

hyph-malagasy

Motifs de césure TeX pour le malgache

Documentation méthodologique et résultats de validation

Auteur : RALAHADY Bruno Bakys

ralahadybru@yahoo.fr

2 septembre 2026

Etat du projet

Prototype en developpement actif. Ce document consigne les decisions linguistiques et techniques prises pendant le developpement, ainsi que les resultats d'executions reelles (pdfTeX/LuaTeX, TeX Live 2023) qui les ont validees ou invalidees.

Mise a jour du 2 septembre 2026 : la contamination croisee identifiee lors des premiers tests (section 4) a ete corrigees et revalidee par execution reelle. Le corpus de 29 mots a en outre ete relu par un locuteur natif du malgache, qui a confirme l'ensemble des mots comme corrects, sans aucune correction proposee (voir section 5). La validation par l'Academia Malgache reste a faire.

Structure du projet

- `source/malagasy-hyphenation.txt` -- corpus annote (notation Liang)
- `source/manual-overrides.txt` -- correctifs manuels ponctuels
- `source/generate_patterns.py` -- generation des motifs + format de test
- `scripts/validate.py` -- validation (clusters, synchronisation, couverture)
- `scripts/extract_hyphenation.py` -- extraction des resultats de compilation
- `tex/hyph-mg.tex` -- motifs bruts generes
- `tex/loadhyph-mg.tex` -- loader conforme hyph-utf8 (`\uselanguage`)
- `test/` -- tests pdfTeX (mot par mot) et LuaTeX (contexte de paragraphe)

1. Base phonologique

Les consonnes prenasalisees (mp, mb, nk, ng, nd, ndr, nj, nts) et les digraphes/trigraphes (ts, tr, dr) sont traitees comme des unites phonemiques uniques en attaque de syllabe, jamais scindees par la censure. Base : Dahl (1952), O'Neill (2015), Rakotofiringa (1982).

Le cluster *ntsar* mentionne a l'origine du projet n'a pas ete confirme comme unite autonome dans la litterature consultee -- traite comme n + ts (deja couvert). **A valider avec un locuteur natif ou l'Academia Malgache avant de le considerer regle.**

2. Contraintes TeX confirmees par execution reelle

INITEX obligatoire. `\patterns{}` n'est chargeable qu'en mode INITEX sous pdfTeX (`tex -ini`). Sous LuaTeX cette limitation n'existe pas : chargement dynamique via `lang.new()` / `objet:patterns()`.

Contamination anglaise. `plain.tex` (TeX Live 2023) precharge l'anglais sur `\language0` (181 operations de trie, `lefthyphenmin=2`, `riquiryphenmin=3`) et `\patterns{}` AJOUTE a la trie existante au lieu de la remplacer. Le malgache doit etre isole sur son propre langage.

Parametres globaux. `lefthyphenmin`/`riquiryphenmin` ne sont PAS stockes par langue dans le moteur TeX de base -- a reappliquer explicitement a chaque changement de langue, comme le fait

`\addlanguage{french}{...}{2}{2}` dans `hyph-utf8`.

Pas de numero fixe. `hyph-utf8` enregistre les langues par nom (`\uselanguage`), pas par numero code en dur ; TeX Live assigne le numero dynamiquement.

Accents dans `\directlua`. des caracteres accentues dans un commentaire Lua a l'interieur d'un bloc `\directlua` cassent le comptage de groupe TeX (erreurs 'Missing \$ inserted').

Commentaires Lua dans `\directlua`. plus grave : tout commentaire Lua (`--`) a l'interieur d'un bloc `\directlua` corrompt SILENCIEUSEMENT l'execution du chunk, sans aucune erreur visible. Regle retenue : aucun commentaire `--` a l'interieur d'un bloc `\directlua` ; documenter en TeX (%) juste avant/apres a la place.

Ligatures typographiques. un glyphe ligature (fi -> caractere unique) doit etre decompose via le champ `.components` lors de l'extraction, sinon on recupere un code de police plutot que les caracteres d'origine.

3. Fiabilite de l'extraction

Deux methodes ont ete testees reellement pour recuperer les coupures produites par le moteur TeX :

Methode	Couverture	Limite
<code>\showhyphens{mot}</code> (pdfTeX)	Mot isole	Simple et fiable, mais ne teste pas le comportement en contexte de paragraphe
<code>Overfull/Underfull \hbox</code> (pdfTeX, <code>\tracingparagraphs</code>)	Lignes en depassement seulement	Angle mort confirme : une ligne de badness finie (<code>Loose \hbox</code>) n'a pas son contenu imprime, meme avec <code>\hbadness=0</code>
Callback <code>post_linebreak_filter</code> (LuaTeX)	Toutes les lignes, sans exception	Necessite LuaTeX ; legerement plus complexe a mettre en place

Conclusion retenue : utiliser le callback LuaTeX pour toute validation en contexte de paragraphe ; reserver `\showhyphens` aux tests mot-par-mot rapides.

4. Limite majeure -- contamination croisee (RESOLUE le 2026-09-02)

Le `PatternExtractor` generait initialement toutes les sous-chaines de longueur 2 a 6 de chaque mot annote, sans elagage (pas de logique `patgen`). Sur le corpus de 33 mots, cela produisait 623 motifs bruts.

Test reel ayant revele le probleme : en appliquant ces 623 motifs a une phrase de test, deux mots absents du corpus (*momba*, *malagasy*) se sont retrouves hyphenes a tort (*mo-mba*, *ma-la-ga-sy*), par coincidence de sous-chaines courtes provenant d'autres mots du corpus (ex. *la* extrait de *vola*).

Correction appliquee : `PatternExtractor` n'emet desormais, par default (`whole_word_only=True`), qu'un seul motif par mot, ancre aux deux bouts par le caractere de frontiere (`.mot.`). Un tel motif ne

peut matcher que ce mot exact dans son integralite -- aucune contamination croisee n'est structurellement possible entre mots. Contrepartie assumee : generalisation nulle a des mots absents du corpus (ils restent simplement non hyphenes -- comportement de repli sur, pas une erreur). Le corpus passe de 623 motifs bruts a 29 motifs mot-entier.

Revalidation reelle apres correctif : les 10 mots-temoins du corpus se coupent toujours identiquement ; momba et malagasy restent intacts, en isolation et en contexte de paragraphe ; la sortie en contexte de paragraphe est identique, motif pour motif, a la reference test/expected/sentences.txt.

L'ancien comportement (whole_word_only=False) est conserve dans le code, desactive par default, documente comme dangereux en l'etat. A ne reactiver qu'apres implementation d'un vrai algorithme d'elagage (patgen ou equivalent) sur un corpus nettement plus large avec des exemples positifs ET negatifs explicites.

5. Etat des fichiers de reference -- relecture native effectuee

Le corpus de 29 mots a ete soumis a un locuteur natif du malgache le 2 septembre 2026 via un guide de relecture sans jargon technique. **Resultat : les 29 mots ont ete valides comme corrects, sans aucune correction proposee.**

Cela ne dispense pas d'une validation complementaire par l'Academia Malgache pour une integration officielle dans hyph-utf8, ni n'etend la couverture du corpus au-dela de ces 29 mots -- mais la reserve initiale ("non relu par un locuteur natif") est levee pour ce qui existe actuellement.

lefthyphenmin=2 / righthyphenmin=2 restent des valeurs de test, pas une decision linguistique formellement validee.