
KU Leuven
{zeno.vandenbulcke, lukas.vermeire}@student.kuleuven.be
miryam.delhoneux@kuleuven.be

Abstract

POS tagging plays a fundamental role in numerous applications. While POS taggers are highly accurate in well-resourced settings, they lag behind in cases of limited or missing training data. This paper focuses on POS tagging for languages with limited data. We seek to identify the characteristics of datasets that make them favourable for training POS tagging models without using any labelled training data from the target language. This is a zero-shot approach. We compare the accuracies of a multilingual large language model (mBERT) fine-tuned on one or more languages related to the target language. Additionally, we compare these results with models trained directly on the target language itself. We do this for three target low-resource languages. Our research highlights the importance of accurate dataset selection for effective zero-shot POS tagging. Particularly, a strong linguistic relationship and high-quality datasets ensure optimal results. For extremely low-resource languages, zero-shot models prove to be a viable option.

1 Introduction

In recent years, a lot of progress has been made in Natural Language Processing (NLP). However, certain fundamental technologies such as *Part-of-Speech* (POS) *tagging* or dependency parsing are still only available for a small part of the world's languages. This is mostly for languages with significant amounts of available data. For languages with limited or no available data (*low-resource* languages), these technologies are highly inaccurate or sometimes even nonexistent (Joshi et al., 2020). Advancements in multilingual language models have shown impressive cross-lingual transfer abilities (Wu and Dredze, 2019). In this paper, we build on these advancements to explore zero-shot POS tagging for low-resource languages.

We investigate two questions:

RQ1 What are the essential characteristics of datasets for effectively fine-tuning *zero-shot POS tagging models* for low-resource languages?

RQ2 Are zero-shot models useful in realistic low-resource settings when compared to models fine-tuned with target language data?

We explore these questions by fine-tuning a multilingual pretrained language model for zero-shot POS tagging, using related languages (which we call *support languages*) to the target language. We start by fine-tuning POS tagging models for Afrikaans, using Dutch, German, and English as support languages. We test the models on Afrikaans and compare the results in an attempt to identify the characteristics of the datasets that affect the performance of the models. We then experiment with two additional target languages: Faroese (supported by Icelandic, Danish, Norwegian and Swedish) and Upper Sorbian (supported by Czech, Polish and Slovak). We aim to determine whether our findings for Afrikaans also apply to these languages.

In relation to **RQ1**, we find that when multiple supporting languages are available, high-quality datasets (Kulmizev and Nivre, 2023) that are closely related to the target language result in better performance. Using the most closely related language leads to consistently better accuracy, especially with a limited number of training sentences. For an optimal training dataset size, using between 100 and 5000 sentences helps to avoid *under-* or *overfitting*.

Regarding **RQ2**, we find that zero-shot POS tagging models can certainly be a viable option for low-resource languages. Nevertheless, models trained on annotated data from the low-resource target language itself remain superior, similarly to pre-

vious findings in the literature (Meechan-Maddon and Nivre, 2019).

As we will discuss in the conclusion, these findings can be translated into concrete guidelines for different scenarios.

This paper starts with background information on low-resource zero-shot POS tagging. We then discuss the technical components and methodology used. Finally, we present the results and attempt to answer the previously mentioned research questions.

2 Background

Part-of-Speech tagging with Universal Dependencies

Part-of-Speech tagging is an essential application within NLP. It is used in machine translation, word meaning disambiguation, parsing, among other applications (Chiche and Yitagesu, 2022). It is a highly researched task for which there are many annotated datasets, in many languages. Universal Dependencies (UD; de Marneffe et al., 2021) is a collection of treebanks which include POS tags for hundreds of languages, with a unified POS tagging scheme. UD distinguishes between 17 different tags, called the Universal POS tags (Zeman et al., 2023). This unified annotation scheme allows the development and comparative evaluation of POS taggers across the languages included in UD. In addition, UD makes it possible to build multilingual POS taggers and dependency parsers which can deal with multiple languages within a single model (e.g. Kondratyuk and Straka, 2019), and this enables cross-lingual transfer. A limitation of UD is that the annotation quality varies considerably across treebanks (Kulmizev and Nivre, 2023). This may negatively impact cross-lingual transfer, a question we investigate in this paper.

Zero-shot learning

A zero-shot model is a learning model that can perform a task without having seen examples or data of that task during the fine-tuning phase. In the context of this paper, a zero-shot POS tagging model refers to a model that is trained to POS tag sentences in one or more support training languages. The performance of this model is then evaluated on data from a different language, also known as the target language. Importantly, the model does not encounter the data of this target language during the fine-tuning phase (although it may have seen target language data during pre-training). One could compare it to a student who has an Afrikaans exam scheduled but is only allowed to prepare by studying, for instance, Dutch and German. This approach is useful in NLP to fill gaps in the availability or correctness of data for a target language. Thus, for *extremely low-resource* languages, or especially languages for which no annotated data is available, POS tagging could be performed using a zero-shot model trained on related languages.

We use the zero-shot strategy here because we are interested in scenarios where no data is available for certain languages. This provides a better understanding of the situation because all models developed for this purpose would, by definition, be zero-shot.

Low-resource languages

While there is no general definition of the term low-resource, researchers have attempted to define it (Joshi et al., 2020). However, this definition has not yet been widely adopted. We consider a dataset, and thus a language, to be low-resource if it contains fewer than 50,000 tokens in UD. This mainly concerns indigenous languages, but can also include languages that are more broadly used. In UD, a token is a syntactic word used for analysis, which might differ from orthographic or phonological words (de Marneffe et al., 2021).

3 Methodology

Through our experiments, we hope to gain insights into the characteristics of datasets that contribute to the performance of zero-shot POS tagging models for low-resource languages.

Our experiments focus on fine-tuning POS tagging models based on mBERT, a *large language model* (LLM; Devlin et al., 2019). This process involves fine-tuning the model on annotated treebanks to enable it to perform POS tagging. The model's performance is then evaluated on a test dataset. We specifically chose mBERT because of its multilingual capabilities. Additionally, mBERT has shown good results in zero-shot scenarios (Pires et al., 2019).

The fine-tuning of mBERT is done using the tool MaChAmp (van der Goot et al., 2021). MaChAmp is a user-friendly tool that enables the fine-tuning of LLMs on various NLP tasks from diverse datasets and languages. The latter functionality in particular is valuable for our research. This enables us

Language	Treebank	# Sents	Rank
Afrikaans	AfriBooms	1.5k	46
Dutch	Alpino	12k	29
German	GSD	15k	19
English	EWT	16k	23
Faroese	FarPaHC	1.6k	62
Icelandic	Modern	3.5k	34
Danish	DDT	5.5k	53
Norwegian	Bokmaal	20k	14
Swedish	Talbanken	6.0k	57
Upper Sorbian	UFAL	0.6k	
Czech	FicTree	12.7k	24
Polish	LFG	17.2k	20
Slovak	SNK	10.6k	42

Table 1: Languages with corresponding treebanks (UD v2.13) ranked according to their relatedness to the target language, including the number of sentences (**# Sents**) in each treebank and their respective quality rankings (**Rank**; Kulmizev and Nivre, 2023).

to jointly train POS tagging models on multiple languages.

Languages and treebanks

We select three different groups from Universal Dependencies v2.13 (Zeman et al., 2023). These groups are shown in Table 1. For each group, we select one low-resource language and a few related languages for which larger treebanks are available.

Our focus is primarily on Afrikaans. For this, we use the AfriBooms treebank (Augustinus et al., 2016). The supporting languages are Dutch, German, and English. To verify our observations of Afrikaans, we use two other clusters. One cluster includes Scandinavian languages, with Faroese as the low-resource language. The related languages are Icelandic, Danish, Norwegian, and Swedish. We also use a West Slavic cluster, with Upper Sorbian as the low-resource language, supported by Czech, Polish, and Slovak.

Characteristics

To answer **RQ1** (*What are the essential characteristics of datasets for effectively fine-tuning zero-shot POS tagging models for low-resource languages?*), we consider two main characteristics: the linguistic relatedness between languages and the quality of the treebank. The relevance of the linguistic relatedness is already evident from previous work (see section 5), but the treebank quality has not been taken into consideration before in spite of being a

clear differentiating factor between UD treebanks (see section 2 & 5). Firstly, we take a look at the linguistic relatedness. The support languages are consistently chosen to be of the same genus as the target language. This results in an intrinsic relatedness. In the first cluster, Dutch shows the closest relatedness to Afrikaans (van Zaanen et al., 2014). In the Scandinavian cluster, Faroese is most closely related to Icelandic (Snæbjarnarson et al., 2023). In the West Slavic cluster, Upper Sorbian is most closely related to Czech (Howson, 2017), followed by Polish.

For the second characteristic of the treebanks, we rely on a ranking developed by Kulmizev and Nivre (2023). This ranking is based on three criteria: how difficult or easy the treebanks are to parse, how much information they contain that is actually usable by a parser, and how sample efficient they are. We report the rank of the languages considered in this work in Table 1. The supporting languages for Afrikaans, for example, can be ranked as follows: German > English > Dutch. Throughout this paper, we refer to this as the 'quality' of a treebank.

As an additional characteristic, we use the size of the dataset. With this, we investigate whether overfitting might occur and determine the optimal number of sentences a model should use.

Experimental setup

We train zero-shot models for each of the three clusters. We fine-tune these models in several ways. First, we fine-tune separate zero-shot models for each distinct supporting language. Then, we fine-tune models based on different combinations of these languages. We repeat this process for the different clusters.

We conduct learning curve experiments to display the performance of the models as they are fine-tuned on increasingly more data. The fine-tuning starts with five sentences and gradually reaches the maximum available number of sentences from the treebanks. For each cluster and each model within the cluster, we repeat this process three times. During each iteration, the sentences of the training dataset of each language are shuffled. This allows for random selection of sentences, which is crucial for ensuring generalisability.

We determine the accuracy of each model using the F1-score, a common metric for assessing the performance of classifiers such as POS tagging

models (Jurafsky and Martin, 2009). The results of these experiments and the associated observations are discussed in the following section.

Finally, we also look at monolingual non-zero-shot models that are fine-tuned on the respective target languages themselves. This means that we fine-tune three distinct models, each using only Afrikaans, Faroese, or Upper Sorbian. This can then provide us with an answer to our second question (**RQ2**): *Are zero-shot models useful in realistic low-resource settings compared to monolingual non-zero-shot models in terms of accuracy?*

4 Results and discussion

First, we analyse the learning curves of all clusters to identify which characteristics of datasets seem to have the most impact on the results and could therefore be more suitable for a zero-shot model. Then, we evaluate the effective usability of our models and results to determine whether the zero-shot approach is effective.

4.1 RQ1: Dataset characteristics

Afrikaans

First and foremost, we take a look at the accuracies of the models that have Afrikaans as the target language and that were trained on one supporting language. This can be seen in Figure 1. It can be clearly seen that when fine-tuning the zero-shot model using a language more closely related to the target language, the initial accuracies are higher. Initial accuracies are the accuracies that occur with a smaller number of training sentences. A one-to-one correspondence can be seen between how closely related the training language is to the target language and how accurate the corresponding model is for a very small subset of the dataset.

As more sentences are added to the training dataset of the models, it can be seen that the accuracies of the three models converge. This is most likely due to the fact that the supporting languages are all Germanic languages and there is a high similarity to the target language. However, it can be seen that the Dutch model performs better overall and also achieves the best accuracy of the three models. The model that performs the worst overall is the English model. This is not surprising, as English is least closely related to Afrikaans among the three supporting languages, and its quality falls in the middle range.

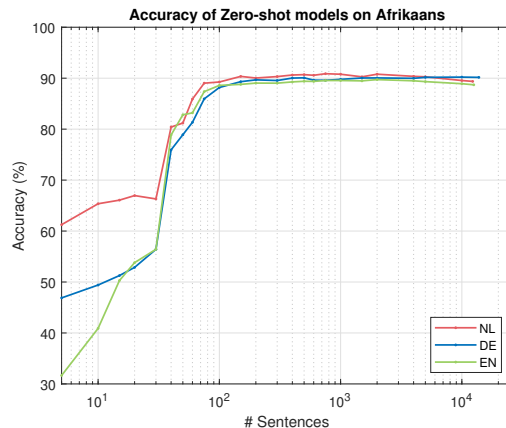Next, we add the models that were trained with

Figure 1: Accuracy of fine-tuned models on Afrikaans, represented through learning curves.
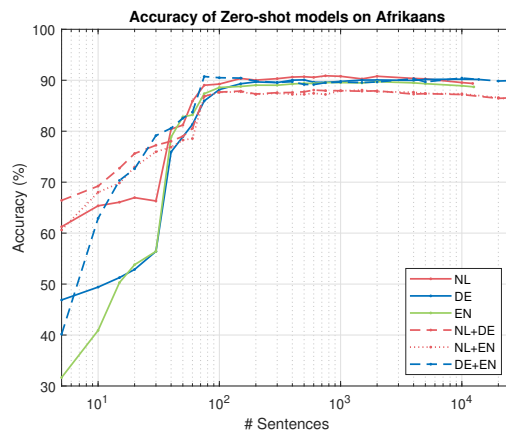

Figure 2: Accuracy of fine-tuned models on Afrikaans, represented through learning curves.

combinations of supporting languages and analyse these separately. This can be seen in Figure 2. It is immediately noticeable that the added models using the most closely related language initially achieve higher accuracies. The model with the highest initial accuracy here is the Dutch-German model. This seems logical since Dutch and German follow each other as most related to Afrikaans. In second place is the Dutch-English model, again because Dutch is in the training dataset and has a significant influence. The worst-performing model is the German-English model. This is not surprising since German and English are the two languages that are least related to Afrikaans.

As the size of the training datasets of the models increases, there is a greater shift between the accuracies of the different models. The model that performs best overall is the German-English model. This is unexpected, given that Dutch is the closest language to Afrikaans (Heeringa et al., 2015). One

140

explanation for this might be that when a model is trained on more than one language, the quality of the datasets becomes more important than the language relatedness. The Dutch-German model performs slightly better across the board than the Dutch-English model, which suggests that relatedness still plays a role. When we look at all the models in Figure 2 globally, we can make several observations:

If a model has a language in the training dataset that is more closely related to the target language, the model has a higher initial accuracy. When multiple closely related languages are used, such as the Dutch-German model, this accuracy increases even further.

The most performant model is one trained on multiple languages. In the case of Afrikaans, this is the German-English model. This can be attributed to the quality of the datasets used. This model quickly achieves better results and consistently maintains a high accuracy.

There seems to be a plateau at which all models achieve accuracies that neither increase nor decrease, usually between 100 and 5000 sentences. What is also notable is that within this interval, the Dutch model generally performs the best, while the German-English model achieved the highest peak accuracy prior to this interval.

Faroese

Secondly, we take a look at all the models we have fine-tuned that have Faroese as the target language. This can be seen in Figure 3. What stands out immediately is that all models that contain the most related language - Icelandic - consistently achieve the best results. This results in two distinct groups: one group with models containing Icelandic, and a second group with the other models. Just as with Afrikaans, it can also be seen here that the models that contain a more related language achieve a higher initial accuracy.

Regarding dataset quality, the model fine-tuned using the highest-quality datasets (Icelandic-Norwegian) ranks among the best performing models, while the model trained on the lowest-quality datasets (Danish-Swedish) ranks among the worst. Interestingly, the model that achieves the overall peak accuracy is the Icelandic-Danish model, again highlighting the importance of language relatedness, not only for lower training sizes, but throughout the entire process.
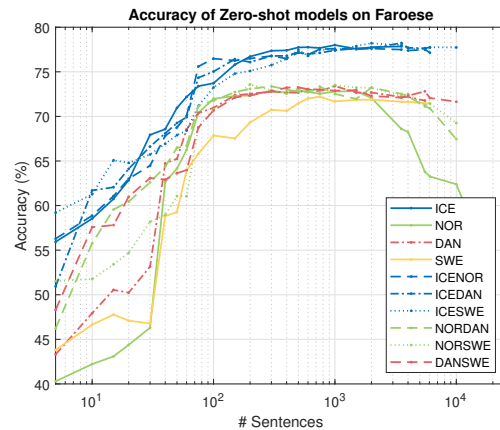
Here, a plateau between about 100 and 5000

Figure 3: Accuracy of fine-tuned models on Faroese, represented through learning curves.

sentences can also be clearly seen. This is especially noticeable in the accuracies of the models that do not include Icelandic. The curves start with an increase, followed by a stagnation and then a decline in accuracy. This is particularly noticeable in the models trained on Swedish and Norwegian. In the case of Swedish, this is not surprising as the treebank is of relatively low quality and not very closely related to Faroese. Norwegian, on the other hand, is closely related and of high quality, which makes this trend all the more striking.

Some further observations also become clear here. As more sentences are added to the training dataset of the models, the accuracies of the models converge. This again highlights the idea of intrinsic relatedness between languages within the same language family. Furthermore, it can be seen that the best performing model is one that is fine-tuned on multiple supporting languages, although the model solely fine-tuned on Icelandic is also among the better performing models.

Upper Sorbian

Lastly, we look at the accuracies of models whose target language is Upper Sorbian. This can be seen in Figure 4. Here, largely the same trends are seen as in the two previous clusters. The initial accuracies of models trained with the most closely related language (Czech) are higher, although Polish takes the lead when the model is trained on a single supporting language. The statement holds true for models trained on multiple supporting languages: the greater the relatedness and the higher the quality, the better the model performs. In addition, the trend between 100 and 5000 sentences can be seen again here, although it is slightly less pronounced.
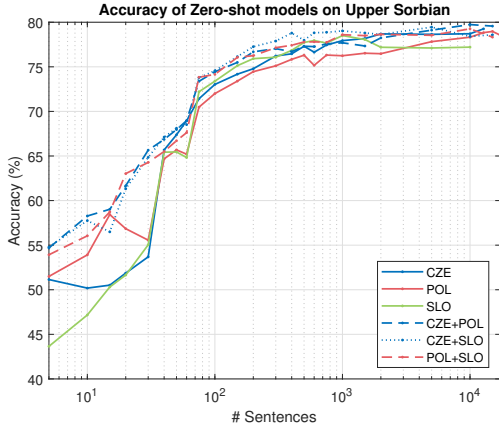
141

Figure 4: Accuracy of fine-tuned models on Upper Sorbian, represented through learning curves.

Language	Zero-shot	Non-zero-shot
Afrikaans	90.9	**98.5**
Faroese	78.2	**97.3**
Upper Sorbian	**79.7**	70.7

Table 2: Peak accuracies of zero-shot & non-zero-shot models.

4.2 RQ2: Usefulness of zero-shot models in realistic settings

In order to answer **RQ2**, we compare our zero-shot models with those that have been fine-tuned directly on the respective target languages (non-zero-shot).

Comparing zero-shot and non-zero-shot performance

Firstly, we examine the practical relevance of zero-shot models in the context of low-resource languages. The peak accuracies for both our zero-shot and non-zero-shot models can be seen in Table 2. For Afrikaans and Faroese, we observe that the non-zero-shot models outperform their zero-shot counterparts, with the Faroese model showing a nearly 20 percentage point improvement over the zero-shot model. This suggests that, given enough training data, fine-tuning on the target language can lead to substantially better results, as also discussed by Meechan-Maddon and Nivre (2019).

However, when we take a look at Upper Sorbian, an *extremely low-resource* language with only 23 training sentences, we observe a different trend. Here, the zero-shot model actually surpasses the non-zero-shot model by 9 percentage points, achieving a peak accuracy of 79.7% compared to the non-zero-shot model's 70.7%. This result sug-
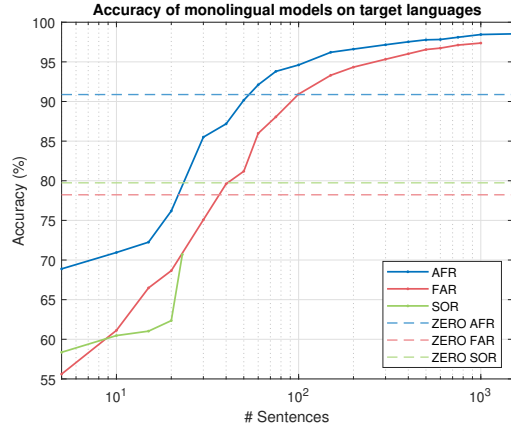

Figure 5: Accuracy of models fine-tuned on the target languages Afrikaans, Faroese, and Upper Sorbian, represented through learning curves accompanied by the peak accuracies of the respective zero-shot models.

gests that our zero-shot models are certainly a viable option for extremely low-resource languages or languages for which no data is available.

Upper Sorbian is not a unique case; we counted 82 languages in UD v2.14 that have fewer than 23 training sentences. This widespread scarcity highlights the importance of zero-shot models in real-world applications where data is often hard to come by.

Amount of annotated data needed to surpass zero-shot performance

Secondly, we take a look at how much annotated data is necessary to improve upon zero-shot performance through monolingual fine-tuning (non-zero-shot). In Figure 5, the accuracies of the POS tagging models are shown when they are fine-tuned on the respective target languages, alongside the peak accuracies of the respective zero-shot models. Using these lines, and the intersection they make with the learning curves of the non-zero-shot models, we can estimate when a non-zero-shot model becomes strictly better than a zero-shot model for the same target language.

For Afrikaans, the intersection occurs between 50 and 60 training sentences, indicating that at least this amount is necessary for the non-zero-shot model to outperform the zero-shot model. Similarly, for Faroese, the intersection point is around 40 sentences, suggesting a slightly lower data requirement to achieve better performance. Again, the Upper Sorbian models are slightly different. The learning curve for the non-zero-shot model does not intersect with the peak accuracy of

the zero-shot model, simply because there is not enough targeted data available.

Meechan-Maddon and Nivre (2019) made similar observations in a similar context. They compared dependency parsers trained on treebanks from several support languages (akin to our zero-shot setting) with those trained solely on target language data (similar to our non-zero-shot setting). They found that the non-zero-shot models required between 100 and 200 sentences to reach the performance of the zero-shot models. This is higher than what we found, but can likely be explained by the fact that they do not make any use of pretraining and use a BiLSTM instead of a transformer. We use mBERT, which has a greater number of parameters than a BiLSTM and which already has acquired cross-lingual transfer capabilities by virtue of being trained on multilingual data. This may reduce the need for annotated target language data. This should be verified in future research, however.

5   Related work

Dealing with the limited availability of training data for low-resource languages is an active area of research within NLP. Thanks to the UD treebanks, a large collection of data in numerous languages with varying data sizes, POS tagging and dependency parsing have become highly researched topics within this context

Closest to our work, de Vries et al. (2021) investigated zero-shot transfer for two target languages: Gronings and West Frisian. They also fine-tuned mBERT on related languages, as well as monolingual language models in related languages. They found the latter to be superior to the former. Relatedly, de Vries et al. (2022) did an extensive evaluation of zero-shot POS tagging across 105 target languages. They fine-tuned mBERT using 65 different support languages, testing all possible combinations of support and target languages, with one support language used each time. They found that related languages are generally the best support languages.

Our work is complementary to these by considering a number of target languages that is in between these two extremes (2 versus 105). It allows a targeted evaluation, looking at learning curves and trying multiple support languages in different combinations, while still providing results that generalize to more than two closely related languages. We confirm that, among related languages, the ones

that are the most closely related to the target language are the best support languages. This finding is consistent with many other earlier works in POS tagging and dependency parsing using different types of models (Smith et al., 2018; Pires et al., 2019; Lauscher et al., 2020).

Our learning curve experiments take inspiration from earlier work in dependency parsing by Meechan-Maddon and Nivre (2019). They investigated zero- and few-shot learning of multilingual parsers to find out how much can be gained from cross-lingual transfer versus annotating target language data. They use a BiLSTM parser trained only on treebank data, in multiple languages, including and excluding target language data. Their results showed that the zero-shot approach is inferior to the other approaches, provided at least 200 training sentences are available from the target language. We confirm this finding in the context of fine-tuning a multilingual transformer model, although we find that fewer training sentences are necessary in this context.

Finally, a dataset property which has not yet been investigated in the context of cross-lingual transfer (to our knowledge) is data quality. Kulmizev and Nivre (2023) thoroughly evaluated the quality of UD treebanks using three different metrics and found that the quality varies considerably across treebanks. They found some treebanks to perform consistently low across metrics, making them practically unusable. This raises the question of how this quality impacts results in cross-lingual transfer: a low-quality treebank may be too noisy to use for cross-lingual transfer. We investigated this question and found a subtle link between the quality of the UD Treebanks and the peak accuracies of the corresponding zero-shot models. Of course, more research is needed to confirm this by investigating a larger set of treebanks.

6   Conclusion

Initially, we can conclude that developing zero-shot POS tagging models is a viable option for low-resource languages. Nevertheless, using the low-resource dataset of a specific language remains superior for constructing a POS tagging model for that language, similar to what Meechan-Maddon and Nivre (2019) found in the context of dependency parsing. If the amount of data for a language is so scarce and/or a zero-shot model is still desired, the following guidelines can be followed:

One language can be used as a support language. In this case, always use the language that is most closely related to the target language. This generally gives better accuracies with a low number of training sentences. Even with larger numbers of training sentences, these models tend to perform well. The quality plays a lesser role here.

Multiple support languages can be used. In this case, use as many languages as possible that are closely related to the target language and are of high quality. High relatedness gives the best results with a limited number of training sentences, and high quality generally gives the best results with higher numbers of training sentences.

What is the most suitable number of training sentences? If enough data is available from the support languages, preferably use a training number of 100 to 5000 sentences. Below 100 sentences, the models are often 'underfitted'. Above 5000 sentences, the models can overfit and the accuracies may decrease.

7 Limitations

This work considered only three target languages, each paired with three to four related source languages, selected somewhat arbitrarily. While this restricted number allowed in-depth analysis, our findings need to be verified using more languages with various degrees of relatedness. To keep the number of languages manageable while ensuring generalizability of the results, a sample of typologically diverse languages could be selected using the recently proposed framework by Ploeger et al. (2024).

Additionally, we relied on the linguistic literature to describe the degree of relatedness between the languages considered. It would be informative to quantitatively define language distances, as done by Ploeger et al. (2024).

Finally, we only evaluated one pretrained language model, mBERT. Our findings should be verified with other models.

Acknowledgments

References

Liesbeth Augustinus, Peter Dirix, Daniel van Niekerk, Ineke Schuurman, Vincent Vandeghinste, Frank Van Eynde, and Gerhard van Huyssteen. 2016. AfriBooms: An online treebank for Afrikaans. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 677–682, Portorož, Slovenia. European Language Resources Association (ELRA).

Alebachew Chiche and Betselot Yitagesu. 2022. Part of speech tagging: a systematic review of deep learning and machine learning approaches. *Journal of Big Data*, 9(1):10.

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

Wietse de Vries, Martijn Bartelds, Malvina Nissim, and Martijn Wieling. 2021. Adapting monolingual models: Data can be scarce when language similarity is high. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4901–4907, Online. Association for Computational Linguistics.

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

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.

Wilbert Heeringa, Febe De Wet, and Gerhard B. van Huyssteen. 2015. Afrikaans and dutch as closely-related languages : a comparison to west germanic languages and dutch dialects. *Stellenbosch Papers in Linguistics Plus*, 47(1):1–18.

Phil Howson. 2017. Upper sorbian. *Journal of the International Phonetic Association*, 47(3):359367.

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.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc., USA.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Artur Kulmizev and Joakim Nivre. 2023. Investigating UD treebanks via dataset difficulty measures. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1076–1089, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Ailsa Meechan-Maddon and Joakim Nivre. 2019. How to parse low-resource languages: Cross-lingual parsing, target language annotation, or both? In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 112–120, Paris, France. Association for Computational Linguistics.

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

Esther Ploeger, Wessel Poelman, Andreas Holck Høeg-Petersen, Anders Schlichtkrull, Miryam de Lhoneux, and Johannes Bjerva. 2024. A principled framework for evaluating on typologically diverse languages. *Preprint*, arXiv:2407.05022.

Aaron Smith, Bernd Bohnet, Miryam de Lhoneux, Joakim Nivre, Yan Shao, and Sara Stymne. 2018. 82 treebanks, 34 models: Universal Dependency parsing with multi-treebank models. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Vésteinn Snæbjarnarson, Annika Simonsen, Goran Glavaš, and Ivan Vulić. 2023. Transfer to a low-resource language via close relatives: The case study on Faroese. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 728–737, Tórshavn, Faroe Islands. University of Tartu Library.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Menno van Zaanen, Gerhard van Huyssteen, Suzanne Aussems, Chris Emmery, and Roald Eiselen. 2014. The development of Dutch and Afrikaans language resources for compound boundary analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1056–1062, Reykjavik, Iceland. European Language Resources Association (ELRA).

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

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Salih Furkan Akkurt, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Erik Andersen, Lene Antonsen, Tatsuya Aoyama, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranzabe, Bilge Nas Arıcan, Ḥórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Katla Ásgeirsdóttir, Deniz Baran Aslan, Cengiz Asmazoğlu, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Mariana Avelãs, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Shabnam Behzad, Juan Belieni, Kepa Bengoetxea, brahim Benli, Yifat Ben Moshe, Gözde Berk, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Kristín Bjarnadóttir, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, António Branco, Kristina Brokaitė, Aljoscha Burchardt, Marisa Campos, Marie Candito, Bernard Caron, Gauthier Caron, Catarina Carvalheiro, Rita Carvalho, Lauren Cassidy, Maria Clara Castro, Sérgio Castro, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Liyanage Chamila, Shweta Chauhan, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Claudia Corbetta, Daniela Corbetta, Francisco Costa, Marine Courtin, Benoît Crabbé, Mihaela Cristescu, Vladimir Cvetkoski, Ingerid Løyning Dale, Philemon Daniel, Elizabeth Davidson, Leonel Figueiredo de Alencar, Mathieu Dehouck, Martina de Laurentiis, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza,

Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Adrian Doyle, Timothy Dozat, Kira Droganova, Magali Sanches Duran, Puneet Dwivedi, Christian Ebert, Hanne Eckhoff, Masaki Eguchi, Sandra Eiche, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Farah Essaidi, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Federica Favero, Jannatul Ferdaousi, Marília Fernanda, Hector Fernandez Alcalde, Amal Fethi, Jennifer Foster, Theodorus Fransen, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Federica Gamba, Marcos Garcia, Moa Gärdenfors, Fabrício Ferraz Gerardi, Kim Gerdes, Luke Gessler, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Kirian Guiller, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Takahiro Harada, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Yidi Huang, Marivel Huerta Mendez, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Artan Islamaj, Kaoru Ito, Sandra Jagodzińska, Siratun Jannat, Tomáš Jelínek, Apoorva Jha, Katharine Jiang, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Ritván Karahóa, Andre Kåsen, Tolga Kayadelen, Sarveswaran Kengatharaiyer, Václava Kettnerová, Lilit Kharatyan, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Petr Kocharov, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Alexey Koshevoy, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Adrian Kuqi, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Kris Kyle, Käbi Laan, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Lauren Levine, Cheuk Ying Li, Josie Li, Keying Li, Yixuan Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Yi-Ju Jessica Lin, Krister Lindén, Yang Janet Liu, Nikola Ljubešić, Irina Lobzhanidze, Olga Loginova, Lucelene Lopes, Stefano Lusito, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, Ilya Makarchuk, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Stella Markantonatou, Héctor Martínez Alonso, Lorena Martín Rodríguez, André Martins, Cláudia Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Tatiana Merzhevich, Niko Miekka,

Aaron Miller, Karina Mischenkova, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Manuela Nevaci, Lng Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Maria das Graças Volpe Nunes, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Hulda Óladóttir, Adéday`ọ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Noam Ordan, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Teresa Paccosi, Alessio Palmero Aprosio, Anastasia Panova, Thiago Alexandre Salgueiro Pardo, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Giulia Pedonese, Angelika Peljak-Łapińska, Siyao Peng, Siyao Logan Peng, Rita Pereira, Sílvia Pereira, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Andrea Peverelli, Jason Phelan, Claudel Pierre-Louis, Jussi Piitulainen, Yuval Pinter, Clara Pinto, Rodrigo Pintucci, Tommi A Pirinen, Emily Pitler, Magdalena Plamada, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Robert Pugh, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andreia Querido, Andriela Rääbis, Alexandre Rademaker, Mizanur Rahoman, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Joana Ramos, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Arij Riabi, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eiríkur Rögnvaldsson, Ivan Roksandic, Mykhailo Romanenko, Rudolf Rosa, Valentin Roca, Davide Rovati, Ben Rozonoyer, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Ezgi Sanıyar, Dage Särg, Marta Sartor, Mitsuya Sasaki, Baiba Saulīte, Agata Savary, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Emmanuel Schang, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Syeda Shahzadi, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Maria Shvedova, Janine Siewert, Einar Freyr Sigurðsson, João Silva, Aline Silveira, Natalia Silveira, Sara Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Haukur Barri Símonarson, Kiril Simov, Dmitri Sitchinava, Ted Sither, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Per Erik Solberg, Barbara Sonnenhauser, Shafi Sourov, Rachele Sprugnoli, Vivian Stamou,

146

Steinhór Steingrímsson, Antonio Stella, Abishek Stephen, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Daniel Swanson, Zsolt Szántó, Chihiro Taguchi, Dima Taji, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Dipta Tanaya, Mirko Tavoni, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Sara Tonelli, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sveinbjörn Hórðarson, Vilhjálmur Horsteinsson, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Elena Vagnoni, Sowmya Vajjala, Socrates Vak, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Uliana Vedenina, Giulia Venturi, Eric Villemonte de la Clergerie, Veronika Vincze, Natalia Vlasova, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Shira Wigderson, Sri Hartati Wijono, Vanessa Berwanger Wille, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Qishen Wu, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Olcay Taner Yıldız, Zhuoran Yu, Arlisa Yuliawati, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, He Zhou, Hanzhi Zhu, Yilun Zhu, Anna Zhuravleva, and Rayan Ziane. 2023. Universal dependencies 2.13. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.