YAD: LEVERAGING T5 FOR IMPROVED AUTOMATIC DIACRITIZATION OF YORÙBÁ TEXT

Akindele Michael Olawole[1,*], Jesujoba O. Alabi[2*], Aderonke Busayo Sakpere[1], David I. Adelani[3]
[1] Department of Computer Science, University of Ibadan, Nigeria
[2] Spoken Language Systems (LSV), Saarland University, Saarland Informatics Campus, Germany
[3] University College London
jalabi@lsv.uni-saarland.de, d.adelani@ucl.ac.uk

Introduction Yorùbá, a language spoken predominantly in West Africa, is renowned for its tonal nature which is characterized by a heavy use of diacritics to signify tone variations. In Yorùbá and many other languages, diacritics play a crucial role in disambiguating word meanings and in word pronunciation, making accurate diacritization essential for effective communication and language processing tasks (Skiredj & Berrada, 2024). However, manual diacritization is time-consuming and requires specialized linguistic expertise, motivating the development of automatic diacritization systems. In recent years, significant progress has been made in natural language processing (NLP) techniques, leading to the exploration of various approaches to automate the diacritization process for languages using diacritics (Náplava et al., 2018; Mubarak et al., 2019; Náplava et al., 2021; Stankevicius et al., 2022, *inter alia*) including Yorùbá (Orife, 2018; Orife et al., 2020). Despite these efforts, there is the absence of a standard benchmark dataset for evaluating and comparing developed Yorùbá diacritization systems. Furthermore, the emphasis has predominantly been on model development rather than assessing the usability of these systems, particularly in the context of lightweight models amidst the prevalence of large neural networks in contemporary NLP research. In this study, we address these gaps by introducing YAD (Yorùbá Automatic Diacritization Dataset), which we curated by leveraging the MENYO-20k (Adelani et al., 2021) dataset. Moreover, we focus on developing lightweight diacritizers for Yoruba, leveraging T5 (Raffel et al., 2020) a text-to-text transformer architecture, to improve usability and efficiency in the era of large neural networks.

Yorùbá language Yorùbá orthography is based on the Latin alphabet. Its alphabets include 25 Latin letters excluding characters c, q, v, x and z, and its has five additional letter (ẹ, gb, ṣ , ọ) representing certain sounds. Diacritics including accents and undots are used in Yorùbá to capture tone and vowel height. The tones include low, middle and high which are denoted by the grave (e.g. "à"), optional macron (e.g. "ā") and acute (e.g. "á") accents respectively.

Yorùbá automatic diacritization dataset (YAD) In this work, we present YAD, a dataset created for evaluating diacritization systems for Yorùbá. YAD is constructed based on the MENYO-20k dataset, which is a benchmark for English-Yorùbá translation, containing 20,000 parallel sentences. We divided the Yorùbá side of MENYO-20k's test data into two halves, allocating one half as development set (Dev.) and the other half as test set (Test). We ensured that both halves of the data had equal representation across all domains present in MENYO-20k. Given these splits, we applied four heuristics on the data, considering the sequence-to-sequence nature of the diacritization task. This task requires a source side and a target side, where the source side can consist of Yorùbá text without diacritics or partially diacritized texts, and the target side is fully diacritized Yorùbá texts. To create and prepare the source side of the data, the following heuristics were applied with specific proportions on the original, well-diacritized texts: (i) removal of all diacritics for 60% of the data, (ii) removal of only tone marks for 20% of the data, (iii) a combination of (ii) and the removal of any random word for 10%, and (iv) a combination of (ii) and swapping any two distinct words (applicable only for sentences with more than two words). Where (iii) and (iv) can be seen as infilling and grammatical error correction tasks respectively. Example 1, shows an example from the development set where (iv) was applied where sibẹsibẹ, and Ṣugbọn were swapped on the source side.

(1) **source**: sibẹsibẹ, Ṣugbọn Mama o gbagbọ.
reference: Ṣùgbọ́n síbẹ̀síbẹ̀, Màmá ò gbàgbọ́

*Equal contribution.

Diacritization model We pre-train T5 models from scratch using Yorùbá WURA corpus (Oladipo et al., 2023) and JW300 (Agić & Vulić, 2019) with different sizes: tiny (number of layers encoder/decoder (L=4), heads (H=4)) , mini (L=4, H=4), small (L=8, H=6), base (L=12, H=12) corresponding to 14M, 18M, 60M and 280M parameters. All models make use of a vocabulary size of 32K except the tiny model. We called our model o**yo**-T5. [1]. All models were pre-trained using HuggingFace Flax code and TPU v3-8 for 100-200K steps, which took around one day each.

After pre-training, we trained our baseline diacritizer models by fine-tuning existing multilingually trained language models, MT5 (Xue et al., 2021), AfriMT5 (Adelani et al., 2022), AfriTeVa-V2 (Oladipo et al., 2023), and UMT5 (Chung et al., 2023). We use MT5-base, Afri-MT5-base, the base and large variants of AfriTeVa-V2, and UMT5. we fine-tune these models using HuggingFace Transformers (Wolf et al., 2020). These eventual diacritization models are evaluated using the test and development SacreBLEU (Post, 2018)[2] and ChrF (Popović, 2015).

Training and evaluation data For our experiments, we trained diacritization models using data from three sources, which are, the Yorùbá Bible, combination of training and development split of MENYO-20k (now referred to as YAD training split from here onwards), and JW300 (Agić & Vulić, 2019). Table 1 shows the number of sentences in each split. Different combinations of these datasets were used for training our models, and unless specified otherwise we implemented the four previously described heuristics on the training data also. We evaluated our models on the YAD development and test splits, as well as on an existing Global Voices (GV) test set from Orife et al. (2020). Additionally, we used a sampled 10% of the Bible as a test set. For reproducibility, we release our code, data and models on GitHub[3].

Baseline models and result We trained diacritization models by fine-tuning MT5, Afri-MT5, AfriTeVa-V2, and UMT5, on the YAD training split. The experimental results as presented in Table 2 shows that AfriTeVa-V2-large outperform other massively multilingual language models. Interestingly, Oyo-T5-base demonstrates competitiveness on both BLEU and CHRF similar to AfriTeVa-V2-large, despite having fewer parameters compared to AfriTeVa-V2-large.

Corpus	Train	Dev	Test
Bible	27739	-	3083
JW300	459,871	-	-
YAD	13,467	3,326	3,330
Previous test set: Orife et al. (2020)			
GV	-	-	619

Table 1: Data split of different corpus.

Model	Size	Dev. eval. BLEU	Dev. eval. CHRF	Test eval. BLEU	Test eval. CHRF
UMT5-base	580M	38.1	62.5	39.3	64.0
MT5-base	580M	48.1	69.4	49.5	71.2
AfriMT5-base	580M	53.3	72.8	54.0	74.2
AfriTeVa-base	313M	56.9	72.9	57.2	74.0
AfriTeVa-large	871M	67.2	80.8	66.8	81.3
Oyo-T5-base	280M	**71.9**	**82.2**	**70.2**	**82.3**

Table 2: Diacritization model evaluation on YAD Dev. and Test sets.

Effect of model scaling Given that Oyo-T5-base is a competitive model despite its size, we chose to train light-weight diacritizers by finetuning the smaller variants of Oyo-T5 model on the YAD training data to check the effect of model size on this task. We evaluated the models on YAD dev. and test sets. As shown in Table 4, bigger models are better for this task. However, Oyo-T5-small with 60M parameters even outperformed AfriTeVa-base with 313M parameters on the same task.

Effect of training data scaling Going further, we choose to evaluate the effect of training data size on the diacritization models. Hence trained new diacritization models on Bible, JW300, and a combination of them and YAD training data using Oyo-T5-base as the backbone model and we evaluate them on the Bible, GV, and YAD test sets. For these test sets, we use the version where all diacritics were removed from the source side without applying the earlier described heuristics [4].

Table 3 shows that training on more data is better. Also training on just JW300 outperformed the best result achieved by Orife et al. (2020) on GV by +11.6 BLEU point. Furthermore, we observed that models trained on a specific domain tend to perform better on that domain. For instance,

[1]Oyo Yorùbá (YO) dialect of T5. Oyo dialect is the standardized dialect for writing Yorùbá

[2]"intl" tokenizer, all data comes untokenized.

[3]`https://github.com/ajesujoba/YAD`

[4]We we also analyze this effect on text with only underdot on the source side and provide the result in Appendix A

models trained on the Bible achieved a $+17.6$ BLEU point improvement compared to the next best performing model, which was trained on JW300. It is important to note that YAD's training data, obtained from the combination of MENYO-20k training and development splits, included Global Voices news. Therefore, we suspect that its high BLEU scores on the GV tests set could be attributed to potential data leakage. Lastly, our findings suggest that more data generally leads to better performance, as evidenced by the results from the model trained on the combined dataset.

Model/Data	**Bible**	**GV**	**YAD**
Data: JW300+Bible+others			
Orife et al. (2020)	-	59.8	-
Bible	88.1	55.0	57.2
JW300	71.7	71.4	66.9
YAD	69.0	82.1	72.0
JW300+Bible+YAD	90.8	91.0	**78.7**

Table 3: Data scale effect on Bible, GV & YAD.

		Dev. eval.		**Test eval.**	
Model	**Size**	**BLEU**	**CHRF**	**BLEU**	**CHRF**
Oyo-T5-tiny	14M	42.1	62.1	36.3	60.4
Oyo-T5-mini	18M	49.0	64.6	47.5	65.6
Oyo-T5-small	60M	63.5	75.8	62.9	76.7
Oyo-T5-base	280M	**71.9**	**82.2**	**70.2**	**82.3**

Table 4: Effect of the model scale on YAD.

Conclusion In this work we present Yorùbá automatic diacritization (YAD) benchmark dataset for evaluating Yorùbá diacritization systems. In addition, we pre-train text-to-text transformer, T5 model for Yorùbá and showed that this model outperform several multilingually trained T5 models. Lastly, we showed that more data and bigger models are better at diacritization for Yorùbá.

ACKNOWLEDGMENTS

Jesujoba Alabi was partially supported by the BMBF's (German Federal Ministry of Education and Research) SLIK project under the grant 01IS22015C. David Adelani acknowledges the support of DeepMind Academic Fellowship programme. This work was supported in part by Oracle Cloud credits and related resources provided by Oracle.

REFERENCES

David Adelani, Dana Ruiter, Jesujoba Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Esther Awokoya, and Cristina España-Bonet. The effect of domain and diacritics in Yoruba–English neural machine translation. In Kevin Duh and Francisco Guzmán (eds.), *Proceedings of Machine Translation Summit XVIII: Research Track*, pp. 61–75, Virtual, August 2021. Association for Machine Translation in the Americas. URL `https://aclanthology.org/2021.mtsummit-research.6`.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3053–3070, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.223. URL `https://aclanthology.org/2022.naacl-main.223`.

Željko Agić and Ivan Vulić. JW300: A wide-coverage parallel corpus for low-resource languages. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3204–3210, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1310. URL `https://aclanthology.org/P19-1310`.

Hyung Won Chung, Xavier Garcia, Adam Roberts, Yi Tay, Orhan Firat, Sharan Narang, and Noah Constant. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=kXwdL1cWOAi`.

Hamdy Mubarak, Ahmed Abdelali, Hassan Sajjad, Younes Samih, and Kareem Darwish. Highly effective Arabic diacritization using sequence to sequence modeling. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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)*, pp. 2390–2395, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1248. URL `https://aclanthology.org/N19-1248`.

Jakub Náplava, Milan Straka, Pavel Straňák, and Jan Hajič. Diacritics restoration using neural networks. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL `https://aclanthology.org/L18-1247`.

Jakub Náplava, Milan Straka, and Jana Straková. Diacritics restoration using BERT with analysis on czech language. *CoRR*, abs/2105.11408, 2021. URL `https://arxiv.org/abs/2105.11408`.

Akintunde Oladipo, Mofetoluwa Adeyemi, Orevaoghene Ahia, Abraham Owodunni, Odunayo Ogundepo, David Adelani, and Jimmy Lin. Better quality pre-training data and t5 models for African languages. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 158–168, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.11. URL `https://aclanthology.org/2023.emnlp-main.11`.

Iroro Orife. Attentive sequence-to-sequence learning for diacritic restoration of yorùbá language text. *Proc. Interspeech 2018*, pp. 2848–2852, 2018.

Iroro Orife, David I. Adelani, Timi Fasubaa, Victor Williamson, Wuraola Fisayo Oyewusi, Olamilekan Wahab, and Kola Tubosun. Improving yorùbá diacritic restoration, 2020.

Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina (eds.), *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL `https://aclanthology.org/W15-3049`.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/W18-6319`.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL `http://jmlr.org/papers/v21/20-074.html`.

Abderrahman Skiredj and Ismail Berrada. Arabic text diacritization in the age of transfer learning: Token classification is all you need, 2024.

Lukas Stankevicius, Mantas Lukosevicius, Jurgita Kapociute-Dzikiene, Monika Briediene, and Tomas Krilavicius. Correcting diacritics and typos with byt5 transformer model. *CoRR*, abs/2201.13242, 2022. URL `https://arxiv.org/abs/2201.13242`.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL `https://aclanthology.org/2020.emnlp-demos.6`.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL `https://aclanthology.org/2021.naacl-main.41`.

A Effect of applying diacritics to undiacritized texts versus text with only underdot

	Bible		**GV.**		**YAD.**	
Model	**BLEU**	**CHRF**	**BLEU**	**CHRF**	**BLEU**	**CHRF**
no accent on source side						
Bible	88.2	94.3	55.2	68.4	57.8	72.9
JW300	64.0	80.9	68.8	81.6	65.0	80.9
Menyo	71.8	83.9	79.3	86.7	72.0	83.7
JW300+Bible+Menyo	89.6	95.8	90.0	95.1	78.4	89.0
no diacritics on source side						
Bible	88.1	94.2	55.0	67.4	57.2	71.6
JW300	71.7	82.6	71.4	83.7	66.9	81.9
Menyo	69.0	82.4	82.1	87.3	72.0	83.3
JW300+Bible+Menyo	90.8	96.0	91.0	95.1	78.7	88.8

Table 5: Effect of applying diacritics to undiacritized texts vs text with only underdots