Dictionnaire électronique (DE) des noms simples issus de verbes

Joro Ny Aina Ranaivoarison

Dictionnaire électronique (DE) des noms simples issus de verbes Les noms issus des alternances *mp-* ou *f-*

Joro Ranaivoarison
Université d'Antananarivo
Centre Interdisciplinaire de Recherche Appliquée au Malgache
Madagascar
jororanaivo@yahoo.fr

Résumé

Cet article décrit la construction d'un dictionnaire électronique de noms issus de verbes du malgache (DEMA-NVS). Ces noms se composent de noms d'agent, de noms de profession, de noms de manière, de noms d'instrument, de noms d'action et de noms exprimant un état. Les structures morphologiques de ces derniers sont détaillées puis décrites à l'aide de transducteurs afin de construire une ressource destinée à des utilisations informatiques – un dictionnaire électronique. On discute dans cet article de la mise en œuvre du dictionnaire, du dictionnaire électronique lui-même et de son évaluation en rapport avec sa couverture lexicale.

Mots-clés : dictionnaire électronique, ressource linguistique, morphologie, malgache, nom

1 Introduction

Ce travail se situe à l'interface entre morphologie descriptive et traitement automatique des langues (TAL). Son objet est le malgache, une langue « peu dotée » en outils et ressources au sens de Berment (2004). Pour développer des outils de TAL qui rendent possible le traitement automatique de cette langue et permettre aux utilisateurs de disposer des moyens pour communiquer dans leur langue, il est nécessaire d'augmenter la couverture lexicale actuelle de celle-ci. En effet, il sera plus facile pour les développeurs d'applications de décider ou non de créer d'applications pratiques (correcteur grammaticale et/ou orthographique, outil d'aide à la traduction) pour le malgache si les ressources crées pour celui-ci ont une couverture correcte de ses lexiques, c'est-à-dire que tous les mots de la langue, du moins ceux se rattachant aux grandes catégories grammaticales (verbes, noms, adjectifs, adverbes, pronoms, etc.) sont insérés dans les ressources. L'objectif de cet article est de construire un dictionnaire électronique (DE) des « noms simples »[1]. Notre travail en cours porte sur 3 200 lemmes verbaux dont nous recensons des dérivés exprimant un état ou servant de noms d'agent, de profession, de manière, d'instrument ou d'action.

Le malgache est une langue agglutinante avec une riche morphologie, qu'il s'agisse de formes fléchies ou de formes dérivées. Dans cet article, une partie de la morphologie nominale est exposée. En effet, dans cette langue, il y a les mots qui sont eux-mêmes "noms" (N) comme *angady* "bêche, pelle", *trano* "maison", *penina* "stylo", *bara* « A. barre qui sépare les mesures en musique. B. Traverse, pièce mise en travers », *baby* « A. Épi de maïs sur la tige. B. Action de porter sur le dos. ». Ensuite, il y a les noms qui sont issus des alternances de l'élément temporel des verbes (V) avec *f-* ou *mp-* comme dans *mpijery* **N.** "spectateur", *fijery* **N.** "manière de regarder", *fijerena* **N.** "action de regarder" issus respectivement des verbes *mijery* **V. actif-statif** (act.-stat.) "regarder" et *ijerena* **V. circonstanciel** (circ.) "regarder". Enfin, il y a les noms qui sont issus des adjectifs (A) comme *hatsara* **N.** ou *fahatsara* **N.** « l'état de ce qui est bon, beau », *hatsarana* **N.** ou *fahatsarana* **N.** « la bonté, la beauté" issus de l'adjectif *tsara* **A.** "bon, qui a de bonnes qualités, beau ». Dans ce qui suit, seuls les noms issus des alternances du préfixe de temps avec *mp-* ou *f-*, qui sont des préfixes formatifs de nom, sont discutés. Ces noms sont issus de formes verbales comme *milalao* **V. act.-stat.** « jouer » dont dérivent *mpilalao* **N.** « joueur » (nom d'agent) et *filalao* **N.** « manière de jouer » (nom de manière) ; ou comme *anendrikendrehana* **V. circ.** « calomnier » dont dérive *fanendrikendrehana* **N.** « action de calomnier » (nom d'action). Toutes les fois que le terme "noms" est utilisé dans ce qui suit, il désigne les noms issus de cette formation.

Dans ce papier, les caractéristiques morphologiques des noms puis les méthodes utilisées (Gross, 1989) pour construire le dictionnaire sont présentées.

[1] Un dictionnaire électronique des verbes contenant 3 200 radicaux verbaux pouvant générer plus de 60 000 formes verbales a été déjà réalisé (Ranaivoarison *et al.*,2013, 2015a, 2016).

Par la suite est décrite la construction des graphes nécessaires au bon fonctionnement du dictionnaire avec Unitex, une plateforme de traitement de corpus écrits par dictionnaires et grammaires (cf. Paumier, 2016). Le dictionnaire électronique des noms issus de verbes simples (DEMA-NVS) et celui des paradigmes flexionnels des radicaux verbaux formant des noms simples (DEMA-NVSflx) sont ensuite présentés, ainsi que les résultats de leur évaluation.

2 Caractéristiques morphologiques des noms

Rajaona (1972, p. 642 - 645) présente les grandes lignes de la structure morphologique des noms issus des alternances du préfixe de temps avec *mp-* ou *f-*, préfixe formatif de noms, en malgache. Généralement, ces noms sont :

- soit des noms d'agent (Nag) ou de profession (Nprof),
- soit des noms de manière (Nman) ou d'état (Nét),
- soit des noms d'instrument (Ninst),
- soit des noms d'action (Nact).

Les noms d'agent et de profession sont à préfixe *mp-* se combinant avec les affixes de l'actif-statif[2] – les affixes de l'actif-statif sont : *i-*, *a-* ou une de ses variantes *an-*, *am-*, *ana-* apparaissant entre le préfixe de temps et le radical (cf. Rajaona, 1972, p. 454) – comme *mpijery* « celui qui regarde » analysé *mp-i-jery*, *mpandraharaha* « administrateur » analysé *mp-an-draharaha*, *mpamoha* « celui qui réveille, qui fait lever » analysé en *mp-am-oha* où *i-*, *an-*, *am-* sont des préfixes à valeur d'actif-statif. Les noms de manière, d'instrument et d'état sont à préfixe *f-* se combinant pareillement avec les affixes de l'actif-statif comme *fanafaingana* « manière d'accélérer » analysé *f-ana-faingana*, *famaky* « hache » analysé *f-am-aky*, *fihanjahanja* « l'état de ce qui est nu » analysé *f-i-hanjahanja* où *ana-*, *am-*, *i-* sont des préfixes de l'actif-statif. Enfin, les noms d'action se forment également sur *f-* avec des affixes à valeur de circonstanciel – les affixes à valeur de circonstanciel sont les affixes parasynthétiques du type *x-…-ana* où *x-* est un préfixe de l'actif-statif (cf. Rajaona, 1972, p. 159) – comme *fivoriana* « réunion, assemblée, séance » analysé *f-i-vori-ana*, *fihantsiana* « action de provoquer » analysé *f-i-hantsi-ana*, *fiverenana* « action de retourner » analysé en *f-i-veren-ana* où l'affixe parasynthétique *i-…-ana* est à valeur de circonstanciel.

Il s'ensuit que *mp-* est un préfixe formatif de noms d'agent et de profession[3] ; et, *f-* peut être :

- soit un préfixe formatif de noms de manière, d'instrument et d'état (quand il se combine avec les affixes de l'actif-statif)
- soit un préfixe formatif de noms d'action (quand il se combine avec les affixes du circonstanciel).

Pour aboutir à une description linguistique précise de chaque élément verbal pouvant former des noms, ces informations linguistiques[4] fournies par Siméon Rajaona (1972), en plus des informations sur les variations de formes des lemmes, sont codées et insérées dans le dictionnaire servant à une analyse morphologique claire et précise des noms de la langue.

3 Codification des noms simples

La morphologie à deux niveaux (Koskenniemi, 1983) a été largement utilisée pour traiter les langues agglutinantes telles que le finnois (Koskenniemi et Church, 1988), le turc (Oflazer, 1993) et même le malgache (Dalrymple *et al.*, 2006). Dans notre approche du traitement automatique du malgache, les méthodes analogues à celles utilisées pour le coréen (Nam, 1994 ; Nam et Paumier, 2014) ont été adoptées. Ces méthodes reposent sur des lexiques construits manuellement par des linguistes et ne sont pas à base de règles de calcul. Si les méthodes à base de calcul et/ou de statistiques ont l'avantage d'être économiques, les méthodes par dictionnaire sont précises et ont l'avantage d'être souples en ce qui concerne la maintenance et la mise à jour. Notre méthode de travail s'inscrit dans cette deuxième catégorie.

Elle se fonde sur les travaux de Gross (1989). La méthode se base sur une description explicite et détaillée de chaque mot de la langue. Rakotoalimanana (2000) mentionne cette approche. Sa description du malgache est explicite et claire et couvre tous les niveaux d'analyse (phonétique, morphologie, syntaxe, sémantique) et toutes les catégories grammaticales en allant dans les détails des découpages des affixes. Cependant, il ne mentionne que quelques exemples de variations morphologiques des mots, et ne vise pas une couverture lexicale substantielle. Par exemple, pour les verbes, son modèle ne prévoit pas d'indiquer pour chaque lemme verbal à quelle voix il peut apparaitre, ni quels affixes il prend parmi ceux affectés à chaque voix. Ce modèle ne prévoit donc pas de façon fiable le découpage morphologique de tous les mots, et il accepte des formes inconnues du malgache.

[2] L'actif-statif et le circonstanciel sont deux des valeurs que peut prendre la voix, une catégorie morphologique, au sens où on parle de voix active et passive en français. Lorsque le verbe passe de la voix active-stative à la voix circonstancielle, un complément circonstanciel passe parallèlement dans la position de sujet. Le malgache possède cinq voix (Ranaivoarison, 2016, p. 98).

[3] Et quelquefois un préfixe formatif de noms exprimant une habitude (Nhab) comme *mpidanadana* « ce qui reste habituellement ouvert ».

[4] Ces informations linguistiques ont été reprises telles quelles pour formaliser la catégorie grammaticale des noms. En effet, elles ont été suffisamment complètes, explicites et cohérentes pour pouvoir les utiliser dans le traitement automatique des langues.

Nous avons choisi de combler cette lacune en recensant systématiquement, d'une part, les variations morphologiques des lemmes, et d'autre part les combinaisons d'affixes avec ces variantes. Dans la pratique, notre description formelle prend la forme de deux activités : la codification de propriétés (catégorie grammaticale, combinaison d'affixes, variation de formes) et la construction de graphes (transducteurs de flexion et grammaires locales). Avant d'aborder la construction des graphes (section 4.), la codification effectuée pour construire le DE des noms est d'abord présentée dans cette section. Premièrement, la codification des catégories grammaticales et valeurs des préfixes formatifs de noms est abordée. Puis sont abordées respectivement la codification des combinaisons des affixes (classes affixales) et des variations de formes des radicaux (classes radicales).

3.1 Codification des catégories grammaticales et valeurs des préfixes formatifs de noms

Les catégories grammaticales et sémantiques qui entrent dans la construction du DE des noms issus de verbes sont listées ci-dessous.

PFN	Préfixes formatifs de noms
PV	Préfixes de voix
SV	Suffixes de voix
V	Verbes
:g	noms d'agent et de profession
:m	noms de manière et d'état
:n	noms d'instrument
:t	noms d'action

3.2 Codification des classes affixales

Une classe affixale est une classe de lemmes qui ont en commun la façon dont ils se combinent avec des affixes. Les codes de classes affixales des noms sont composés de trois cases.

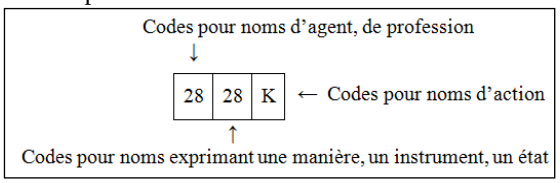

Figure A : Schéma général des codes de classes affixales des noms

- La première case indique les noms d'agent et de profession à préfixe *mp-* + préfixe de l'actif-statif, comme *mpanendy* analysé *mp-an-endy* « celui qui fait frire », *mpanjono* analysé *mp-an-jono* "pêcheur", *mpamboly* analysé en *mp-am-boly* « celui qui plante, jardinier, cultivateur ».
- La deuxième case indique les noms de manière, d'instrument et d'état à préfixe *f-* + préfixe de l'actif-statif, comme *fanadala* analysé *f-an-adala* "manière de duper", *fitaratra* analysé *f-i-taratra* "miroir", *fangatsiaka* analysé *f-an-gatsiaka* « l'état de celui qui a froid, de ce qui est froid »
- La troisième case est celle des noms d'action à préfixe *f-* + préfixe du circonstanciel, comme *fivahinianana* analysé *f-i-vahinian-ana* "action de voyager, de séjourner", *fanendasana* analysé *f-an-endas-ana* « action de faire frire, de rôtir, de griller ; poêle, marmite, rôtissoire », *fanabeazana* analysé en *f-ana-beaz-ana* « action d'agrandir, d'augmenter, d'élever, d'éduquer ».

La première et la deuxième cases ne peuvent recevoir que des chiffres et la troisième case des lettres en majuscules. Le code « v » est le seul utilisé pour chacune de ces trois cases si la case indique une absence de termes de noms d'agent, de profession, de manière, d'instrument, d'état ou d'action. Ci-dessous, ces types d'informations sont développés dans cet ordre.

3.2.1 Codes de noms d'agent et de profession à préfixe *mp-* + actif-statif

Les noms à préfixe *mp-* sont obtenus par alternance du préfixe de temps[5] avec *mp-*, préfixe formatif de nom d'agent et de profession. D'une manière générale, ces éléments obtenus par alternance reposent sur la voix active-stative comme dans *manjono* **V.act.-stat.** « pêcher » / *mpanjono* **N.** « pêcheur », *miady* **V.act.-stat.** « combattre » / *mpiady* **N.** « guerrier, combattant », *manafaingana* **V. act-stat.** « accélérer » / *mpanafaingana* **N.** « celui qui accélère ». Ci-après les codes de combinaison des affixes de l'actif-statif se combinant avec *mp-*.

Codes	Affixes
1	*Ø-*
2	*i-*
3	*an-*
4	*ana-*
7	*i-/an-*
21	*i-/an-/ana-*
23	*a-*
26	*i-/am-*
28	*i-/ana-*
30	*an-/ana-*
32	*ana-/anka-*
37	*anam*

Codes	Affixes
38	*i-/Ø-*
43	*am-*
60	*i-/am-/ana-*
61	*anka-*
62	*aha-*
63	*am-/ana-*
65	*an-/ana-/ian-*
66	*i-/a-*
67	*a-/anka-*
68	*Ø-/an-*
69	*i-/anam-*

Tableau 1 : Codes utilisés pour les noms formés sur l'actif-statif

Ces codes se placent en première position dans la chaîne des codes et sont composés uniquement de chiffre.

Si le radical à l'origine des noms ne fournit pas de noms d'agent et de profession alors un code "v" est utilisé pour marquer cet absence comme pour *màndra* "épuiser une terre par une incessante production" qui

[5] Les préfixes de temps dont il s'agit ici sont ceux combinables avec l'actif-statif, c'est-à-dire /*m-* « présent » : *n-* « passé » : *h-* « futur »/ comme pour *lèha* « marcher » : *mandeha* au présent, *nandeha* au passé et *handeha* au futur.

a pour code v4E, la langue n'atteste pas l'existence du nom d'agent ou de profession **mpanamondra* mais fournit les formes comme *fanamondra* "manière d'épuiser la terre par une incessante production" (Nman) et *fanamondrana* "action d'épuiser la terre" (Nact) .

3.2.2 Codes de noms de manière, d'instrument, d'état à préfixe *f-* + actif-statif

Les mêmes codes de l'actif-statif utilisés au 3.2.1 sont utilisés pour former les noms de manière, les noms d'instrument et les noms exprimant un état. Les noms comme *fijery* **N.** « manière de regarder » issu de *mijery* **V. act.-stat.** « regarder », *fiendrinendrina* **N.** « l'état de stupidité » issu de *miendrinendrina* **V. act.-stat.** « être stupide », *fihogo* **N.** « peigne » issu de *mihogo* **V. act.-stat.** « peigner, se peigner » sont respectivement des noms exprimant une manière, un état, un instrument. En effet, les deux formations, l'une avec *mp-* et l'autre avec *f-* reposent toutes deux sur les affixes de l'actif-statif. Il s'ensuit que cette deuxième case est renseignée également par les chiffres présentés au tableau 1.

Si cette deuxième case n'est pas renseignée pour une entrée donnée alors elle est renseignée par le code "v" comme pour *hèry* 2 "1. A. Être fort, courageux, puissant, brave, zélé, faire bien, faire beaucoup. B. Gagner, l'emporter, vaincre, avoir un excédent, un surplus. 2. Rendre fort, fortifier, encourager. 3. Devenir fort, se fortifier, prendre courage" qui a pour code 67vXX, la langue n'atteste pas l'existence des noms de manière ou d'état **fahery* ou **fankahery* mais fournit les formes *mpahery* "habituellement vainqueur, un brave" (Nhab) [6] , *mpankahery* "celui qui fortifie" (Nag), *faherezana* "le courage, la force, la vigueur, l'entrain" et *fankaherezana* "action de fortifier" (Nact).

3.2.3 Codes de noms d'action à préfixe *f-* + circonstanciel

Les noms d'action sont formées sur le préfixe *f-*, préfixe formatif de noms, se combinant avec les affixes du circonstanciel comme *filalaovana* **N.** « action de jouer » issu du circonstanciel *ilalaovana* **V. circ.** « jouer », *fanadihadiana* **N.** « action de scruter, information » issu du circonstanciel *anadihadiana* **V. circ.** « scruter », *fieritreretana* **N.** « action de réfléchir » issu du circonstanciel *ieritreretana* **V. circ.** « réfléchir » . Ils sont obtenus par alternance du préfixe de temps[7]avec *f-*. Les codes des préfixes de la voix circonstancielle sont résumés dans le tableau ci-contre.

Codes	Affixes
A	*Ø-*
B	*i-*
C	*am-*
D	*an-*
E	*ana-*
F	*Ø-/an-*
G	*aha-*
H	*i-/Ø-*
I	*i-/am-*
J	*i-/an-*
K	*i-/ana-*

Codes	Affixes
L	*am-/ana-*
N	*an-/ana-*
O	*i-/an-/ana-*
S	*i-/am-/ana-*
T	*i-/an-/aha-*
U	*a-*
W	*i-/a-*
Z	*i-/anam-*
CC	*ana-/anka-*
XX	*a-/anka-*
ZZ	*an-/ana-/ian-*

Tableau 2 : Codes utilisés pour les noms formés sur le circonstanciel

Si cette troisième case n'est pas renseignée pour une entrée donnée alors elle est renseignée par le code "v" comme pour *zò* "tomber sur" qui a pour code 33v, la langue n'atteste pas l'existence du nom d'action **fanjoana* mais fournit les formes *mpanjo* "ce qui tombe sur" (Nag) et *fanjo* "manière de tomber sur" (Nman).

3.3 Codification des classes radicales

Une classe radicale est une classe de lemmes qui ont en commun la façon dont varie leur radical. Les codes de classes radicales des noms sont composés de trois cases comme pour les verbes (Ranaivoarison, 2016, p. 218). Ces mêmes codes de classes radicales employés pour les verbes sont réutilisés car les noms sont également issus de verbes. Ci-dessous les principes utilisés pour ces codes sont résumés.

- La première case désigne les finales des radicaux verbaux qui peuvent être « 0 », « 1 », « 2 » ou « 3 ».
- La deuxième case désigne la compatibilité des radicaux verbaux avec le suffixe *-ina* et peuvent être « a » ou « i ».
- La troisième case indique les phénomènes[8] qui peuvent apparaître au niveau des radicaux verbaux lorsque ceux-ci sont entrent en contact avec les affixes.

Les codes des classes radicales sont introduites par la lettre V désignant les verbes. Ils sont aux alentours de 170 correspondant à des transducteurs de flexion (4.1) qui permettent de générer les paradigmes flexionnels et les relier aux affixes.

4 Construction des graphes de noms

Deux types de graphes sont associés aux codes de classes affixales et codes de classes radicales. Ces deux types de graphes sont présentés ci-après en exa-

[6] Voir note 3.

[7] Les préfixes de temps dont il s'agit ici sont ceux combinables avec le circonstanciel, c'est-à-dire *Ø-* « présent »/*n-* « passé »/*h-* « futur » comme pour *lèha* « marcher » : *andehanana* au présent, *nandehanana* au passé et *handehanana* au futur.

[8] Ces phénomènes sont par exemple de phénomènes de suppression ou de remplacement de la première lettre d'un radical, d'insertion d'une lettre au début ou d'utilisation d'un élargissement, etc.

minant premièrement ceux qui sont rattachés aux codes de classes radicales et deuxièmement ceux rattachés aux codes de classes affixales.

4.1 Transducteurs de flexion

Les transducteurs de flexion sont les graphes qui se rattachent aux codes de classes radicales. Ils fournissent à l'aide du programme de génération de formes d'Unitex les variantes morphologiques des radicaux formant des noms. Pour un radical comme *lèha* « marcher » par exemple, le transducteur de flexion V0ibe permet de générer automatiquement les variantes morphologiques de *lèha* comme *dèha*[9] dans *mpandeha* « voyageur, passant » ou dans *fandeha* « manière de marcher, démarche », et comme *dehán* dans *fandehanana* « action de marcher, marche, chemin » en indiquant les affixes qui vont avec les variantes. Ci-après, le graphe de transducteur de flexion V0ibe est fourni.

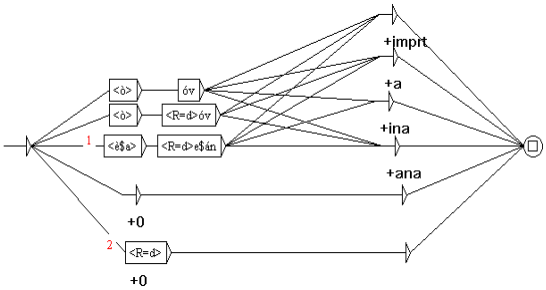

Figure B : Transducteur de flexion V0ibe

Les chemins 1 et 2 dans la figure B. permettent de générer les formes *dehán* et *dèha* ; ils indiquent respectivement les affixes avec lesquels ils se combinent. Le chemin 1 fournit par exemple la forme *dehán* et lui associe une propriété codée **+ana** indiquant qu'il se combine avec l'affixe *-ana* et se retrouve dans la forme *fandehanana* « action de marcher, marche, chemin » pour les noms. Les autres propriétés (**+imprt**, **+a**, **+ina**) pour ce chemin sont utilisées pour les formes verbales (Ranaivoarison, 2016, p. 227.). La boîte avec **+0** indique qu'après la variante morphologique il n'y a plus de suffixe comme dans le chemin 2 (Fig. B). En effet, après la variante morphologique *dèha*, il n'y a plus de suffixe, comme dans les formes nominales *mpandeha* « voyageur, passant » et *fandeha* « manière de marcher, démarche ».

4.2 Graphes de grammaires locales

Dans l'état actuel de notre recherche, 67 graphes de grammaires locales ont été créés. Ils correspondent aux codes de classes affixales (3.2). Ces graphes permettent l'analyse morphologique des noms issus des verbes. Ci-contre, le graphe de grammaire locale v2B pour les radicaux verbaux qui n'ont pas de noms d'agent ni de profession mais ont toutes les autres formes nominales (noms de manière ou d'état et noms d'action) est fourni.

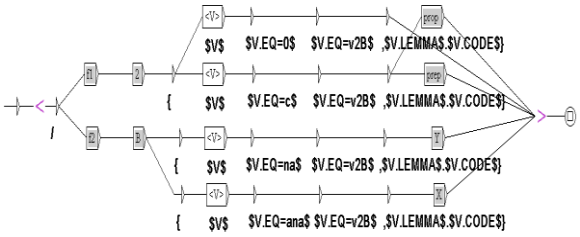

Figure C : Graphe de grammaire locale v2B

Comme exemple se rattachant à ce graphe, nous avons *zozozòzo* « bourdonner, bruire, siffler », d'où *fizozozozo* « manière de bourdonner, de bruire, de siffler » (Nman), *fizozozozoana* « bourdonnement, bruissement, sifflement » (Nact). Ce type de graphe peut aussi être utilisé par des programmes de génération de formes non plus pour découper les formes reconnues mais pour construire, indépendamment d'un corpus donné, des listes de formes nominales. Rakotoalimanana (2000, p. 378) expose un exemple de ce programme de génération de formes avec les formes verbales. Il y présente un prototype d'Analyseur – Générateur des Termes prédicatifs Malgaches (AGTM) implémenté en langage Prolog.

5 Les dictionnaires de noms

Les codes de classes affixales et radicales sont insérés dans le dictionnaire de noms et opèrent directement sur le dictionnaire à l'aide des transducteurs de flexion et des graphes de grammaire locale. Dans cette section, le dictionnaire électronique des noms issus de verbes (DEMA-NVS) est présenté en premier lieu ; ensuite, le dictionnaire des variantes morphologiques des radicaux (DEMA-NVSflx) est abordé en second lieu.

5.1 DE des noms issus des verbes (DEMA-NVS)

Les entrées du DEMA-NVS sont les radicaux verbaux. Dans l'état actuel de notre recherche, elles sont au nombre de 1500 ; toutes les entrées commençant par A – J, M, N, Z ont été codées. Ci-après un extrait de ce dictionnaire.

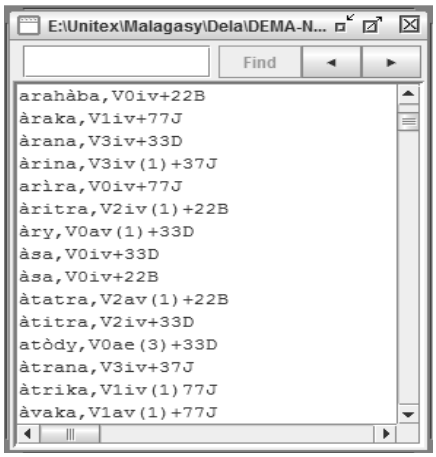

Figure D : DEMA-NVS

[9] L'accent graphique note une information fournie par le dictionnaire sur l'accent tonique, mais il n'est généralement pas précisé dans les textes écrits.

Avec les conventions d'Unitex, les articles du dictionnaire sont séparés des entrées par une virgule et les codes après la virgule sont les articles du dictionnaire. Les avantages d'un dictionnaire construit par les linguistes sont qu'il est précis et facile à mettre à jour. Il fournit des informations jugées pertinentes soit pour les futurs programmes d'aide à la construction de dictionnaires usuels, soit pour les logiciels et applications destinées aux utilisateurs finaux.

5.2 DE des variantes morphologiques des noms (DEMA-NVSflx)

Les variantes morphologiques des radicaux verbaux formant des noms sont rangées dans un autre dictionnaire appelé DEMA-NVSflx. À proprement parler, le dictionnaire n'est pas un dictionnaire de formes fléchies de noms, il est un dictionnaire morphologique contenant les variantes morphologiques du radical, et indiquant par des codes les affixes se combinant avec ces variantes. Ci-après une image du DEMA-NVSflx.

Figure E : DEMA-NVSflx

Pour fournir un véritable dictionnaire de formes fléchies, un autre programme de génération automatique de termes est requis. Il servira plus tard à de nombreuses applications pratiques.

6 Test du dictionnaire

Des procédures d'évaluation du dictionnaire ont été mises au point sur un extrait du corpus journalistique du malgache contemporain (cjmc) de Diwersy (2009) qui n'a pas été utilisé pour construire le dictionnaire. Le dictionnaire a été testé sur les 50 premières phrases du cjmc1[10] qui comportent 35 noms différents. Parmi les 35 noms différents :

- 6 sont reconnus et découpés par Unitex en utilisant les ressources codées
- 29 ne sont pas reconnus car ils sont absents du dictionnaire. Parmi ces derniers :
    - Toutes les classes radicales existent déjà dans les ressources
    - Pour les classes affixales :
        - 24 noms non-reconnus correspondent en réalité à 6 classes affixales existantes dans le fichier des grammaires locales
        - pour les 5 autres noms non-reconnus, les classes affixales sont à insérer dans les ressources

En termes de classes radicales, le texte est à 100% couvert tandis qu'en termes de classes affixales, il est à 86% couvert. D'une manière générale, la plupart des classes radicales et affixales des radicaux ont déjà été construites dans Unitex au cours du travail. Il s'agit ensuite d'enrichir le dictionnaire de radicaux verbaux et le dictionnaire peut couvrir le lexique des noms issus de verbes.

7 Couverture lexicale

Une fois que le dictionnaire est enrichi des radicaux verbaux formant des noms, Unitex est capable de faire les analyses morphologiques des noms d'agent, de profession, de manière, d'instrument, d'état et d'action dérivés de ces radicaux. Il peut reconnaître également d'une part ces noms couplés avec des pronoms personnels du type *fijeriko* « mon regard », *filalaoko* « ma manière de jouer », *fisaorako* « mon remerciement » ou avec des prépositions comme *mpamilin'* « le chauffeur de » et d'autre part les variantes morphologiques de ces noms au début des radicaux au contact d'un trait d'union comme *pifamoivoizana* (de *fifamoivoizana* « action de circuler, circulation ») dans *lozam-pifamoivoizana* « accident de la circulation » dans les mots composés. Les transducteurs de flexion et les graphes de grammaires locales construits fonctionnent correctement et le codage des entrées pour constituer un DE complet des noms issus des verbes est en cours. Si dans l'état actuel de notre recherche, nous sommes à 1500 entrées de ce dictionnaire, il reste 53% des entrées qui ont besoin d'être insérées dans le dictionnaire. Une fois l'enrichissement du dictionnaire complet, un dictionnaire DEMA-NVS des noms issus de verbes du malgache sera disponible, ce qui augmentera d'une manière assez considérable la couverture lexicale du malgache.

8 Conclusion

La construction de dictionnaire électronique des noms issus de verbes est en phase de constitution au Centre Interdisciplinaire de Recherche Appliquée au Malgache. S'il reste des entrées manquantes qui doi-

[10] Cjmc 1 est une partie du corpus journalistique du malgache contemporain de Diwersy (2009) dont nous avons divisé en quatre parties (voir Ranaivoarison, 2016, p. 260). Cjmc1 comporte 180 000 mots et 12 700 phrases.

vent être insérées dans le dictionnaire pour constituer un dictionnaire complet, ce dictionnaire est déjà utilisable pour certaines applications. Une fois que la construction de ce dictionnaire sera terminée, la construction des dictionnaires de noms issus d'adjectifs et de noms simples constituerait les prochaines priorités pour former un dictionnaire de noms simples qui tend à l'exhaustivité du vocabulaire.

L'extension de ce dictionnaire aux autres catégories grammaticales (adjectifs, adverbes, et les autres catégories à faible variation de formes telles que les conjonctions, les prépositions, etc.) permettra d'avoir un dictionnaire morphologique électronique complet du malgache qui servira d'accès aux dictionnaires de mots composés et d'un lexique-grammaire représentant systématiquement les propriétés syntaxiques des mots de la langue. Ces informations seront ensuite utilisées dans d'autres programmes informatiques qui ont pour finalité la génération de formes, la normalisation, la correction orthographique et/ou grammaticale. En d'autres termes, elles serviront à la construction d'outils de TAL performants et accessibles aux grands publics.

Références

Berment, V. (2004). *Méthodes pour informatiser des langues et des groupes de langues « peu dotées »*. Thèse de doctorat. Université Jean Fourier, Grenoble 1.

Dalrymple, M., Liakata, M., Mackie, L. (2006). Tokenization and morphological analysis for Malagasy. In: *Computational Linguistics and Chinese Language Processing* 11 (4), pp. 315-332. Taipei: Institute of Linguistics, Academia Sinica.

Diwersy, S. (2009). *Corpus journalistique du malgache contemporain.* Romance Philology Department University of Cologne.

Gross, M. (1989). La construction de dictionnaires électroniques. In : *Annales des télécommunication, tome 44* N°1, 2. Issy-les-Moulineaux/lannion : CNET.

Koskenniemi, K. (1983). *Two-Level Morphology: A general Computational Model for Word-Form Recognition and Production.* Department of General Linguistics, University of Helsinki.

Koskenniemi, K. and Church, K.W. (1988). Complexity, two-level morphology and Finnish. In: *COLLING'88.*

Nam, J. S. (1994). Construction d'un lexique électronique des noms simples en coréen. In : *Lexiques-grammaires comparés et traitements automatiques*. Université du Québec à Montréal : Jacques Labelle, pp. 219-245.

Nam, J. S., Paumier, S. (2014). Un système de dictionnaire de mots simples du coréen. Fryni Kakoyianni-Doa. Penser le Lexique-Grammaire. Perspectives actuelles, Honoré Champion, pp.481-490, 2014, Collection Colloques, congrès et conférences. Sciences du Langage, histoire de la langue et des dictionnaires. 30th International Conference on Lexis and Grammar (Nicosia, Cyprus, 2011), 978-2-7453-2512-9.

Oflazer, K. (1993). Two-level Description of Turkish Morphology. In: *EACL'06.* Netherlands, Utrecht.

Paumier, S. (2016). *Unitex 3.1. Manuel d'utilisation.* Université Paris-Est Marne-la-Vallée. Version française.

Rajaona, S. R. (1972). *Structure du malgache.* Antananarivo : Ambozontany.

Rakotoalimanana, H. D. (2000). *Structure morpho-syntaxique et modélisation informatique.* Thèse de doctorat. Université Nancy 2.

Ranaivoarison, J., Laporte, É., Ralalaoherivony, B. S. (2013). Formalisation of Malagasy conjugation. In: *Language and Technology Conference*. Poznan, Poland. pp.457-462.

Ranaivoarison, J. (2015a). *Description du dictionnaire électronique des verbes simples du malgache*. Session Poster. Colloques Jeunes Chercheurs. Montpellier.

Ranaivoarison, J. (2016). *Construction de dictionnaire électronique des verbes du malgache*. Deutschland : Editions Universitaires Européennes.