VoxMg: An Automatic Speech Recognition Dataset for Malagasy

Falia Ramanantsoa
Laboratoire d'Informatique et de Mathématiques
Université d'Antananarivo
Antananarivo, Madagascar
falia.rmt@gmail.com

Abstract

African languages are not well-represented in Natural Language Processing (NLP). The main reason is a lack of resources for training models. Low-resource languages, such as Malagasy, cannot benefit from modern NLP methods if no datasets are available. This paper presents the curation and annotation of VoxMg, a speech dataset for Malagasy that consists of 3873 audio files totaling 10.80 hours. We also run a baseline, which is the first Automatic Speech Recognition (ASR) model ever built in this language and obtained a Word Error Rate (WER) of 33%.

Automatic Speech Recognition, Data Collection, Low-Resource Languages, Malagasy

1 Introduction

Automatic Speech Recognition (ASR) involves transcribing human speech automatically. Over the past few years, it has made significant progress and has become more ubiquitous in various aspects of our daily lives such as voice assistants and automatic captioning. However, these advancements mainly pertain to high-resource languages like English. The training of ASR models requires substantial amounts of speech data to achieve optimal performance (Moore, 2003). The data should also include a variety of speakers to capture the nuances of different speech patterns and accents (Garnerin et al., 2021). Additionally, the data should be noise-free for better performance, while not forgetting that a model trained solely on a clean dataset may not perform well in real-world environments (Openshaw & Masan, 1994). Therefore, creating an ASR system for low-resource languages is a challenging task (Magueresse et al., 2020). Despite the limited data availability, many efforts have been made to tackle this issue (Gales et al., 2014; Tüske et al., 2014; Bansal et al., 2018; Khare et al., 2021), such as the cross-lingual training approach (Babu et al., 2021).

Our main contributions are the creation of a Malagasy speech dataset called VoxMg and the training of a baseline ASR model. We present all the details concerning the curation and annotation of the dataset as well as the model fine-tuning process.

The rest of the paper is organized as follows: In Section 2, we review related works on the advancement of NLP for African languages. In Section 3, we outline the process used to create our own Malagasy dataset, named VoxMg. In Section 4, we provide details of the model development for the ASR task. In Section 5, we present the results of our experiments and analysis. Finally, in Section 6, we discuss potential future improvements and conclusions.

2 Related Works

NLP for African languages poses a significant challenge due to the scarcity of large and high-quality datasets. Efforts to address this issue include initiatives such as Mozilla Common Voice (Ardila et al., 2019) and the Masakhane project (Orife et al., 2020), which aim to advance NLP research in fostering a community-driven approach to data collection and sharing. Both initiatives have created and shared speech datasets for various African languages. Recently, a text and speech dataset for

Luganda, Runyankore-Rukiga, Acholi, Lumasaaba and Kiswahili was also made by Babirye et al. (2022).

In late 2020, a speech dataset for spoken language identification task known as VoxLingua107 (Valk & Alumäe, 2021) was made publicly available, it contains a total of 6628 hours of data for 107 languages, including Malagasy. This dataset is the only one that contains Malagasy speech, 109.21 hours of untranscribed speech, distributed over 44115 audio files in WAV mono format, 16kHz, whose duration ranges from 1 to 20 seconds. This dataset has been used as a starting point for our efforts related to speech recognition for Malagasy. Another valuable resource for the Malagasy language is the website nybaiboly.net [1], which offers free access to Malagasy biblical media, including the Malagasy Bible, radio broadcasts, biblical teachings, and a bilingual French-Malagasy Bible. The audio version of the Bible provided on the website is from a single-speaker and covers the Bible in its entirety.

3 Creation of the VoxMg Malagasy Speech Dataset

We obtained untranscribed audio from two sources, as described in Section 2: VoxLingua107 and nybaiboly.net. However, our ability to transcribe these data was limited due to resource constraints. Transcribing a large amount of data requires many annotators, but we were unable to secure enough participants. This section outlines the steps we took to process the obtained data.

3.1 The Malagasy Speech Data from VoxLingua107

We used Label Studio [2] to transcribe audio from VoxLingua107. It is an open source user-friendly data labeling tool. We selected 57% of the 9-second files, giving us 1499 files with a total of 3.75 hours of audio. Additionally, we chose 87% of the 10-second files, resulting in 1976 files with a total of 5.49 hours of audio. This gave us a total of 3475 files to transcribe, with a combined duration of 9.24 hours.

3.1.1 Transcription Rules

The following points are the rules we applied during transcription:

1. The texts are in lower case;
2. Punctuations are only used for apostrophes (“ ’ ”) and hyphens (“ - ”), as they are widely used in Malagasy. In this language, letters omission and compound words occupy a major place and follow specific rules that use these two punctuation marks. For example, the word for “peace” is “fandriam-pahalemana”, “interest” is “zana-bola”, and “afternoon” is “tolak’andro”;
3. Audios containing excessive foreign words are excluded;
4. Audios with very loud background music are excluded;
5. Audios with excessive noise are excluded;
6. Audios with too much voice overlap are excluded;
7. Audios with incomprehensible voice are excluded;
8. Audios that do not use the official Malagasy language are excluded;
9. Hybrid words are written with a hyphen (“ - ”) in between. Malagasy people have a tendency to use French words as radicals to create new words. For example, they may conjugate French verbs following the Malagasy rules of conjugation. Although these new words are not official, audios containing them were still included in the dataset as their ratio was small. For example, the French verb “lancer” (meaning “to throw”) creates the hybrid word “no-lancer-ny” (meaning “he throwed”), whereas the true Malagasy word is “natsipiny”;
10. Acronyms are not separated by punctuation;

[1] Malagasy Biblical Media
[2] Label Studio: Open Source Data Labeling Tool

11. The filled pauses used are: “ah”, “anh”, “eh”, “euh”, “oh”, “hmh”, “hein”, “huh”, “mh”;
12. Laughter is represented by “hahaha”, “hehehe”, “hihihi”, “huhuhu”;
13. Numbers are written in full form;
14. Truncated words are transcribed as they were heard. For example, if the word should have been “fianarana”, but only “narana” was heard, then “narana” is transcribed;
15. Hesitations are transcribed as they were heard. For example, “fia fianarana”, “mi mi” and so on.

3.1.2 Quality Assurance

For the first quality assurance task, we enlisted the help of 10 volunteers. The audio-transcription pairs were evenly distributed among them. They were given instructions to:

- Correct any spelling mistakes;
- Add any omitted words;
- Remove any words not present in the audio;
- Follow the transcription rules outlined in Section 3.1.1.

The second quality assurance task was performed by us. We thoroughly re-checked all the audio-transcription pairs, adhering to the rules we established, to ensure their accuracy and minimize errors. Our participants were volunteers and may not have been fully invested in the work, potentially leading to a less-than-perfect verification.

3.2 The Malagasy Speech Data from nybaiboly.net

Unlike the audio recordings from VoxLingua107, the audio recordings from nybaiboly.net were recorded in optimal conditions with no background noise and a clear voice, but they were recorded by only one male speaker. To prevent overfitting to this single speaker, we selected only 30 chapters from Genesis.

The transcription process involved aligning the text version with the audio version. We used a free software called Annotation Pro [3] for this task, which allowed us to easily fragment the recordings and match the corresponding annotations. We split the recordings by verse, resulting in 820 files that ranged between 3 to 20 seconds, totaling 2.20 hours after excluding some verses that did not contain valuable phrases.

The transcripts we obtained were not yet cleaned since we just took the texts from the Bible as they are and aligned them with the audios. The transcripts contained yet punctuations and uppercase words. The cleaning task was made along with the data obtained from VoxLingua107 in the postprocessing step.

3.3 Postprocessing of the Entire Data

After the hand labelling process of the data from VoxLingua107 and nybaiboly.net, further verification and processing were carried out on both using a Python script that included the following steps:

1. Normalization to lower case;
2. Removal of all punctuations, except for “ ’ ” and “ - ”;
3. Removal of duplicate spaces and line breaks;
4. Removal of digit numbers (some verse numbers were accidentally taken during the alignment);
5. Checking if there were still digit numbers that should be converted to full form;
6. Checking for spelling mistakes after sorting all words present in the whole transcripts;

[3] Annotation Pro: Let’s Annotate Together

7. Checking for words not present in the Malagasy vocabulary and for nonsense words to determine if they were mistakes, acronyms, or proper nouns.

3.4 Final Dataset

The final dataset consists of 3873 audio files, totaling 10.80 hours in duration. More details can be found in Table 1. To determine the number of male and female speakers in the data we took from VoxLingua107, we grouped the recordings based on the cosine distance between the speaker embeddings, specifically x-vectors (Snyder et al., 2018) generated using the XVectorSincNet pretrained model from Pyannote (Bredin et al., 2020). We then manually verified if each recording within each class was from the same speaker.

Table 1: VoxMg's statistics.

Corpus	**VoxLingua107**	**nybaiboly.net**	**Total**
Treated (hrs)	9.24	2.58	11.82
Validated (hrs)	8.60	2.20	10.80
Audio Chunks	3053	820	3873
Audio Chunks per Gender (M—F)	2229—824	820—0	3049—824
Num. Speakers (M—F)	306—153	1—0	307—153

3.5 Split

We split VoxMg such that the speakers in the training and validation sets do not appear in the test set. We used classes of speakers that contain a maximum of 4 audio chunks to construct the test partition. Further details are presented in Table 2.

Table 2: Statistics of train/val/test partitions of VoxMg.

Subset	**Train**	**Validation**	**Test**
Duration (hrs)	8.63	1.09	1.08
Audio Chunks per Gender (M—F)	2446—655	299—87	304—82
Num. Speakers (M—F)	113—85	86—55	195—57

4 Baseline Model

4.1 Wav2Vec2-XLS-R Model

Given the fact that we dealt with a low-resource language, a cross-lingual pretrained model was a good option for our work. Cross-lingual training has shown good efficiency on low-resource languages (Babu et al., 2021), and so we chose the pretrained Wav2Vec2-XLS-R model and fine-tuned it on the VoxMg Malagasy dataset we created. Furthermore, since VoxLingua107 was part of the datasets used for pretraining the Wav2Vec2-XLS-R model and it contained Malagasy speech, this supported our choice in using this model.

Wav2Vec2-XLS-R is pretrained on 436k hours of unlabeled speech data and has three variants depending on the number of parameters: the 300M version, the 1B version, and the 2B version. For this work, we used the 300M version. The other two versions require much more computational resources, but we were limited to the GPU provided by Kaggle [4].

[4] Kaggle: Your Machine Learning and Data Science Community

4.2 TOKENS

The tokenization process consists in splitting a phrase into smaller units called tokens and that could be characters, subwords or words.

We used character-level tokenization to fine-tune the Wav2Vec2-XLS-R model, with a token vocabulary of 45 unique units, as shown in Table 3. The "|" token represents silence in audio, "[UNK]" stands for unknown tokens, and "[PAD]" is a padding token that corresponds to the "blank token" used in Connectionist Temporal Classification (CTC) (Graves et al., 2006; Hannun, 2017).

Table 3: List of tokens used for fine-tuning the Wav2Vec2-XLS-R model.

<table>
<tr><td>|</td><td>’</td><td>-</td><td>a</td><td>b</td><td>c</td><td>d</td><td>e</td><td>f</td></tr>
<tr><td>g</td><td>h</td><td>i</td><td>j</td><td>k</td><td>l</td><td>m</td><td>n</td><td>o</td></tr>
<tr><td>p</td><td>q</td><td>r</td><td>s</td><td>t</td><td>u</td><td>v</td><td>w</td><td>x</td></tr>
<tr><td>y</td><td>z</td><td>à</td><td>â</td><td>ä</td><td>ç</td><td>è</td><td>é</td><td>ê</td></tr>
<tr><td>ì</td><td>î</td><td>ï</td><td>ò</td><td>ô</td><td>û</td><td>ỳ</td><td>[UNK]</td><td>[PAD]</td></tr>
</table>

The Malagasy alphabet consists of 21 letters[5], but since our dataset includes foreign words from English and French, we have additional tokens as shown in Table 3.

4.3 LEXICON

A lexicon is a dictionary that contains words from a language and the mapping of these words to their corresponding pronunciations (phoneme lexicon) or spellings (grapheme lexicon) (Dines & Magimai Doss, 2008).

In this work, we used a grapheme lexicon. During inference, this restricts the model's decoder search space to only words from the lexicon. The lexicon should cover as many words as possible. We built our lexicon by following the steps below:

- Taking all words from the train set;
- Taking all words from the entire Bible;
- Taking all Malagasy radicals and their derivatives from the online Malagasy dictionary "mondemalgache.org";
- Removing duplicates;
- Mapping all words to their corresponding spellings in the format of one word per line, followed by its space-split tokens. For example, "trano t r a n o".

4.4 LANGUAGE MODEL

In the ASR task, a language model (LM) is useful in modeling sequences of words. By assigning probabilities to word sequences, it allows for the selection of the most likely sequence that corresponds to the audio. The LM is an important part of the entire ASR system, especially when dealing with a low-resource language, as it can generally decrease the WER (Nakatani, 2019).

For our LM, we used the corpus from the train set and the entire Bible, excluding those in the test set. We then built 2-grams to 5-grams KenLM (Heafield, 2011) based language models for the experiments.

[5]For more details about the Malagasy alphabet and phonetic system, see Appendix A

5 Experiments and Results

The Wav2Vec2-XLS-R model was trained on the VoxMg dataset using Kaggle's NVIDIA Tesla P100 GPU with 16GB and the Huggingface Transformers library (Wolf et al., 2019). The training lasted for a total of 46000 optimization steps and was performed in several sessions due to Kaggle's GPU usage quota of 30 hours per week with a maximum of 12 hours per session. During each session, the model was trained for 5000 to 7000 steps and a checkpoint was saved at the end to resume training the next session. The batch size was set at 8 with a gradient accumulation of 2, and the AdamW optimizer (Loshchilov & Hutter, 2017) was used with a linear learning rate warm-up from 0 to 3e-04 in the first 500 steps followed by linear decay to zero. We used wandb [6] for monitoring the training progress and saving checkpoints. We employed beam search decoding (Dahlmeier & Ng, 2012; Scheidl et al., 2018). We tuned some of the decoder's parameters individually by testing various values on each one and keeping the others constant. The details are displayed in Table 4. The tuning was performed using the validation set.

Table 4: Beam Search Decoder parameters.

Parameters	Description	Tested values	Best values
lm	n-gram	2-grams, 3-grams, 4-grams, 5-grams	4-grams
beam_size	maximum number of best hypotheses to hold after each decoding step	10, 50, 100, 500, 1500, 3000	1500
beam_size_token	number of tokens to consider for expanding each hypothesis at the decoding step	5, 10, 20, 30, len(tokens)	> 10
word_score	score to add when word finishes	-10, -7, -4, -3, -1, 0, 1, 2, 3, 5	1
lm_weight	weight to assign to the language model score which to accumulate with the acoustic model score for determining the overall scores	1, 2, 3, 4, 5, 10	5

Table 5 presents the performance of the baseline model with and without the LM. The best hyperparameters selected from Table 4 were used.

Table 5: WER on the test set with and without a LM.

Model	WER
Wav2Vec2-XLS-R	44.39
Wav2Vec2-XLS-R + LM	33.43

[6] Weights & Biases: Developer tools for ML

5.1 Discussions

The LM significantly impacted the performance of the model, reducing the error rate by 10.96% on the test set. Although the WER improved, the LM can also worsen the model performance in some cases, such as when high values are attributed to the language model weight. Other factors, such as the specificity of some samples in the dataset and the context covered by the LM, also play a role. To make the LM more robust, adding a more diverse corpus from various domains can help.

On the test set, the base model achieved a WER higher than 50% on 32.64% of the samples. An error analysis on these samples revealed that:

- 42.06% contains foreign words;
- The overall quality of the majority of the audios are very poor;
- Some audios have background music;
- Some audios contain voice overlapping;
- Some segments are unintelligible;
- Some transcription mistakes occurred at the beginning and at the end of some audios. Because of some truncated words at these sections, the transcription was more complex. In some of these cases, predictions are correct but the labels are wrong;
- In some predictions, morphemes are omitted. For example, "ilah" written instead of "ialahy";
- In some predictions, some words are merged together. For example, "tonolona" written instead of "itony olona";
- In some predictions, some morphemes are separated. For example, "hala an'ireo" written instead of "halan'ireo".
- The model struggles to correctly predict certain Malagasy words, such as person names, due to their low ratio in the training set.

The dataset mainly consists of spontaneous speech, which can negatively impact the model's performance, especially with truncated words, mispronounced words, interjections, and unstructured sentences (Ward, 1991). However, this also means that the model can learn to handle spoken language and not just prepared speech.

6 Conclusion and Future Works

In this work, we presented the steps we took to create an Automatic Speech Recognition dataset for Malagasy. We also presented a baseline model which achieved a WER of 33% with a language model. To the best of our knowledge, this is the first ready-to-use Malagasy dataset and ASR system ever built. Our initiative will open doors for many Malagasy researchers and developers. Future works include increasing both the quantity and diversity of the dataset. There are still multiple Malagasy dialects that are not represented in the dataset that we plan to include.

Acknowledgments

We thank Nathanaël Carraz Rakotonirina for providing valuable feedback. We thank the members of LIMA (Laboratoire d'Informatique et de Mathématiques) and the students from MISA (Mathématiques, Informatique et Statistique Appliquées) for discussion, advice and for their valuable contribution to this project.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

Claire Babirye, Joyce Nakatumba-Nabende, Andrew Katumba, Ronald Ogwang, Jeremy Tusubira Francis, Jonathan Mukiibi, Medadi Ssentanda, Lilian D Wanzare, and Davis David. Building text and speech datasets for low resourced languages: A case of languages in east africa. 2022.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*, 2021.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. *arXiv preprint arXiv:1809.01431*, 2018.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7124–7128. IEEE, 2020.

Daniel Dahlmeier and Hwee Tou Ng. A beam-search decoder for grammatical error correction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 568–578, 2012.

John Dines and Mathew Magimai Doss. A study of phoneme and grapheme based context-dependent asr systems. In *Machine Learning for Multimodal Interaction: 4th International Workshop, MLMI 2007, Brno, Czech Republic, June 28-30, 2007, Revised Selected Papers 4*, pp. 215–226. Springer, 2008.

Mark JF Gales, Kate M Knill, Anton Ragni, and Shakti P Rath. Speech recognition and keyword spotting for low-resource languages: Babel project research at cued. In *Fourth International workshop on spoken language technologies for under-resourced languages (SLTU-2014)*, pp. 16–23. International Speech Communication Association (ISCA), 2014.

Mahault Garnerin, Solange Rossato, and Laurent Besacier. Investigating the impact of gender representation in asr training data: a case study on librispeech. In *3rd Workshop on Gender Bias in Natural Language Processing*, pp. 86–92. Association for Computational Linguistics, 2021.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pp. 369–376, 2006.

Awni Hannun. Sequence modeling with ctc. *Distill*, 2(11):e8, 2017.

Kenneth Heafield. Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pp. 187–197, 2011.

Shreya Khare, Ashish R Mittal, Anuj Diwan, Sunita Sarawagi, Preethi Jyothi, and Samarth Bharadwaj. Low resource asr: The surprising effectiveness of high resource transliteration. In *Interspeech*, pp. 1529–1533, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*, 2020.

Roger K Moore. A comparison of the data requirements of automatic speech recognition systems and human listeners. In *Eighth European Conference on Speech Communication and Technology*, 2003.

Tomohiro Nakatani. Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration. In *Proc. Interspeech*, volume 2019, 2019.

John P Openshaw and JS Masan. On the limitations of cepstral features in noise. In *Proceedings of ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pp. II–49. IEEE, 1994.

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

Harald Scheidl, Stefan Fiel, and Robert Sablatnig. Word beam search: A connectionist temporal classification decoding algorithm. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 253–258. IEEE, 2018.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5329–5333. IEEE, 2018.

Zoltán Tüske, Pavel Golik, David Nolden, Ralf Schlüter, and Hermann Ney. Data augmentation, feature combination, and multilingual neural networks to improve asr and kws performance for low-resource languages. In *Fifteenth Annual Conference of the International Speech Communication Association*. Citeseer, 2014.

Jörgen Valk and Tanel Alumäe. Voxlingua107: a dataset for spoken language recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 652–658. IEEE, 2021.

Wayne Ward. Understanding spontaneous speech: The phoenix system. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, pp. 365–367. IEEE Computer Society, 1991.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

A Appendix

A.1 Malagasy Alphabet

Table 6 displays the letters of the Malagasy alphabet, along with their pronunciations based on the International Phonetic Alphabet (IPA). Note that the letters **c**, **q**, **u**, **w** and **x** are not included in the Malagasy alphabet.

Table 6: The 21 letters of the Malagasy Alphabet.

a [a]	b [b]	d [d]	e [e]	f [f]	g [g]	h [h]
i [i]	j [d͡z]	k [k]	l [l]	m [m]	n [n]	o [u]
p [p]	r [r]	s [s]	t [t]	v [v]	y [i]	z [z]

A.2 Phonetic System

- The final vowel **a** is mostly silent, while **e** and **o** are always pronounced at the end of words; e.g. lalana [lalan] (road), rano [ranu] (water).
- The letter **y** is used as a substitute for **i** at the end of words; e.g. vary (rice) instead of vari.
- **h** is almost never pronounced; e.g. ahy [ai] (mine).
- **ts**, **tr** and **dr** are pronounced as [t͡s], [ʈʂ] and [ɖʐ], respectively; e.g. betsaka [bet͡saka] (many), trano [ʈʂanu] (house), andro [anɖʐu] (day).
- **g**, **h** and **k** are palatalized when preceded by **i**; e.g. isika [isikia] (us).
- **ao** and **oa** are pronounced as [o] when they occur within a word, but are pronounced as [aw] and [ua], respectively, when they appear at the end of a word; e.g. taona [tona] (year), toaka [toka] (alcohol), manao [manaw] (faire), roa [rua] (two).
- The grave accent is used for the letters **à** and **ỳ** in Malagasy to indicate stressed syllables; e.g. tanàna (town), tanana (hand), arỳ (there), ary (and).