Evaluating zero-shot transfers and multilingual models for dependency parsing and POS tagging within the low-resource language family Tupían

Frederic Blum
Institut für deutsche Sprache und Linguistik
Humboldt-Universität zu Berlin
frederic.blum@hu-berlin.de

Abstract

This work presents two experiments with the goal of replicating the transferability of dependency parsers and POS taggers trained on closely related languages within the low-resource language family Tupían. The experiments include both zero-shot settings as well as multilingual models. Previous studies have found that even a comparably small treebank from a closely related language will improve sequence labelling considerably in such cases. Results from both POS tagging and dependency parsing confirm previous evidence that the closer the phylogenetic relation between two languages, the better the predictions for sequence labelling tasks get. In many cases, the results are improved if multiple languages from the same family are combined. This suggests that in addition to leveraging similarity between two related languages, the incorporation of multiple languages of the same family might lead to better results in transfer learning for NLP applications.

1 Introduction

For most of the 7000 languages of the world, no NLP resources exist (Joshi et al., 2020; Mager et al., 2018). As a response to this situation, more and more initiatives emerged in recent years that work on NLP applications for underrepresented and low-resource languages (Orife et al., 2020; Nekoto et al., 2020; Mager et al., 2021). Despite those advances, access to tools like machine translation still is hindered by a large language barrier. Most of those languages do not have large text corpora, which have been used for the recent advantages in NLP like the building of large transformer models (Vaswani et al., 2017). Annotated data and parallel corpora thus remain an important but scarce tool for many of them. Yet, annotating this data is a challenge itself, and might be aided through the transfer of models from languages with more available resources.

The idea to leverage existing databases and models for cross-lingual transfer is not new (Aufrant et al., 2016; Duong et al., 2015; Lacroix et al., 2016; Vania et al., 2019; Wang et al., 2019). However, many studies even in this area remain within the environment of high-resource languages, and benchmarks with a typological sample as representative as possible - common nowadays in linguistic typology - are rarely found (Bender, 2009; de Lhoneux, 2019; Ponti et al., 2019). The main goal of this contribution is to replicate previous findings on cross-lingual transfer in low-resource settings (Meechan-Maddon and Nivre, 2019) within an underpresented language family, Tupían.

2 Data and Hypotheses

The data used for this study is taken from the Tupían Dependency Treebanks project (TuDeT, Gerardi et al., 2021)[1], which is openly available under a CC-BY-SA-4.0 License and is already partially present in the Universal Dependencies database. The author is not part of the team that developed these treebanks. There are currently seven languages in the dataset, which belong to different branches of the Tupían family (Hammarström et al., 2021). Except Tupinambá, which is extinct, the languages are spoken in Brazilian territory. All languages but Guajajára have SOV word order, while the former has VSO. The datasets are summarized in Table 1. There are some important differences with respect to the distribution of annotations data. For example, adjectives are absent for nearly all languages but Karo, either because they do not have adjectives and use stative verbs instead like Guajajára (Harrison, 2010), or because of low sample size. There are some tags, like NUM and INTJ, which are quite unevenly distributed between the available treebanks for the respective languages. As a consequence, this will result in low macro-f1

[1] https://github.com/tupian-language-resources/tudet

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Student Research Workshop, pages 1 - 9

Language	Code	Branch	Word order	Tokens	Utterances	Tokens per utterance
Akuntsu	aqz	Tuparic	SOV	408	101	4.04
Guajajára	gub	Tupi-Guarani	VSO	3571	497	7.18
Kaapor	urb	Tupi-Guarani	SOV	366	83	4.41
Karo	arr	Ramarama	SOV	2318	674	3.44
Makuráp	mpu	Tuparic	SOV	146	31	4.71
Mundurukú	myu	Mundurukuic	SOV	828	124	6.68
Tupinambá	tpn	Tupi-Guarani	SOV	2576	353	7.30

Table 1: Treebanks used in the dataset

scores, making accuracy the more relevant measure for this research question. A detailed description of the distribution of UPOS-tags in the dataset is given in Appendix A, the distribution of dependency relations is given in Appendix B.

In this study, I primarily test the utility of cross-lingual transfer for POS-taggers and dependency parsers with special attention given to language phylogeny. Language phylogeny can be seen as a proxy to typological features, given that closely related languages usually show many structural similarities. Previous studies have shown that even a comparably small treebank from a closely related language will improve the results of annotation considerably (Meechan-Maddon and Nivre, 2019).

Recent studies suggest to leverage phylogenetic proximity in a more efficient way than simply comparing languages based on the language family they belong to (Dehouck and Denis, 2019). Which model generalizes best over the different treebanks used in this sample, and what role does language phylogeny play in this? In this study, 'closeness' of two languages is defined based on the proximity of their phylogenetic clades. This is used as a proxy to their typological similarity. Especially for languages which do not have extensive descriptive material available, such similarities cannot easily be computed from typological databases. Based on phylolinguistic inferences about Tupían (Galucio et al., 2015; Gerardi and Reichert, 2021), the following explicit hypotheses are postulated:

1. Guajajára and Tupinambá should provide the best results for the evaluation of Kaapor, given that all three are part of the Tupi-Guarani branch of the Tupían language family.

2. Despite belonging to three different branches, the remaining four languages are quite close to each other in networks of lexical similarity. Here, Mundurukú is closer to Akuntsú than to Makuráp, and Karo is closer to Makuráp than to Akuntsú. The results should mirror this relation.

3 Experiments

One of the challenges for NLP applications with low-resource languages is the lack of language-specific resources on which embeddings can be trained on (Mager et al., 2018). Even though there are useful pipelines which can sometimes be used to crawl monolingual data from published sources (Bustamante et al., 2020), those are not always available or accessible. The embeddings used for the experiments in this contributions are based on the jw300-corpus (Agić and Vulić, 2019). This corpus is derived specifically from 343 low-resource languages and shows greater typological diversity than most dominating multilingual models. The embeddings are implemented in flair (Akbik et al., 2018). They have been fine-tuned for the pooled set of source languages. Transformer word embeddings mBERT (Devlin et al., 2019) and ROBERTA (Conneau et al., 2020) were also evaluated for the model, but rarely surpassed 40% accuracy for the source languages and have thus been discarded from further experiments for now. This results further call into question the utility of such large models for typologically diverse languages, and strengthens previous findings that even the largest multilingual transformer models do not show good results when transferring to typologically different languages (Ahmad et al., 2019; Lauscher et al., 2020; Pires et al., 2019). However, the exact reasons for their failure in this experiment are not entirely clear and need further research with more typologically diverse low-resource languages.

The experiments will be done for both POS tagging and dependency parsing and include a zero-shot setting. Also, models trained on individual source languages will be compared against models

trained on multiple datasets, with the evaluation set being the remaining treebanks of the dataset. Given the small amoung of training data and the models chosen, all model runs combined did not need more than three hours on CPU. The evaluation was done within the provided utilities by flair and SuPaR, respectively. All code is available on OSF.[2]

3.1 POS-tagging

For all experiments, the datasets have been separated into source (Guajajára, Karo, Tupinambá) and target languages (Akuntsu, Kaapor, Makuráp, Mundurukú). The split has been made according to the availability of data, and all treebanks with over 2000 annotated tokens have been used as source language. The main reason for this is to assure that the training sets have sufficient data for training and evaluating the models. Every treebank in the source set was further split into training, test and dev data (80/10/10). Given the scarcity of the data, all models were trained including the dev-set. The model itself a BiLSTM-CRF sequence tagger implemented using the *flair*-framework (Akbik et al., 2019, Version 0.10),[3] trained with a hidden size of 512. The following models were run:

1. training on the combined source set (tupi3)
2. training on the individual source languages Guajajára (gub), Karo (arr) and Tupinambá (tpn)
3. fine-tuning the tupi3 model for each Akuntsu (tupi3-aqz) and Mundurukú (tupi3+myu) on 50% of of the respective data, with the remaining part of the data used as evaluation
4. using a model pre-trained for 12 European UD languages, implemented in *flair* (Akbik et al., 2018).[4] This model was trained on treebanks from Czech, Danish, Dutch, English, Finnish, French, German, Italian, Norwegian, Polish, Spanish, and Swedish

The pre-trained model for European languages was used in order to provide a baseline of transferability of models based on unrelated, high-resource languages. All models were evaluated on each target language. Each model was run five times, and the average results are presented in Table 2. In case of the fine-tuning experiment, training accuracy describes the result on the test set, while the language-specific column gives the result for the overall treebank. The evaluation column is a summary over the evaluation set, without considering the source language. The best result for each of the languages in the evaluation set is boldfaced.

Unsurprisingly, the experiment conditions with fine-tuning for a specific language show the best results for the respective language. In both cases, the results for the other language were also improved, confirming the hypothesis that the results of Akuntsu and Mundurukú should be closely related. This could motivate training a model on Akuntsu and Mundurukú combined. The close relationship between Akuntsu and Makuráp, on the other hand, does not seem to lead to better results. The best predictions for Makuráp are instead based on the model trained for Karo, a relationship that was predicted by the second hypothesis, even though only as the second strongest effect. Despite those results, it should be considered that Makuráp has by far the smallest treebank available with only 146 annotated tokens, so no final evaluations should be made. This also reflects in the low overall accuracy in all settings for Makuráp, never surpassing 40%.

3.2 Dependency parsing

The experiment settings were mostly identical for the dependency parsing experiment. The main difference is that no pre-trained model for European languages is available for the dependency parser that was used for the experiments. For the same reason, no fine-tuning for the tupi3 setting is implemented so far. Instead, a single model for Mundurukú was added for further evaluation of Hypothesis 2. As model architecture, an implementation of the deep biaffine dependency parser (Dozat and Manning, 2017) from SuPar (Version 1.01) was used (Zhang et al., 2020).[5] The results are shown in Table 3. In case the language was the source language, the evaluation score only reflects the evaluation of the test split. This is the case for the tupi3 setting as well as the individual languages. All other languages in each row were evaluated against the entire dataset. As the main evaluation criteria, Labelled Attachment Scores (LAS) were chosen.

[2] https://doi.org/10.17605/OSF.IO/ZHDMP
[3] https://github.com/flairNLP/flair, MIT License
[4] https://huggingface.co/flair/upos-multi
[5] https://github.com/yzhangcs/parser, MIT License

Model	TrainAcc	TrainF1	EvalAcc	EvalF1	aqz	mpu	myu	urb
arr	0.84	0.68	0.30	0.10	0.35	**0.36**	0.30	0.24
gub	0.91	0.76	0.44	0.19	0.45	0.29	0.48	0.41
tpn	0.87	0.81	0.42	0.17	0.43	0.25	0.49	0.34
tupi3	0.86	0.64	0.46	0.20	0.49	0.35	0.47	**0.42**
tupi3+aqz	0.56	0.31	0.48	0.19	**0.52**	0.32	0.51	0.40
tupi3+myu	0.55	0.22	0.48	0.19	0.51	0.34	**0.53**	0.39
multi			0.33	0.13	0.38	0.23	0.36	0.23

Table 2: Average training and evaluation accuracy and F1-scores over five runs of the POS tagging experiment

Model	aqz	arr	gub	mpu	myu	tpn	urb
arr	0.00	**64.10**	0.00	**25.00**	0.00	0.00	0.00
gub	12.90	14.50	**73.30**	9.00	8.90	10.30	14.20
myu	**19.09**	14.98	10.65	7.64	**65.28**	7.85	13.89
tpn	13.30	0.00	20.90	14.30	0.00	**46.40**	15.80
tupi3	9.50	62.60	72.70	11.80	8.90	42.90	**21.80**

Table 3: Labelled Attachment Scores (LAS) of the dependency parsing experiment

4 Discussion

4.1 Discussing the POS tagging experiment

Against Hypothesis 1, the best result for Kaapor is not achieved by Guajajára or Tupinambá, but by the combined model trained on the pooled treebanks. However, the model of Guajajára is only 0.01% behind the pooled model and should be considerd equal, as it is well within the standard deviation of the average result (upos 0.02, gub 0.01). It should also not be forgotten that two of the three languages in the pooled set, including Guajajára itself, are part of the Tupí-Guarani branch, which can be reasonably postulated as part of the reason that tupi3 scores so high. Instead of a single language of that branch, it might just be the combination of two languages from the same branch that shows such strong results.

This leads to another result that should be highlighted, namely the overall usefulness of the multilingual Tupían model. While the European multilingual model had, perhaps expectedly without any fine-tuning, low results for most evaluations, the Tupían model was competitive in most settings. For both Makuráp and Kaapor it was basically equal with the best individual model, for Akuntsu it was second best behind the fine-tuned models, and even for Mundurukú it showed good results, even though it showed weaker predictions in this case. While previous studies suggested that at least 200 annotated utterances are sufficient to improve the results of a multilingual model considerably (Meechan-Maddon and Nivre, 2019), the results in this contribution suggest that as few as 50 or 60 training utterances could already provide a considerable improvement of the evaluation scores. These are only approximate numbers, and definitely need more experiments with other datasets in order to be confirmed.

All in all, the POS tagging experiment shows that language phylogeny is a strong, but not a deterministic predictor for the transferability of models. Given the low amount of training data for the models even in the combined tupi3 setting, the zero-shot transfer results are better than perhaps expected.

4.2 Discussing the Dependency Parsing experiment

Overall, the transfer LAS are much lower than the accuracy in the previous experiment. Given the complexity of dependency parsing compared to POS tagging, this is hardly surprising. This is also true for the training scores, never surpassing 75%. With regard to Hypothesis 1, we see again that both Guajajára and Tupinambá show better results for Kaapor than Karo and Mundurukú. The model hugely improves in the tupi3 setting, indicating again that both larger training treebanks and combining different closely related languages might show considerable effects to the evaluation of a new language. This has already been the case for the POS tagging, and will result in an additional experiment in the next phase of this study.

Hypothesis 2 is also largely confirmed. Karo

was hypothesized to achieve the best results for the evaluation of Makuráp, and this predictions is met strongly, with a LAS difference over 10%. As Mundurukú outperforms the other languages in the evaluation of Akuntsu, the second part of the hypothesis is also confirmed. The results for Mundurukú itself further show that even with a small treebank of only $\sim$ 100 utterances, good predictions can be achieved.

At the current state of this paper, an important gap is the missing detailed error analysis. One important source of errors for the models is the uneven distribution of dependency relations between the treebanks, as shown in Table 5. Partially due to the low amount of data and due to language-specific differences, some tags are distributed unevenly among languages, or are not present at all in some of them. However, even when accounting for this differences, the exact factors that determine failure and success of the transfer remain not fully explained. For example, whether the overall success of the combined model of various languages (tupi3) is due to the higher amount of training data, or whether there are other factors involved when combining data from multiple languages that could be leveraged for the development of NLP applications for low-resource languages, cannot be answered by this contribution.

5 Conclusion

This study further confirms previous findings that cross-lingual transfer of dependency parsers and POS taggers is a viable option in low-resource settings if a closely related language is available (Vania et al., 2019; Meechan-Maddon and Nivre, 2019). This extends previous evidence for phylogenetically informed transfer from Indo-European and Uralic (Dehouck and Denis, 2019) to Tupían. Further experiments on other language families should be conducted in order to confirm the exact features that make successful transfer possible.

Further, this study provided further evidence for extending the phylolinguistically informed combination of source languages. In all experiment settings of this study, the pooled source language set had very good results, and a targeted combination will likely further improve the results. Further follow-up experiments will consist of targeted combinations of annotated data from different languages, including an incorporation of typological features and delexicalized transfer. In preliminary experiments, CRF2o dependency parsing (Zhang et al., 2020) showed promising results for transfer results as well. Especially in the dependency parsing experiment the transfer scores were quite low, and further improving the training data as well as comparing different models should be a viable solution for this challenge.

References

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. 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 2440–2452, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Lauriane Aufrant, Guillaume Wisniewski, and François Yvon. 2016. Zero-resource dependency parsing: Boosting delexicalized cross-lingual transfer with linguistic knowledge. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 119–130, Osaka, Japan. The COLING 2016 Organizing Committee.

Emily M. Bender. 2009. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece. Association for Computational Linguistics.

Gina Bustamante, Arturo Oncevay, and Roberto Zariquiey. 2020. No data to crawl? monolingual corpus creation from PDF files of truly low-resource

languages in Peru. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2914–2923, Marseille, France. European Language Resources Association.

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.

Miryam de Lhoneux. 2019. *Linguistically informed neural dependency parsing for typologically diverse languages*. Ph.D. thesis, Acta Universitatis Upsaliensis.

Mathieu Dehouck and Pascal Denis. 2019. Phylogenic multi-lingual dependency parsing. 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 192–203, Minneapolis, Minnesota. 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.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *ICLR 2017*.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. A neural network model for low-resource Universal Dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 339–348, Lisbon, Portugal. Association for Computational Linguistics.

Ana Vilacy Galucio, Sérgio Meira, Joshua Birchall, Denny Moore, Nilson Gabas, Sebastian Drude, Luciana Storto, Gessiane Picanço, and Carmen Reis Rodrigues. 2015. Genealogical relations and lexical distances within the tupian linguistic family. *Boletim do Museu Paraense Emílio Goeldi. Ciências Humanas*, 10:229–274.

Fabrício Ferraz Gerardi and Stanislav Reichert. 2021. The tupí-guaraní language family. *Diachronica*, 38(2):151–188.

Fabrício Ferraz Gerardi, Stanislav Reichert, Carolina Aragon, Lorena Martín-Rodríguez, Gustavo Godoy, and Tatiana Merzhevich. 2021. Tudet: Tupían dependency treebank.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2021. glottolog/glottolog: Glottolog database 4.5.

Carl H. Harrison. 2010. Verb prominence, verb initialness, ergativity and typological disharmony in guajajara. In Desmond C. Derbyshire and Geoffrey K. Pullum, editors, *Volume 1 Handbook of Amazonian languages: Volume 1*, pages 407–439. De Gruyter Mouton.

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.

Ophélie Lacroix, Lauriane Aufrant, Guillaume Wisniewski, and François Yvon. 2016. Frustratingly easy cross-lingual transfer for transition-based dependency parsing. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1058–1063, San Diego, California. 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.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Manuel Mager, Arturo Oncevay, Annette Rios, Ivan Vladimir Meza Ruiz, Alexis Palmer, Graham Neubig, and Katharina Kann, editors. 2021. *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*. Association for Computational Linguistics, Online.

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.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo,

Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Z. Abbott, Vukosi Marivate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan Van Biljon, Arshath Ramkilowan, Adewale Akinfaderin, Alp Öktem, Wole Akin, Ghollah Kioko, Kevin Degila, Herman Kamper, Bonaventure Dossou, Chris Emezue, Kelechi Ogueji, and Abdallah Bashir. 2020. Masakhane - machine translation for africa. *CoRR*, abs/2003.11529.

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

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Clara Vania, Yova Kementchedjhieva, Anders Søgaard, and Adam Lopez. 2019. A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. 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 1105–1116, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual BERT transformation for zero-shot dependency parsing. 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 5721–5727, Hong Kong, China. Association for Computational Linguistics.

Yu Zhang, Zhenghua Li, and Min Zhang. 2020. Efficient second-order TreeCRF for neural dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3295–3305, Online. Association for Computational Linguistics.

A POS-tags used in the dataset

	UPOS	Akuntsu	Guajajára	Kaapor	Karo	Makuráp	Mundurukú	Tupinambá
1	ADJ	2		3	103		5	
2	ADP	29	79	25	36	27	126	73
3	ADV	32	68	101	42	137	29	76
4	AUX	7	9	16	75	14	12	4
5	DET	49	24	8		41	5	20
6	INTJ	5	3			14	2	8
7	NOUN	429	250	240	244	219	408	338
8	NUM	15	1		2		2	4
9	PART	39	132	101	129	103	25	42
10	PRON	78	32	172	129	75	59	48
11	PROPN	42	41	55	5		4	34
12	PUNCT	88	176	16	1	14	115	209
13	VERB	184	181	246	222	329	179	140
14	CCONJ		2	11		27	2	1
15	SCONJ		2	5	10		23	1
16	X				2		4	1

Table 4: POS tags per 1.000 Tokens used in TuDeT

B Dependency relations used in the dataset

	deprel	Akuntsu	Guajajára	Kaapor	Karo	Makuráp	Mundurukú	Tupinambá
1	advmod	39	65	101	80	137	25	62
2	amod	5		25	29		6	1
3	appos	15	6		2		10	24
4	aux	2	9	16	57	14	8	4
5	case	34	56	19	36	27	121	62
6	ccomp	2	16	5	3		4	5
7	conj	15	8	5	3	21	12	30
8	dep	17	11		29	116	25	24
9	discourse	39	139	87	26	89	14	42
10	dislocated	2						1
11	iobj	2	14	14				1
12	nmod	135	52	63	60	48	63	94
13	nsubj	150	91	202	127	62	95	45
14	nummod	12	0		2		1	4
15	obj	91	55	156	65	82	54	42
16	obl	59	113	22	31	27	175	99
17	parataxis	44	5		3	96	34	32
18	punct	88	176	16	1	14	115	209
19	root	248	139	227	291	212	150	137
20	advcl		16	8	1	14	41	54
21	compound		1	5	19			1
22	det		18	3	3		4	3
23	flat		1				1	
24	list		2					
25	mark		7	5	47		28	0
26	orphan		1					
27	cc			11		21	1	2
28	csubj			5				
29	xcomp			3	8	21	1	7
30	acl				2			4
31	clf				66		10	
32	cop				9		1	
33	goeswith							1
34	obl:obj							3
35	obl:subj							5
36	vocative							1

Table 5: Dependency relations per 1.000 tokens used in TuDeT