Western Papers in Linguistics / Cahiers linguistiques de Western

Volume 1
Issue 1 *Dialectal microvariation in Madagascar*

Article 19

Traitement informatique de variations terminologiques malgaches. Cas du domaine de l'environnement

Tantely Harinjaka Ravelonjatovo
Université d'Antananarivo, tantelyh@gmail.com

TRAITEMENT INFORMATIQUE DE VARIATIONS TERMINOLOGIQUES MALGACHES. CAS DU DOMAINE DE L'ENVIRONNEMENT*

Tantely Harinjaka Ravelonjatovo
Université d'Antananarivo

Cette étude est axée sur la contribution de la terminologie textuelle (Bourrigault et Slodzian, 1999) dans le traitement des variations terminologiques du cas du domaine de l'environnement malgache. La variation (L'Homme, 2004) diffère de la synonymie en ce sens qu'elle est un phénomène relatif au contexte linguistique alors que la synonymie renvoie à des termes morphologiquement différents sans rapport avec le contexte linguistique. Quels sont les apports de la terminologie textuelle dans le traitement de ces variations ? Pour pouvoir y répondre, nous avançons l'hypothèse selon laquelle la variation terminologie est fonction des contextes linguistiques et extralinguistiques. Ainsi, il sera abordé dans la première section le concept de variation selon les deux principales écoles de pensée à savoir la terminologie conceptuelle de vienne et la terminologie textuelle. Viennent ensuite l'explication et application de la méthodologie qui est la terminologie textuelle. La section sur les résultats obtenus dont les méthodes de détection des variations dans un vaste corpus électronique, la typologie de variations terminologiques dans les textes de spécialités maglaches et l'annotation informatique en guise de traitement de ces variations terminera l'article.

1. La variation dans la terminologie

« *La variation terminologique concerne les changements qu'un terme subit dans les textes spécialisés. Ces changements sont fonctions de son utilisation en contexte linguistique* » (LHomme, 2004)

Daille (1994) partage cette même définition et distingue quatre types de variation : la variation graphique qui se manifeste par l'ajout de signe diacritique ou changement de la casse (ex : système expert et système-expert), la variation flexionnelle qui se manifeste par le changement selon le nombre pour les noms et la conjugaison pour les verbes (imprimante à jet d'encre et imprimante à jets d'encre), la variation syntaxiques faibles qui est le changement des éléments grammaticaux des termes complexes (traitement de parole et traitement de la parole) et la variation morphosyntaxique qui est passage d'une partie de discours à une autre sans changement de sens (génétique et génétiquement).

D'un autre point de vue, la variation terminologique peut se manifester différemment. Elle peut être anaphorique en ce sens que le terme est remplacé par un substitut dans un texte. Jacques (2003) parle également de phénomène de réduction qui signifie qu'un terme (complexe) réutilisé dans un texte peut apparaître différemment (réduit) selon le contexte.

Actes du colloque « Variation dialectale à Madagascar ».
Proceedings of the workshop "Dialectal variation in Madagascar".

1.1 La variation selon la terminologie de Vienne

Né du cercle de Vienne, une formation philosophique dont le siège était fondé à Vienne depuis 1923 et dont le but est de fonder une nouvelle science de rigueur basée sur la logique (positivisme logique) et l'expérience physique, la Théorie Générale de la terminologie de Vienne (VGTT) est normative. Ainsi, leurs principes fondamentaux s'appuient sur la nouvelle perception de la philosophie du monde reposant sur les faits ou la tautologie de la pensée et sur le rejet de la métaphysique. Selon Slodzian (2006), la VGTT vise l'unification de la connaissance du monde dans la terminologie et elle cherche à établir la nomenclature universelle de domaine technique tout en se basant sur l'approche philosophique logique. Le principe universaliste de cette théorie justifie son positionnement par rapport au concept de variation.

La VGTT est basée sur l'approche conceptuelle. En effet, l'objet d'étude de la terminologie est le concept, unité stable et non susceptible de variation dans le domaine de spécialité. Ce principe, dénommé « biunivocité », signifie que le terme entretient une relation univoque avec le concept et inversement.

La VGTT, se souciant de l'efficacité de la communication dans un domaine de spécialité évite toute forme de variation. La synonymie et la polysémie sont également mises de côté par les tenants de cette théorie.

1.2 La variation selon la terminologie textuelle

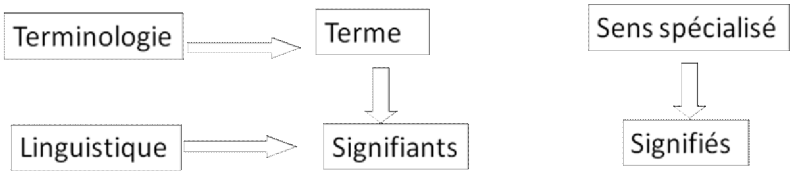
« *La terminologie textuelle, dont le refus du référentialisme est plus ou moins marqué selon les écoles, déplace la problématique de la terminologie aux relations entre signifiés et à la spécificité du fonctionnement des signifiés dans les textes à caractère technique et scientifique; elle s'appuie essentiellement sur les méthodes de la linguistique de corpus pour proposer des listes de candidats termes, sans a priori ontologique* » *(Slodzian, 2006)*

Selon l'auteur suscité, la terminologie textuelle étudie le terme avec une approche différente. Le terme, étant considéré comme une unité linguistique est abordé dans son contexte linguistique naturel attesté dans le corpus. Deux principes forment la terminologie textuelle :

- principe 1 : le texte est le point de départ et est le point d'arrivé de la terminologie.
- principe 2 : le terme est un construit en ce sens qu'il appartient au terminologue, essentiellement textuel, de décider son statut de terme.

Selon le schéma suivant, le terme est considéré comme un signe linguistique à part entière et devrait être analysé comme tel.

(1)



En tant qu'unité linguistique le terme subit naturellement le phénomène de variation en contexte. La terminologie textuelle, à travers l'utilisation de la

linguistique de corpus comme base d'analyse, offre la possibilité d'étudier la variation terminologique. Cette dernière concerne en amont les textes du corpus qui sont issus des différents auteurs et en aval, les variantes terminologiques.

1.3 La démarche de la terminologie textuelle

La démarche appliquée dans la terminologie textuelle est basée sur le texte. Trois étapes fondamentales sont nécessaires dans la terminologie textuelle (L'Homme, 2004) (Ravelonjatovo, 2012) à savoir la constitution de corpus, l'identification de termes et l'étude de terme

La constitution de corpus commence par la collecte des textes spécialisés ou collecte de la documentation qui est une étape fondamentale en terminologie textuelle. Elle consiste en la recherche des textes spécialisés qui contiennent des termes et des informations y afférentes. Elle doit être faite selon des critères linguistiques et extralinguistiques préalablement définis en fonction de l'objectif visé. Les textes collectés doivent être organisés selon les niveaux de spécialité de l'auteur de manière à ce que l'on puisse tenir compte de l'homogénéité et/ou de la variété des textes. A cela s'ajoute l'organisation technique selon les exigences techniques de lisibilité (nom des fichiers, extension des fichiers, nettoyage, etc.)

L'étape suivante est l'identification ou extraction des termes du corpus. Étant donné que le corpus de spécialité est à la fois électronique et volumineux, l'utilisation de logiciel est importante. Selon le logiciel utilisé, l'identification ou extraction peut être entièrement automatique ou semi-automatique. Tous les logiciels concordanciers peuvent servir à l'extraction sémi-automatique tandis que l'extraction automatique nécessite des logiciels spécifiques comme LEXTER, ACABIT, ANA, etc. Les logiciels d'extraction utilisent des méthodes linguistiques formelles et/ou des méthodes statistiques lors de la détection des termes dans le corpus. Il est plus efficace d'utiliser les techniques linguistiques pour la détection des variantes terminologiques.

2. Le corpus *TONTONA* et ses contextes de variation (Ravelonjatovo, 2012)

Le terme « TONTONA » est la contraction du terme *Tontolo iainana.* Il s'agit de l'équivalent malgache du terme français « environnement ». La signification française de l'acronyme est la cagnotte. Le corpus a été constitué pour l'étude systématique des termes du domaine de l'environnement et notamment l'étude de la formation des termes.

Trois critères résument la constitution du corpus : les critères linguistiques/extralinguistiques, les critères terminologiques

Les critères linguistiques, de collecte des textes pour la constitution du corpus TONTONA sont récapitulés dans le tableau suivant :

Tableau 1 : Critères linguistique de constitution

Critères linguistiques	**Critères extralinguistiques**
Langue malgache vs langue étrangère	Textes écrits au format électronique
Langue nationale vs variante régionale	Textes écrits par plusieurs auteurs
Langue synchronique vs langue diachronique	Textes produits par des locuteurs natifs
Langue écrite vs langue orale ou parlée	Textes issus de documents différents
Textes traduits et textes originaux	
Textes complets vs extraits de textes	

Le corpus est constitué de textes écrits en langue malgache. Le malgache est écrit avec le caractère latin. Cela signifie qu'il n'existe officiellement aucune forme de lettre spécifique au malgache. En revanche, par rapport à l'alphabet français et l'anglais, celui du malgache n'a pas les lettres « c », « q », « u », « w » et « x ».

Les 21 lettres ne permettant pas l'écriture de tous les sons, il est ajouté en malgache, officieusement, des lettres accentuées, « à » et « ô », avec les accents grave et circonflexe. L'accent grave, sur la lettre « a » sert à différencier la place de l'accent pour la prononciation d'un mot. Ainsi, pour faire la distinction entre [‘tanana] qui singifie main et [ta'nana] qui signifie village, l'accent grave est mis sur le « a » de la deuxième syllabe pour [ta'nana]. Ainsi, on a *tanana* (main) et *tanàna* (village). L'accent circonflexe, sur la lettre « ô », sert à distinguer les sons [u] et [o] représentés respectivement par les syllabes françaises « ou » et « o ». Comme la lettre malgache « o » correspond à la fois à [o] et à [u], certains utilisateurs choisissent « ô » pour transcrire le [o]. Par exemple, la transcription en malgache des mots empruntés au français « moto » et « modèle » sont pour certains *môtô* et *môdely*. Cette règle officieuse est mise en pratique par un grand nombre d'utilisateurs mais soulève des discussions chez ceux qui réfléchissent à la normalisation et à la standardisation de la langue écrite.

Ces formes d'écritures, normalisées selon les 21 lettres de l'alphabet ou légitimées selon l'usage sont présent dans le corpus TONTONA et constitueraient les premiers aspects de la variation de la terminologie de la langue malgache.

Tableau 2 : Critères terminologiques

Critères linguistiques	**Critères extralinguistiques**
Textes spécialisés vs textes généraux	Textes écrits sur le domaine de l'environnement
	Textes écrits avec des niveaux de spécialités différents
	Textes représentatifs du domaine
	Textes issus de sous-domaines différents

Le domaine de l'environnement est très vaste et souffre de plusieurs acceptions selon les points de vue des chercheurs. L'environnement qui nous concerne ici est en rapport avec le PNAE (Programme National d'Action pour l'Environnement). Depuis 1990, des textes sur l'environnement ont été produits en concert avec ce programme national. Ainsi, les textes collectés sont sur le rapport entre la vie quotidienne des malgaches et leur environnement naturel. En d'autres termes, ils relateraient les projets sur la protection de l'Environnement mis en œuvre à Madagascar pendant 15 ans (1990-2005) et leurs effets auprès des riverains et toute la nation.

Etant donné l'étendue du domaine, les niveaux de spécialisations (L'Homme, 2004) des textes décrits en fonction de ces des auteurs s'imposent. Trois niveaux ont été identifiés :

Niveau 1 - Expert à expert : les textes contenant des termes très spécifiques à un sous-domaine particulier tel que la botanique, biologie, zoologie, etc. Cela concerne par exemple, les articles parus sur Internet dans une revue scientifique des plantes.

Niveau 2 - Expert à un intermédiaire : A ce niveau, LHomme (2004) propose deux niveaux différents ; le niveau expert à expert d'un autre domaine et celui d'expert à un spécialiste en devenir. Ce dernier est appelé également « texte didactique ». Les mémoires de maîtrise sont inclus dans cette catégorie. Ces textes sont les produits écrits par des apprentis-environnementalistes et peuvent être adressés aux autres spécialistes en devenir. On y trouve aussi des textes d'information sur telle technique, tel organisme, telle actualité environnementale intéressant un public plus restreint. Il est à remarquer que le public cible de ces textes n'est pas le grand public. Il se trouve entre le spécialiste du domaine et le grand public non-initié.

Niveau 3 - texte de vulgarisation (texte écrit par un expert ou un non expert à une personne ne possédant pas à priori les connaissances abordées dans le domaine). Ce registre renferme, par exemple, des articles de journaux dans lesquels l'utilisation des termes est dominée par l'emploi des vocabulaires du domaine général. Ce niveau permet d'étudier une partie l'interrelation entre les termes utilisés en environnement et en développement durable. En effet, le côté économique est dominant dans le développement durable par rapport à l'environnement. A titre d'exemple, le terme « environnement » appartient-il au terme « développement durable » ou inversement ? En d'autres termes, environnement est-il l'hyperonyme de « développement durable » ou inversement? La pareille question se posera sur les termes contenus dans les deux sous-ensembles.

Le corpus TONTONA a été organisé selon ces variétés observées en amont. D'autres critères physiques liés au format électronique des documents ont été également pris en considération. Les fichiers sources sont classés dans un répertoire tandis que les fichiers txt contenant des textes bruts nécessaires à l'exploration se trouvent dans un autre répertoire. L'organisation physique du corpus est récapitulée dans le tableau ci-après.

Tableau 3 : Récapitulatif de TONTONA

TONTONA			TOTAL
CORPUS			1 corpus
SCIENTIFIQUE	JURIDIQUE	JOURNALISTIQUE	3 sous corpus
56 textes 33 auteurs	09 textes 09 auteurs	15 textes 11 auteurs	80 textes 52 auteurs
178324 mots	384213 mots	8812 mots	571349 mots

Ce tableau montre que l'on peut admettre que le corpus TONTONA contient en amont trois contextes de variation à savoir le contexte linguistique (langue malgache et ses variations dans l'écriture), le contexte terminologique (les sous domaines de l'environnement et les niveaux de spécialités) et le contexte extralinguistique (variations relatives aux caractères informatiques des textes).

Le contexte linguistique se rapporte aux variations de discours utilisé dans chaque texte. Le discours utilisé dans le sous corpus journalistique est différent de ceux utilisés dans les sous corpus juridique et scientifique. En ce qui concerne le contexte terminologique, les formes de termes utilisés peuvent variés d'un auteur à l'autre, d'un sous-domaine à l'autre. Le contexte extralinguistique se rapporte au nombre de mots pour chaque sous-corpus et/ou texte, la longueur et la taille de chaque texte. Le corpus, réservoir à termes, est également un conteneur de variations et micro-variations. On y voit le bon usage et faux usage, les normes et écarts, les termes et leurs variantes, etc.

3. Types de variations terminologiques

Compte tenu du fait qu'il n'existe pratiquement pas de logiciel malgache d'extraction automatique des termes, il nous a été nécessaire de passer par l'exploration semi-automatique de corpus pour le repérage de termes. Pour cela, le logiciel Nooj a été utilisé. Il s'agit d'un logiciel TALN (Traitement Automatique de Langues Naturelles) ou NLP (Natural Langage Processor). L'objectif classique en est l'exploration et traitement automatique de textes. Malgré la puissance et la mise à jour systématique de Nooj, son efficacité et la vitesse du traitement sont conditionnées par le prétraitement des textes sur lesquels on veut travailler.

Pour chaque requête, nous avons commencé par rechercher le mot exact, comme motif, puis nous avons observé son contexte linguistique immédiat dans un texte avant l'observation dans d'autres textes.

Cette méthode d'exploration terminologique nous a permis d'identifier plusieurs termes et variantes terminologiques. Dans cet article, nous allons nous focaliser surtout sur les types suivants : la variation graphique, la variation syntaxique, la variation morphosyntaxique et la variation lexicoterminologique

3.1 La variation (ortho)graphique

Adaptée à la langue malgache, nous avons mis variation « (ortho)graphique » au lieu de variation « graphique » car cela comprend à la fois le changement de la casse, les typographiques et les erreurs orthographiques.

(2) fitantanana
Fitantanana

FITANTANANA
'gestion'

Le changement de la casse est important dans la mesure où le logiciel doit tenir compte de toutes les formes d'un terme aussi bien en majuscule qu'en minuscule. Le cas est très fréquent dans les titres de paragraphe mais concerne également l'acronyme et son développement. Par exemple :

(3) VOI
Vondron'Olona Ifotony
'COBA : COmmunauté de BAse'

En ce qui concerne le signe diacritique, en malgache, il est utilisé différemment et tient une fonction syntaxique dans la composition.

(4) voa-janahary
voajanahary
'naturel'

Le trait d'union remplace ici la dernière syllabe *–tra* de *voatra* qui tombe en contact avec le deuxième mot *zanahary*. Il résulte du même contact la variation combinatoire *z~j* pour *zanahary* et *janahary*. Sur le plan syntaxique, le terme complexe *voa-janahary* (lit. 'naturel) signifie *namboarin'ny zanahary* (lit. 'fabriqué par Dieu')

(5) voatr'olombelona
voatrolombelona
'artificiel'

Le même phénomène se produit dans l'exemple ci-dessus. L'apostrophe remplace la lettre *-a* qui tombe après le contact avec le mot *olombelona*. La variation combinatoire ne concerne pourtant pas les voyelles et *o-* de *olombelona* reste inchangé.

3.2 La variation syntaxique

La variation syntaxique se manifeste par le changement de la première base lexicale et/ou de la deuxième base lexicale. Par exemple

(6) a. faritra voajanahary
faritra voary
'zone naturelle'
b. harena nomen-janahary
Harena voajanahary
'Patrimoine naturel'

Le changement concerne uniquement la deuxième base lexicale. Par contre, les deux exemples suivants font preuve de changement de la première base.

(7) harena voajanahary azo havaozina
loharanon-karena voajanahary azo havaozina
'ressources naturelles renouvellables'

Le phénomène est plus complexe et difficile à identifier par un logiciel quand le changement touche aussi bien la première que la deuxième base lexicale. Par exemple :

(8) Harena voajanahary mety havaozina
loharanon-karena voajanahary azo havaozina
'ressources naturelles renouvellables'

3.1 La variation morphosyntaxique

Dans la variation morphosyntaxique, on assiste au changement de partie de discours en cotexte alors que le terme reste le même dans le même domaine de l'environnement. Par exemple :

(9) maharitra (adjectif) > fitantanana maharitra
'durable' dans gestion durable
maharitra (adverbe) > mitantana maharitra
'durable' dans 'gère de manière durable'

3.1 La variation lexicoterminologique

En revanche, on assiste à la variation lexicoterminologique quand le phénomène est lié à la terminologisation ou determinologiation. En effet, le terme passe d'un domaine à un autre ou du domaine général à un domaine de spécialité. Par exemple :

(10) manokana (adjectif) > tahiry manokana
'special (adjectif)' > reserve special (domaine : environnement)
manokana (verbe) > mila manokana vola (domaine général)
classer (verbe) > il faut classer (mettre de côté) de l'argent

Dans l'exemple ci-dessus, on assiste à un cas de terminologisation où création de nouveau terme de l'environnement (manokana) à partir d'un terme existant dans le domaine général (manokana). Le problème devient plus complexe car ces deux mots ayant des sens différents se trouvent tous les deux dans le corpus dans des contextes différents. Le logiciel saurait-il distinguer le terme de l'environnement du mot du domaine général ?

4. Les apports du TAL pour le traitement des variations terminologiques

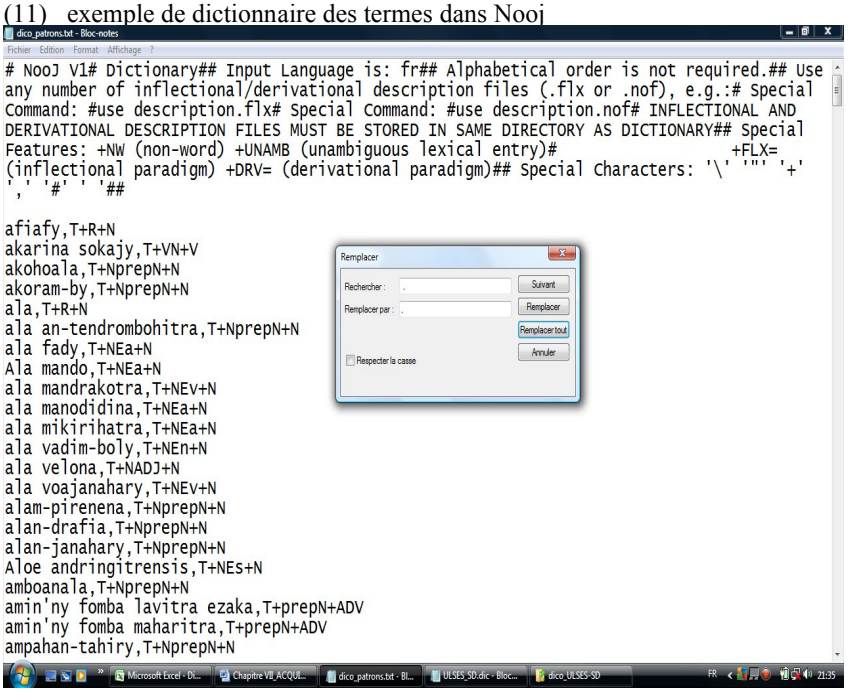
En quoi le logiciel TAL peut-il aider le terminologue à traiter les variations terminologiques? Deux solutions sont proposées dans cet article à savoir la création d'un dictionnaire des termes, la création d'un dictionnaire de lemmatisation.

4.1 Dictionnaire des termes

Nous entendons par dictionnaire des termes, au sens se rapportant au traitement informatique de langue, une liste des termes du corpus avec la description de sa catégorie grammaticale et celle des contextes/situation de communication. Le dictionnaire des termes permet l'analyse contextuelle des termes dans l'ensemble mais également l'analyse des occurrences linguistiques et statistiques des termes dans le corpus et dans chaque sous-corpus.

En tant que dictionnaire électronique, le dictionnaire doit respecter la syntaxe du logiciel auquel il va être implémenté. Pour Nooj, la forme est illustrée par le schéma suivant :

(11) exemple de dictionnaire des termes dans Nooj



La situation de communication peut s'ajouter après le troisième élément de la description. Ainsi, lors de la requête on peut commencer par rechercher tous les termes en tapant la lettre « T » suivi du code de description de la situation de communication. Cela permet de trouver si un terme apparaît sans variation dans les trois sous-corpus différents (juridique, scientifique, journalistique).

On peut également ajouter dans le dictionnaire les informations sur le niveau de spécialisation de l'auteur des textes ou des textes.

4.2 Dictionnaire de lemmatisation

L'objectif de la lemmatisation est que le logiciel dispose d'une seule forme canonique qui peut correspondre dans le syntagme à une ou plusieurs formes linguistiques. Certaines langues comme le français et l'anglais ont connu beaucoup de projet de lemmatisation d'une façon ou d'une autre. En effet, le fait que le choix de lemme est arbitraire implique que chaque langue peut adopter son système de lemmatisation. Par exemple, le verbe français a comme lemme la forme à l'infinitive, le nom a comme lemme la forme au masculin singulier. Les autres langues où il n'y pas de variation flexionnelle par rapport au genre et nombre adoptent d'autre système de lemmatisation de leur lexie. Pour le cas de variation terminologique malgache, on peut adopter le format suivant :

(12) a. Lemme : variante1+variante2+variante3+variante4

Lemme : forme canonique relatif à un terme comme entrée du dictionnaire
Variante 1 : variante (ortho)graphique
Variante 2 : variante syntaxique
Variante 3 : variante morphosyntaxique
Variante 4 : variante lexicoterminologique

Par exemple :

b. Harena voajanahary, harena voa-janahary+harena nomen-janahary+E+E
Terme : Harena voajanahary
Variante (ortho)graphique : harena voa-janahary
Variante syntaxique : harena nomen-janahary
Variante morphosyntaxique : vide
Variante lexicoterminologique : vide

4.3 Conclusion

La variation terminologique est inhérente à toutes les langues spécialisées du monde y compris la langue malgache. L'utilisation de la terminologie textuelle basée sur la linguistique textuelle ne fait que corroborer ce propos. Appliquée à la langue malgache, cette approche permet de confirmer que la variation terminologique est fonction des contextes. Quatre types de variations ont été identifiés à savoir la variation (ortho)graphique, la variation syntaxique, la variation morphosyntaxique et la variation lexicoterminologique. Grâce aux dictionnaires implémentés dans Nooj, il est possible d'identifier aussi bien les termes que leurs variantes.

Références

Bourrigault, Didier, et Slodzian, Monique. 1999. Pour une terminologie textuelle. Dans *Terminologies nouvelles 19 Spécial TIA99*. 29-32.

Daille, Béatrice. 1994. *Approche mixte pour l'extraction automatique de terminologie : statistique lexicale et filtre linguistique.* Doctorat en Informatique, Université Paris 7.

Jacques, Marie Paule. 2003. *Approche en discours de la réduction des termes complexes dans les textes spécialisés.* Thèse de doctorat en Sciences du langage, Université de Toulouse II.

L'Homme, Marie Claude. 2004. *Terminologie : principes et techniques.* Montréal : Presse Universitaire de Montréal, 2004.

Ralalaoherivony, Baholisoa Simone. 2004. *Création d'outils méthodologiques pour la déscription du malgache en vue de sa modernisation.* Habilitation à diriger des recherches, Université d'Antsiranana.

Ravelonjatovo, Tantely. 2012. *Constribution à la methodologie d'analyse systematique des termes malgaches. Cas du domaine de l'environnement*. Thèse de doctorat en linguistique appliquée, Université d'Antananarivo

Slodzian, Monique. 2006. *La terminologie, historique et orientations*. Dans *Textes et Connaissances*. Résumé de la conférence dans le cadre de la journée.