Université de Montréal

Élaboration d'un vocabulaire de base du kirundi écrit
Problématique, description et applications

par

Pascal Ntirampeba

Département de linguistique et de traduction
Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de
Philosophiæ doctor (Ph.D.)
en linguistique

Août 1999

© Pascal Ntirampeba, 1999

P
25
U54
2000
v.006

Université de Montréal
Faculté des études supérieures

Cette thèse intitulée :

Élaboration d'un vocabulaire de base du kirundi écrit.
Problématique, description et applications

présentée par

Pascal Ntirampeba

a été évaluée par un jury composé des personnes suivantes :

Gilles Bélanger	Président-rapporteur
Louise Dagenais	Directrice de recherche
Nathan Ménard	Codirecteur de recherche
Nazam Halaoui	Membre du jury
Pierre Lafon	Examinateur externe

Thèse acceptée le 3 décembre 1999

SOMMAIRE

La présente recherche porte sur l'élaboration d'un vocabulaire de base du kirundi écrit. Le kirundi étant une langue agglutinante, l'étude table sur la complémentarité entre deux niveaux d'analyse : le niveau lexical et le niveau morphologique.

Après avoir situé notre recherche en statistique lexicale et en morphologie computationnelle, nous consacrons le chapitre 1 à la définition du vocabulaire de base, aux critères mis en avant pour le sélectionner et aux matériaux à partir desquels il est élaboré.

Dans le chapitre 2, nous présentons notre cheminement méthodologique depuis l'échantillonnage jusqu'à l'analyse des résultats. Nous indiquons les problèmes rencontrés au moment de la saisie des données, de l'indexation et de la reconnaissance morphologique automatique. Sont également abordés les problèmes liés à la norme lexicologique. Nous présentons nos décisions en rapport avec la détermination du nombre de mots-formes (N) et de celui des vocables (V). Nous décrivons ensuite notre dictionnaire de référence (Rodegem 1970) et expliquons enfin le calcul des indices statistiques que nous retenons à savoir les indices de fréquence, de dispersion et d'usage.

Nous consacrons le chapitre 3 à une présentation générale de la morphologie du kirundi dans une perspective lexicométrique et à des fins de reconnaissance automatique des morphèmes. Nous présentons les règles morphophonologiques à partir desquelles fonctionne l'analyseur morphologique et fournissons une description des mots-formes et des morphèmes de la langue.

Les résultats de notre recherche sont présentés au chapitre 4. Ils sont abordés selon les deux niveaux d'analyse considérés.

Au niveau lexical, une attention particulière est réservée aux vocables dont l'indice d'usage est supérieur ou égal à 3, vocables qui constituent le noyau du vocabulaire de base du kirundi écrit. Nous analysons par la suite la fréquence et la stabilité des catégories grammaticales et comparons les 50 vocables les plus fréquents en français, en anglais et en kirundi.

Au niveau morphologique, cinq types de morphèmes font l'objet de comptages : les morphèmes aspectuels du verbe, les préfixes de classes du substantif, les suffixes de dérivation des verbes et des substantifs ainsi que les dérivatifs thématiques nominaux.

Il restait à envisager l'utilisation des vocables et des morphèmes sélectionnés en didactique du kirundi. C'est l'objet du chapitre 5.

En didactique du kirundi langue maternelle, le vocabulaire de base peut servir notamment dans l'élaboration de phrases-types qui constituent la base de l'apprentissage de la lecture et de l'écriture, dans la mise au point de textes simplifiés pour la lecture et dans la vulgarisation scientifique.

En didactique du kirundi L2, nous proposons des activités qui visent à développer les habiletés à l'expresion orale, à la lecture et à l'écriture; ces activités doivent prendre en compte la tonalité et la quantité vocalique.

Mots-clés : lexicométrie, morphologie computationnelle, vocabulaire de base, langues africaines, kirundi.

TABLE DES MATIÈRES

Identification du jury..... .i
Sommaire iii
Table des matières..... .v
Liste des tableaux, des graphiques et des schémas..... .xviii
Liste des symboles et abréviations..... .xx
Remerciements xxiii
Dédicace..... .xxiv
Correspondances grapho-phonétiques..... .xxvi

INTRODUCTION..... .1

1. INTÉRÊT ET MOTIVATION DE L'ÉTUDE..... .1

2. CHAMP D'ÉTUDE ET APPLICATIONS..... .2
2.1. La statistique lexicale.....4
2.1.1. Les études de statistique lexicale descriptives.....4
2.1.2. Les études de statistique lexicale à visées pédagogiques.....6
2.1.2.1. Les recherches psycholinguistiques.....7
2.1.2.2. Les vocabulaires de base.....8
2.1.3. L'évolution des technologies en statistique lexicale.....9
2.2. La statistique morphologique.....11
2.3. La morphologie computationnelle..... ..12
2.3.1. Datation du syntagme « x computationnel ».....12
2.3.2. La morphologie computationnelle en traitement automatique des langues naturelles.....12
2.3.2.1. Les principales difficultés 14

2.3.2.2. Les applications..... 16
2.3.2.2.1. Les recherches sur la parole..... 16
2.3.2.2.2. Les recherches sur la langue écrite 16
2.3.2.2.3. La recherche documentaire..... 17
2.3.2.2.4. Le traitement de textes 17
2.3.2.3. Les analyseurs morphologiques..... 18

3. HYPOTHÈSES DE RECHERCHE..... 20
3.1. Hypothèse générale..... 20
3.2. Hypothèses spécifiques..... 20
3.2.1. Hypothèse sur la stabilité des fréquences des catégories grammaticales..... 21
3.2.2. Hypothèse sur la hiérarchie des fréquences des catégories grammaticales..... 21
3.2.3. Hypothèse méthodologique..... 21

4. DIFFICULTÉS ET LIMITES DE L'ÉTUDE.....22

5. PLAN.....23

CHAPITRE 1 : ÉLABORATION DES VOCABULAIRES DE BASE : ÉTAT DE LA QUESTION.....24

1. QU'EST-CE QU'UN VOCABULAIRE DE BASE ?..... 24

2. UN VOCABULAIRE DE BASE FONDÉ SUR L'ÉCRIT : JUSTIFICATION 25

3. CRITÈRES DE SÉLECTION DES VOCABULAIRES DE BASE..... 29
3.1. La fréquence 29
3.2. La répartition..... 29
3.3. La dispersion..... 30
3.4. La disponibilité 30
3.5. L'usage..... 31

4. LES SOURCES DES VOCABULAIRES FONDAMENTAUX..... 32
4.1. Les corpus produits par des répondants au cours d'enquêtes sur la ..... disponibilité..... 32
4.2. Les compilations de listes de mots ou de recueils 32
4.3. Les textes..... 32
4.4. Les listes élaborées à partir de langues différentes 35
4.5. La liste du Basic..... 35

5. TERMINOLOGIE UTILISÉE 37
5.1. Le mot, le lexème et le vocable 37
5.1.1. Au niveau du discours : mot, forme, occurrence et mot-forme..... 39
5.1.2. Au niveau de la langue : lexème et lexie..... 40
5.1.3. Le vocable et le lemme 42
5.1.4. La lemmatisation 43
5.2. Le morphe et le morphème.....45

CHAPITRE 2 : MÉTHODOLOGIE DE L'ÉTUDE..... 47

1. LE CORPUS ET L'ÉCHANTILLONNAGE 47
1.1. L'échantillonnage aléatoire simple 49
1.2. L'échantillonnage aléatoire stratifié proportionnel 50

2. LE TRAITEMENT DES DONNÉES..... 55
2.1. La saisie des données avec *WordPerfect* 5.1 57
2.1.1. La désambiguïsation des mots-formes homographes 57
2.1.1.1. L'homographie lexicale..... 58
2.1.1.1.1. L'homographie verbe / verbe..... 58
2.1.1.1.2. L'homographie substantif à base verbo-nominale / substantif à base verbo-nominale 58

2.1.1.1.3. L'homographie substantif à base verbo-nominale / substantif à base nominale 58
2.1.1.1.4. L'homographie substantif à base nominale / substantif à base nominale 59
2.1.1.1.5. L'homographie nom propre / substantif à base verbo-nominale 59
2.1.1.1.6. L'homographie mot grammatical / mot grammatical 59
2.1.1.2. L'homographie syntaxique 59
2.1.1.2.1. L'homographie mot grammatical / verbe 60
2.1.1.2.2. L'homographie substantif à base verbo-nominale / verbe 60
2.1.1.2.3. L'homographie substantif à base nominale / verbe 60
2.1.1.2.4. L'homographie nom propre / verbe 60
2.1.1.2.5. L'homographie substantif à base nominale / mot grammatical 61
2.1.1.3. L'homographie morphologique 61
2.1.1.3.1. L'homographie liée au préfixe personnel 62
2.1.1.3.2. L'homographie liée aux suffixes de dérivation 62
2.1.1.3.2.1. L'applicatif [-ir-] 62
2.1.1.3.2.2. Le causatif indirect [-ish-] 63
2.1.1.3.2.3. L'associatif [-an-] 63
2.1.1.3.2.4. L'opposif [-ur-] 63
2.1.1.3.3. L'homographie liée aux morphèmes compléments 63
2.1.1.3.4. L'homographie liée aux particules adverbales 64
2.1.1.3.4.1. Le morphème actualisateur [-ra-] 64
2.1.1.3.4.2. Le morphème négateur [-tá-] 64
2.1.1.3.4.3. Le morphème de syndèse [-ka-] 64
2.1.1.3.4.4. Le morphème de syndèse [-ki-] 65
2.1.1.3.5. L'homographie liée au prédicatif verbal [-oo-] 65
2.1.1.3.6. L'homographie liée au morphème aspectuel accompli [-ye] 66
2.1.1.3.7. L'homographie liée au morphème réfléchi 67
2.1.2. Les autres modifications apportées au corpus 68
2.1.2.1. La séparation de la particule dicto-modale *nti* 'ne...pas' et du verbe 68

2.1.2.2. La séparation des morphèmes locatifs et du verbe 69
2.1.2.3. La soudure d'une variante lexicale bi-morphématique 70
2.1.2.4. La restitution des voyelles élidées aux mots-formes 70
2.1.2.5. Les variantes orthographiques 70
2.1.2.6. Le type *mweénewáanyu* 'ton frère' / 'ta soeur' 71
2.1.3. La délimitation des mots-formes 72
2.1.3.1 Les mots composés 72
2.1.3.2. Les locutions 73
2.1.3.3. Les sigles 74
2.1.3.4. Les titres 75
2.1.3.5. Les onomatopées 75
2.1.4. La détermination du nombre de vocables 76
2.1.5. Le dictionnaire de référence : Rodegem (1970) 77
2.1.5.1. Les vocables recensés et exclus 77
2.1.5.2. Le traitement de la polysémie / homonymie 77
2.1.5.3. Le traitement des dérivés 80
2.1.5.3.1. Les dérivés verbaux 80
2.1.5.3.2. Les substantifs prévisibles 81
2.1.5.3.3. Le traitement des adjectifs 83
2.1.5.3.4. Le traitement des mots grammaticaux 84
2.1.5.3.4.1. Les mots-formes grammaticaux courts *versus* longs 85
2.1.5.3.4.2. Les mots-formes grammaticaux homographes 85
2.1.5.3.5. Les vocables pour les interjections, les conjonctions et adverbes 88
2.2. L'indexation avec *WordCruncher* 4.23 88
2.3. L'analyseur morphologique 90
2.3.1. L'interface de l'analyseur morphologique 91
2.4. Le logiciel statistique SPSS 7.5 93
2.5. Les indices calculés avec SPSS 7.5 94
2.5.1. L'indice de fréquence (*Fo*) 94
2.5.2. L'indice de dispersion (*D*) 96

2.5.3. L'indice d'usage (*U*)..... 97
2.6. Les résultats attendus 98

3. L'ANALYSE DE LA LISTE DES VOCABLES..... 99
3.1. Le volet quantitatif..... 99
3.2. Le volet qualitatif 101

4. CONCLUSION..... 101

CHAPITRE 3 : LA MORPHOLOGIE DU KIRUNDI..... 103

1. INTRODUCTION..... 103

2. LES MODÈLES MORPHOLOGIQUES APPLIQUÉS AU KIRUNDI 104

3. LA STRUCTURE MORPHOLOGIQUE DES MOTS-FORMES 106
3.1. Aspects généraux 106
3.1.1. Aspects tonals..... 106
3.1.2. Durée vocalique..... 108
3.1.3. Classes d'accord..... 109
3.1.3.1. Homographie à l'intérieur d'une classe d'accord 111
3.1.3.1.1. Homographie totale..... 111
3.1.3.1.2. Homographie partielle 111
3.1.4. Aspects morphophonologiques 114
3.1.4.1. Règles phonologiques..... 115
3.1.4.1.1. Assimilation régressive de la consonne nasale /n/ ..... devant une consonne labiale..... 115
3.1.4.1.2. Consonantisation des voyelles..... 116
3.1.4.1.3. Assimilation progressive de /r/ après la nasale /n/..... 116
3.1.4.1.4. L'assimilation progressive de /h/ devant la nasale /n/..... 116

3.1.4.1.5. Troncation de /a, i/ après /C/ devant une voyelle 117
3.1.4.2. Règles morphophonologiques..... 117
3.1.4.2.1. Chute de consonne 117
3.1.4.2.2. Spirantisation 118
3.1.4.2.3. Contraction des voyelles..... 118
3.1.4.2.4. Loi de Dhal..... 119
3.1.4.2.5. Palatalisation 119
3.1.4.2.6. Épenthèse..... 119
3.1.3.2.7. Harmonie vocalique..... 120
3.2. Mots-formes lexicaux 121
3.2.1. Mot-forme verbal 121
3.2.1.1. Radicaux et morphèmes aspectuels 122
3.2.1.2. Autres morphèmes du mot-forme verbal..... 124
3.2.1.3. Suffixes de dérivation 128
3.2.1.3.1. Dérivatèmes de contact..... 128
3.2.1.3.2. Dérivatèmes de manière..... 129
3.2.1.3.3. Dérivatèmes divers 131
3.2.2. Mot-forme substantif 133
3.2.2.1. Substantif à base verbo-nominale 133
3.2.2.2. Substantif à base nominale 136
3.2.3. Mot-forme adjectif 136
3.3. Mots-formes grammaticaux..... 138
3.3.1. Mots-formes grammaticaux fléchis 138
3.3.1.1. Mots-formes grammaticaux fléchis homographes..... 138
3.3.1.1.1. Démonstratifs 138
3.3.1.1.2. Indéfinis 140
3.3.1.1.3. Interrogatifs .....
3.3.1.1.4. Numéraux cardinaux..... 142
3.3.1.1.5. Locatifs 143
3.3.1.1.6. Interpellatifs..... 144
3.3.1.1.7. Adverbes..... 144
3.3.1.2. Mots-formes grammaticaux fléchis hétérographes 144

3.3.1.2.1. Possessifs 144
3.3.1.2.2. Connectifs 145
3.3.1.3. Les mots-formes courts *versus* les mots-formes longs 147
3.3.1.3.1. Pronoms allocutifs 147
3.3.1.3.2. Pronoms substitutifs 148
3.3.1.3.4. Numéraux ordinaux 149
3.3.1.3.5. Adverbes 150
3.3.2. Mots-formes grammaticaux non fléchis 150
3.3.2.1. Prédicatifs nominaux 150
3.3.2.2. Prépositions 151
3.3.2.3. Conjonctions 151
3.3.2.3.1. Conjonctions de coordination 152
3.3.2.3.2. Conjonctions de subordination 152
3.3.2.4. Adverbes 153
3.3.2.5. Interjections 154
3.3.2.6. Onomatopées et idéophones 154

CHAPITRE 4 : LE VOCABULAIRE DE BASE DU KIRUNDI ÉCRIT 157

1. RÉSULTATS QUANTITATIFS GÉNÉRAUX 158

2. ANALYSE DU VOCABULAIRE DE BASE DU KIRUNDI ÉCRIT 161
2.1. Les vocables dont U ≥ 3 161
2.1.1. Les vocables verbaux dont U ≥ 3 163
2.1.1.1. Aspects morphologiques 163
2.1.1.1.1. Les vocables verbaux morphologiquement défectifs 164
2.1.1.1.2. Les vocables verbaux de structure CV 164
2.1.1.2. Aspects syntaxiques 165
2.1.1.2.1. Les auxiliaires virtualisants 166
2.1.1.2.2. Les auxiliaires actualisants 166
2.1.1.2.3. Les auxiliaires réalisants 167

2.1.1.3. Aspects sémantiques 168
2.1.1.3.1. Les vocables verbaux lexicaux thématiques 169
2.1.1.3.2. Les vocables verbaux lexicaux athématiques 170
2.1.1.3.3. Les relations lexico-sémantiques 172
2.1.1.3.3.1. La synonymie 172
2.1.1.3.3.2. L'antonymie 176
2.1.2. Les vocables nominaux à base nominale 177
2.1.2.1. Les vocables métadiscursifs 178
2.1.2.2. Les vocables thématiques 179
2.1.2.3. Les vocables athématiques 180
2.1.2.4. Les relations lexico-sémantiques 182
2.1.2.4.1. La synonymie 182
2.1.2.4.2. Les relations d'hyperonymie / hyponymie 185
2.1.3. Les vocables nominaux à base verbo-nominale dont U ≥ 3 186
2.1.3.1. Les vocables thématiques 186
2.1.3.2. Les vocables athématiques 187
2.1.3.2.1. Les vocables généraux 187
2.1.3.2.2. Les vocables métadiscursifs 189
2.1.2.3. Vocables nominaux à base verbo-nominale et radicaux verbaux 190
2.1.4. Les vocables adjectivaux dont U ≥ 3 192
2.1.5. Les vocables grammaticaux dont U ≥ 3 193
2.1.5.1. Les vocables grammaticaux flexionnels 194
2.1.5.1.1. Connectifs 195
2.1.5.1.2. Démonstratifs 198
2.1.5.1.3. Allocutifs 200
2.1.5.1.4. Possessifs 201
2.1.5.1.5. Locatifs 203
2.1.5.1.6. Indéfinis 203
2.1.5.1.7. Numéraux cardinaux 205
2.1.5.1.8. Substitutifs 206
2.1.5.1.9. Adverbes 207

2.1.5.2. Les vocables grammaticaux non flexionnels 208
2.1.5.2.1. Prépositions 209
2.1.5.2.2. Prédicatifs nominaux 210
2.1.5.2.3. Conjonctions 210
2.1.5.2.3.1. Conjonctions de coordination 210
2.1.5.2.3.2. Conjonctions de subordination 211
2.1.5.2.4. Adverbes 212
2.1.5.2.4.1. Adverbes de manière 212
2.1.5.2.4.2. Adverbes de comparaison 212
2.1.5.2.4.3. Adverbes de lieu 213
2.1.5.2.4.4. Adverbes de temps 213
2.1.5.2.4.5. Adverbes d'intensité 214
2.1.5.5.4.6. Adverbes d'affirmation et de négation 214
2.1.5.2.4.7. Adverbes d'interrogation 215
2.2. Les vocables dont U est compris entre 3 et 0 216
2.2.1. Verbes 217
2.2.2. Noms à base nominale 218
2.2.2.1. Noms composés 218
2.2.2.2. Emprunts 219
2.2.2.2.1. Emprunts au français 219
2.2.2.2.2. Emprunts au swahili 220
2.2.2.2.3. Emprunts à diverses langues 221
2.2.3. Vocables adjectivaux 221
2.2.4. Vocables grammaticaux 221
2.3. Les vocables dont U < 0 223
2.3.1. Vocables verbaux 223
2.3.2. Vocables nominaux à base verbo-nominale 225
2.3.3. Vocables nominaux à base nominale 226
2.3.3.1. Noms composés 226
2.3.3.2. Emprunts 227
2.3.4. Vocables adjectivaux 228
2.3.5. Vocables grammaticaux 229

2.4. Les « autres » mots-formes du corpus 230
2.4.1. Noms propres..... 230
2.4.1.1. Anthroponymes 230
2.4.1.2. Toponymes 230
2.4.1.3. Mois 231
2.4.2. Abréviations..... 232
2.4.3. Mots-formes appartenant à d'autres langues 233
2.5. Fréquence des catégories grammaticales..... 234
2.6. Stabilité des fréquences des catégories grammaticales 239
2.7. Les 50 vocables les plus fréquents 242

3. STATISTIQUES MORPHOLOGIQUES..... 245
3.1. Les morphèmes objets de comptages 245
3.2. Les limites de ce volet de recherche 247
3.3. Les problèmes en morphologie computationnelle du kirundi..... 247
3.3.1. Le choix du lemme..... 248
3.3.1.1. La réduplication du radical..... 248
3.3.1.2. Les variantes lexicales 249
3.3.2. La complexité morphophonologique..... 249
3.3.3. Paires de radicaux dont l'un ressemble formellement à un ..... dérivé de l'autre 250
3.4. Fréquence des marqueurs aspectuels du verbe..... 252
3.5. Les morphèmes propres au substantif..... 254
3.5.1. Les préfixes de classe..... 254
3.5.1.1. Le préfixe de classe [-ri-] 258
3.5.1.2. Le préfixe de classe [-mu-] 259
3.5.1.3. Le préfixe de classe [-ba-]..... 261
3.5.1.4. Le préfixe de classe [-n-]..... 261
3.5.1.5. Le préfixe de classe [-ki-]..... 262
3.5.1.6. Le préfixe de classe [-ma-]..... 262
3.5.1.7. Le préfixe de classe [-bu-] 264

3.5.1.8. Le préfice de classe [-mi-] 264
3.5.1.9. Le préfice de classe [-bi-] 265
3.5.2. Les dérivatifs thématiques nominaux 266

3.6. Les suffixes de dérivation 268
3.6.1. Les suffixes de dérivation dans les mots-formes verbaux 269
3.6.2. Les suffixes de dérivation dans les mots-formes substantifs à base verbo-nominale 272

CHAPITRE 5 : QUELQUES APPLICATIONS PÉDAGOGIQUES DU VOCABULAIRE DE BASE DU KIRUNDI ÉCRIT 278

1. INTRODUCTION 278

2. ASPECTS GÉNÉRAUX DE L'ENSEIGNEMENT DU VOCABULAIRE 278
2.1. Pour et contre l'utilisation des listes de fréquence en didactique des langues 278
2.1.1. Les vocables fréquents sont pauvres en information 279
2.1.2. Les vocabulaires de base constituent des listes de formes et pas de sens 279
2.1.3. Les listes diffèrent pour une même langue 280
2.1.4. La fréquence est un critère insuffisant 281
2.2. Qu'est-ce que connaître un vocable ? 282
2.3. L'enseignement systématique du vocabulaire 283
2.3.1. Le contexte linguistique authentique 283
2.3.2. Les thèmes 284
2.3.3. Les approches méthodologiques 286
2.3.3.1. Le champ sémantique 286
2.3.3.3. Les relations lexico-sémantiques paradigmatiques 291
2.3.3.3.1. La synonymie et l'antonymie 291
2.3.3.3.2. L'hyponymie et l'hypéronymie 291

2.3.3.4. Les collocations 292

3. VOCABULAIRE DE BASE ET DIDACTIQUE DU KIRUNDI L1 294
3.1. À l'école primaire..... 294
3.1.1. Utilisation du vocabulaire de base dans les phrases-types 294
3.1.2. Vocabulaire de base et textes simplifiés 296
3.2. Au secondaire..... 298
3.2.1. L'utilisation des vocables à faible indice d'usage 298
3.2.2. Les vocables métadiscursifs 299
3.2.3. Les vocables reliés à un domaine d'activité..... 299
3.2.4. La structure morpho-sémantique des vocables 300
3.3. En alphabétisation des adultes 301
3.3.1. Quoi lire?..... 302
3.3.2. Quoi écrire ?..... 303

4. VOCABULAIRE DE BASE ET DIDACTIQUE DU KIRUNDI L2 303
4.1. Écouter et parler 305
4.2. Lire et écrire..... 306

CONCLUSION GÉNÉRALE..... 310

RÉFÉRENCES..... 315

ANNEXES..... 330
ANNEXE 1 : Liste des numéros, pages et colonnes saisis par sous-corpus..... i
ANNEXE 2 : Exemple de résultats fournis par l'analyseur morphologique xv
ANNEXE 3 : Occurrences de quelques radicaux adjectivaux dans les 16 sous-corpus..... xvii
ANNEXE 4 : Typologie des thèmes..... xviii
ANNEXE 5 : Écart type, coefficient de variation et fréquences théoriques de quelques vocables à indice de dispersion négatif..... xix
DISQUETTE INFORMATIQUE xx

LISTE DES TABLEAUX, DES GRAPHIQUES ET DES SCHÉMAS

1. TABLEAUX

Tableau 1 - Taille des corpus et des vocabulaires de base 33
Tableau 2 - Terminologie de Muller (1977 : 6-7), Mel'cuk (1993), Mel'cuk *et al.* (1995) et Ntirampeba (1999) 38
Tableau 3 - Nombre de numéros utilisés par année 48
Tableau 4 - Échantillonnage 48
Tableau 5 - Nombre de pages et de colonnes des journaux dépouillés 51
Tableau 6 - Structure thématique de l'échantillon 52
Tableau 7 - Structure de l'échantillon 53
Tableau 8 - Liste des sous-corpus 54
Tableau 9 - Typologie de l'homographie en kirundi 57
Tableau 10 - Interface de l'analyseur morphologique 91
Tableau 11 - Représentation hiérarchique d'un verbe selon Goldsmith et Sabimana (1989) 105
Tableau 12 - Les préfixes d'accord du kirundi 110
Tableau 13 - Homographie partielle des préfixes d'accord 111
Tableau 14 - Incompatibilité des morphèmes dans le mot-forme verbal 126
Tableau 15 - Déterminants allocutifs du kirundi 147
Tableau 16 - Les classes de V selon U et fréquence des mots-formes 159
Tableau 17 - Les vocables dont $U \geq 3$ 162
Tableau 18 - Les vocables grammaticaux flexionnels selon U et N 194
Tableau 19 - Les classes des vocables grammaticaux flexionnels dont $U \geq 3$ 195
Tableau 20 - Les vocables grammaticaux non flexionnels selon U et N 208
Tableau 21 - Les classes des vocables grammaticaux non flexionnels dont $U \geq 3$ 209
Tableau 22 - Les classes de vocables dont $U < 3 \leq 0$ 216
Tableau 23 - Les classes de vocables dont $U < 0$ 223

Tableau 24 - Fréquence des catégories grammaticales..... 234
Tableau 25 - Fréquence des catégories grammaticales en français, en anglais et en kirundi..... 235
Tableau 26 - Fréquences absolues des catégories grammaticales..... 240
Tableau 27 - Stabilité des fréquences des catégories grammaticales 241
Tableau 28 - Les morphèmes objets de comptages..... 246
Tableau 29 - Fréquences absolues des morphèmes aspectuels 252
Tableau 30 - Les marqueurs aspectuels selon v, D et U..... 253
Tableau 31 - Fréquences absolues des préfixes de classes..... 255
Tableau 32 - Les préfixes de classes selon v, D et U..... 256
Tableau 33 - Fréquences absolues des dérivatifs thématiques nominaux..... 266
Tableau 34 - Les dérivatifs thématiques nominaux selon v, D et U..... 267
Tableau 35 - Fréquences absolues des suffixes de dérivation dans les mots-formes verbaux 270
Tableau 36 - Les suffixes de dérivation dans les mots-formes verbaux selon v, D et U..... 271
Tableau 37 - Fréquences absolues des suffixes de dérivation dans les mots-formes substantifs à base verbo-nominale..... 273
Tableau 38 - Les suffixes de dérivation dans les mots-formes substantifs à base verbo-nominale selon v, D et U..... 274
Tableau 39 - Répartition des vocables selon le niveau des apprenants..... 304

2. GRAPHIQUES

Graphique 1 - Rapports entre les catégories de vocables dont U ≥ 3 162
Graphique 2 - Fréquence des catégories grammaticales en français, en anglais et en kirundi..... 236

3. SCHÉMAS

Schéma 1 - Le traitement des données..... 56
Schéma 2 - Relations entre vocables d'un paragraphe..... 288
Schéma 3 - Relations entre vocables tirés d'un paragraphe et de la liste du vocabulaire de base..... 290

LISTE DES SYMBOLES ET ABRÉVIATIONS

1. SYMBOLES STATISTIQUES

D	dispersion
F_O	fréquence observée
N	nombre de mots-formes
σ	écart type
R	répartition
U	usage
V	nombre de vocables
v	coefficient de variation
$\overline{X}$	moyenne

2. AUTRES SYMBOLES

C	consonne
MF	mot-forme
MG	mots grammaticaux
R	radical
SBN	substantif à base nominale
SBVN	substantif à base verbo-nominale

3. ABRÉVIATIONS

act.	actualisateur
adj.	adjectif
adv.	adverbe
appl.	applicatif
art.	article
asp.	aspect

ass.	associatif
augm.	augment
caus.	causatif
cf.	confer
chap.	chapitre
cl.	classe
compl.	complément
conj.	conjonction
dém.	démonstratif
dér.th.n.	dérivatif thématique nominal
dét.	déterminant
dmod.	dicto-modal
ex.	exemple
fact.	factitif
fut.	futur
ind.	indirect
interrog.	interrogatif
loc.	locatif
n.	nom
nég.	négateur
num.	numéral
opp.	oppositif
p.	page
perd.	perduratif
pers.	personne
pers.sg.	personne du singulier
pl.	pluriel
préd.nom.	prédicatif nominal
préd.v.	prédicatif verbal
préf.	préfixe
préf.adj.	préfixe adjectival
préf.cl.	préfixe de classe

préf.dét.	préfixe déterminatif
préf.inf.	préfixe infinitif
préf.nég.	préfixe négateur
préf.obj.	préfixe objet
préf.pers.	préfixe personnel
préf.pron.	préfixe pronominal
préf.réfl.	préfixe réfléchi
préf.v.suj.	préfixe verbal sujet
prép.	préposition
pron.	pronom
rad.	radical
réfl.	réfléchi
subj.	subjonctif
subst.d.	substantifs déverbaux
subst.n.d.	substantifs non déverbaux
substit.	substitutif
suff.	suffixe
suff.appl.	suffixe applicatif
suff.apt.	suffixe aptitif
suff.ass.	sufixe associatif
suff.caus.	suffixe causatif
suff.dér.	suffixe de dérivation
suff.intens.	suffixe intensif
suff.intens.ampl.	suffixe intensif ampliatif
suff.intens.dur.	suffixe intensif duratif
suff.intens.péj.	suffixe intensif péjoratif
suff.intrans.	suffixe intransitif
suff.pass.	suffixe passif
suff.perd.	suffixe perduratif
suff.répét.	suffixe répétitif
suff.stat.	suffixe statif
suff.trans.	suffixe transitif

suff.transmut.	suffixe transmutatif
synd.	syndèse
v.	verbe

REMERCIEMENTS

Cette recherche a pu être réalisée grâce à une bourse du Gouvernement du Burundi. Je voudrais lui exprimer ma gratitude.

Mes remerciements vont ensuite à Louise Dagenais, directrice de cette thèse. J'ai beaucoup appris de sa rigueur scientifique. Merci aussi pour son soutien pendant les moments où le moral était bas.

Je remercie très sincèrement Nathan Ménard, mon codirecteur de recherche. Outre son appui financier à cette recherche, il m'a fait profiter de sa grande culture et de la profondeur de ses réflexions.

Je remercie particulièrement Nazam Halaoui pour l'information grammaticale qu'il m'a donnée et les conseils de description qu'il m'a prodigués pour mes futurs travaux sur le kirundi.

Je tiens également à remercier Christine et les enfants pour leurs encouragements et les sacrifices consentis.

Mille fois merci à Tharcisse Batungwanayo qui, au Burundi, a bravé tous les dangers pour rassembler le corpus.

Il me faut absolument remercier la Faculté des études supérieures de l'Université de Montréal pour la bourse accordée de 1993-1994 à 1994-1995.

Enfin, merci à mes amis qui m'ont permis de passer d'agréables moments entre les périodes de labeur.

À mes parents
qui m'ont envoyé à l'école
et à
Christine
qui m'a permis d'y retourner

CORRESPONDANCES GRAPHO-PHONÉTIQUES

1. LES GRAPHÈMES

	Graphie	*Phonétique*	*Exemple en kirundi*	*Représentation phonétique*	*Glose*
Voyelles brèves					
	a	[a]	***a**zi*	[azi]	'il sait'
	e	[e]	*amahe**r**a*	[amahera]	'l'argent'
	i	[i]	***i**si*	[isi]	'mycoses'
	o	[o]	*in**o***	[ino]	'orteille'
	u	[u]	*if**u***	[ifu]	'farine'
Voyelles longues					
	aa	[a:]	*kur**aa**ta*	[kura:ta]	'vanter'
	ee	[e:]	*gus**ee**sa*	[guse:sa]	'renverser'
	ii	[i:]	*gus**ii**ba*	[gusi:ßa]	's'absenter'
	oo	[o:]	*kum**oo**ta*	[kumo:ta]	'sentir'
	uu	[u:]	*gus**uu**ra*	[gusu:ra]	'fouiller'
Semi-voyelles					
	w	[w]	*a**w**a*	[awa]	'tant pis'
	y	[j]	*o**y**a*	[oja]	'non'
Consonnes					
	p	[p]	***p**ikip**p**iki*	[ipikipiki]	'une moto'
	b	[ß]	***b**a*	[aßa]	'il habite'
	t	[t]	*i**t**u*	[itu]	'paquet'
	d	[d]	*ku**d**oha*	[kudoha]	'grossir'

k	[k]	***k****u*	[ku]	'sur'
g	[g]	*umu****g****abo*	[umugaβo]	'personne mâle'
h	[h]	***h****ari*	[hari]	'il y a'
s	[s]	***s****i*	[si]	'ce n'est pas'
z	[z]	*akaz****i***	[akazi]	'travail'
sh	[ʃ]	*is****h****u*	[iʃu]	'chou'
j	[Z]	*ku****j****a*	[kuʒa]	'aller'
r	[r]	***r****i*	[ari]	'il est'
f	[f]	***f****u*	[ifu]	'farine'
v	[v]	***v****uga*	[avuga]	'il parle'
pf	[pf]	***pf****a*	[apfa]	'il meurt'
ts	[ts]	*umusha****ts****i*	[umuʃatsi]	'cheveux'
c	[tʃ]	*umu****c****eri*	[umutʃeri]	'le riz'
n	[n]	***n****o*	[ino]	'orteille'
m	[m]	***m****u*	[mu]	'dans'
ny	[ɲ]	***ny****ama*	[iɲama]	'viande'

2. SUITES DE GRAPHÈMES NON VOCALIQUES

	Graphie	*Phonétique*	*Exemple en kirundi*	*Représentation phonétique*	*Glose*
Avec les semi-voyelles /w/ et /y/					
	bw	[bg]	kuga**b** **w**a	[kugabga]	'être dominé'
	cw	[tʃgw]	i**c** **w**ende	[itʃgwe:nde]	'pot à beurre'
	pw	[pkw]	gucá**p** **w**a	[gucápᵏwa]	'être dessiné'
	tw	[tʰw]	kuraa**t** **w**a	[kura:tʰwa]	'être vanté'
	dw	[dgw]	ind **w**i	[indgwi]	'sept'
	kw	[kʰw]	gusa**k** **w**a	[gusakʰwa]	'être fouillé'
	gw	[gw]	umu**g** **w**i	[umugwi]	'groupe'
	hw	[hw]	kuri**h** **w**a	[kurihwa]	'être remboursé'

	sw	[skw]	umus**w**i	[umuskwi]	'poussin'
	zw	[zgw]	a**zw**i	[azwi]	'il est connu'
	shw	[shgw]	as**hw**aana	[aʃgwa:na]	'il se dispute'
	jw	[dʒgw]	**ijw**í	[iʒgwí]	'voix'
	rw	[rgw]	umu**rw**a	[umurgwa]	'ville'
	cw	[tʃgw]	gucac**w**a	[gutʃatʃgwa]	'être testé'
	nw	[nw]	umun**w**a	[umunwa]	'la bouche'
	mw	[mw]	kum**w**a	[mwa]	'raser les poils'
	nyw	[nyw]	an**yw**a	[aɲwa	'il boit'
	fy	[fj]	guf**y**ina	[gufjina]	'badiner'
	my	[mj]	kugum**y**a	[kugumja]	'garder'
	ry	[rgj]	**ry**a	[rgja]	'mange'
	sy	[sgj]	gus**y**a	[gusja]	'moudre'
	vy	[vgj]	kuv**y**ibuha	[kuvgjiβuha]	'grossir'
	zy	[zgj]	iz**y**ondi	[izgjo:ndi]	'minuscule'
Avec les nasales /m/ et /n/					
	mb	[mb]	ija**mb**o	[iʒa:mbo]	'parole'
	nc	[ntʃ]	**nc**a	[ntʃa]	'je passe'
	mf	[mf]	**mf**úra	[imfúra]	'aîné'
	mp	[mp^{h}]	**mp**amba	[imp^{h}a:amba]	'provision'
	mpf	[mp^{h}f]	**mpf**ú	[imp^{h}fú]	'types de décès'
	mv	[mv]	**mv**o	[imvo]	'cause'
	nd	[nd]	kugo**nd**a	[kugo:nda]	'plier'
	ng	[ŋ]	guse**ng**a	[guse:ŋa]	'prier'
	nj	[nʒ]	**nj**a	[nʒa]	'je vais'
	nk	[nk^{h}]	kuro**nk**a	[kuro:nk^{h}a]	'recevoir'
	ns	[ns]	**ns**aba	[nsaʃa]	'je demande'
	nsh	[nʃ]	**nsh**aaka	[nʃa:ka]	'je veux'
	nt	[nt]	**nt**aaha	[nta:ha]	'je rentre'
	nz	[nz]	guso**nz**a	[guso:nza]	'avoir faim'

Avec nasale et semi-voyelle

mbw	[mbg]	i**mbw**á	[imbgá]	'chien'
mfy	[mfj]	**mfy**ina	[mfjina]	'je badine'
mpw	[mp^{h}w]	**mpw**eera	[mp^{h}weera]	'je meurs'
mvy	[mvj]	**mvy**ina	[mvjina]	'je danse'
ncw	[ntʃgw]	guca**ncw**a	[guca:ntʃgwa]	'être vacciné'
ndw	[ndgw]	i**ndw**i	[indgwi]	'sept'
nkw	[nkw]	**nkw**aama	[nkwa:ma]	'je tombe en panne'
ngw	[ngw]	**ngw**a	[ŋwa]	'je tombe'
njw	[nʒgw]	**njw**iira	[nʒgwi:ra]	'je crie'
nsw	[nsw]	**nsw**aaga	[nswa:ga]	'je découpe'
nshw	[nʃw]	**nshw**aana	[nʃwa:na]	'je me dispute'
nsy	[nsgj]	**nsy**a	[nsgja:na]	'je mouds'
ntw	[ntw]	**ntw**aara	[ntwa:ra]	'j'emporte'
nyw	[ɳ̟w]	**nyw**a	[ɳ̟wa]	'je bois'
nzw	[nzgw]	**nzw**i	[nzgwi]	'je suis connu'
nzy	[nzj]	i**nzy**a	[inzja]	'poils du pubis'

INTRODUCTION

1. INTÉRÊT ET MOTIVATION DE L'ÉTUDE

Le kirundi est une langue africaine bantoue, à classes et à tons, parlée au Burundi. Elle a été l'objet de plusieurs études dont celle de Meeussen (1959) qui en a fourni la première description d'envergure. Les recherches ont surtout porté sur la phonologie (Ndayishinguje 1978) et la morphologie (Nkanira 1971, 1984). La syntaxe et le lexique du kirundi ont été moins étudiés. En lexicographie, il y a lieu de noter le dictionnaire de Rodegem (1970). On trouvera dans Ntirampeba (1993) un long descriptif de la morphologie de la langue dans l'esprit d'un traitement lexicologique.

Hormis Ntirampeba (1988), nous n'avons connaissance d'aucune étude de lexicométrie sur le kirundi[1] ou sur toute autre langue bantoue. La présente recherche veut combler cette lacune. Elle veut fournir une analyse la plus exhaustive possible des problèmes posés par les langues à classes et à tons dans l'élaboration des vocabulaires fondamentaux en même temps qu'elle en propose quelques solutions.

Plus spécifiquement, cette recherche se veut une étude de référence du kirundi écrit qui pourrait fonder la sélection des éléments lexicaux, grammaticaux et morphologiques aux fins notamment de la didactique de la langue, de l'alphabétisation, de la postalphabétisation et de la vulgarisation scientifique.

---

[1] Luc Bouquiaux du LACITO (Laboratoire de Langues et Civilisations à Tradition Orale) au CNRS nous confiait en mars 1998, ne pas avoir connaissance de travaux portant sur la statistique lexicale du kirundi ou sur le traitement automatique de la morphologie du kirundi ou d'autres langues bantoues et que ses collègues du LACITO n'en étaient pas non plus au courant. Nous remercions M. Etienne Tiffou qui nous a permis ce contact.

Une première étude substantielle quantitative du kirundi suppose qu'on résolve un problème de fond : déterminer la forme et le contenu des deux unités classiques de la lexicométrie, soit le mot-forme et le vocable. Il s'agit d'abord de définir ces deux unités et de sélectionner ensuite, à l'aide de critères rigoureux, une liste de vocables constitutifs du noyau fondamental de la langue écrite.

En outre, la morphologie complexe du kirundi, qui s'articule autour de radicaux à contenu lexical stable, invite à pénétrer en deçà des frontières du mot-forme. Nous estimons nécessaire d'inclure dans notre recherche l'établissement d'index de fréquences de morphèmes sélectionnés en fonction des objectifs pédagogiques qui nous guident.

2. CHAMP D'ÉTUDE ET APPLICATIONS

La présente recherche s'inscrit dans le cadre de la linguistique quantitative. Celle-ci constitue une branche de la linguistique mathématique.

L'on peut distinguer en linguistique mathématique, en fonction des méthodes utilisées, deux grands types d'études sur le langage. Le premier type rassemble des études qui utilisent des méthodes quantitatives (statistique, calcul des probabilités, etc.) et le second celles qui se fondent sur des méthodes non-quantitatives (algèbre, théorie des graphes, logique, etc.). Notre étude appartient à la première catégorie.

La linguistique quantitative (ou statistique linguistique) se propose d'aborder les unités des langues naturelles dans leur aspect numérique et opère à tous les niveaux où il est possible de concevoir des unités discrètes de diverses natures (phonologie, morphologie, etc.).

On peut citer pour le niveau phonétique / phonologique les études de Deweye (1923) et Roberts (1965) pour l'anglais, celles de Hug (1979) pour le français et celle de De Colombel (1986) pour l'ouldémé, une langue de l'Afrique occidentale. Ndayishinguje (1978) consacre quant à lui un chapitre à la fréquence des phonèmes du kirundi.

En statistiques grammaticales, les études sont menées suivant deux volets : le premier porte sur la morphologie, le second sur la syntaxe. En morphologie

quantitative, l'on peut citer les travaux de Robert (1960), Dubois (1962 a) et Lapierre (1972) pour le français ainsi que celui de Baayen (1989) pour l'anglais.

Aucune étude n'a fourni, à notre connaissance, de données quantitatives sur la morphologie du kirundi. En étudiant la fréquence des morphèmes sur la base d'un corpus d'environ 100 000 mots, la présente étude se veut une contribution à une meilleure connaissance de cet aspect de la langue. Nous y revenons au § 2.2.

Parmi les études de syntaxe quantitative, on peut citer entre autres l'étude de Barth (1961) sur le français, l'anglais et l'espagnol, la deuxième partie de l'étude de Gougenheim *et al.* (1964), de même que Roy (1976), Jolivet (1982) et Greidanus (1990), toutes des études sur le français.

Les recherches statistiques en sémantique sont encore embryonnaires : la sémantique à été restreinte sinon évacuée des recherches en linguistique quantitative. Elle bute notamment sur la nature des unités à quantifier (sèmes, noèmes, primitifs sémantiques) qui ne semblent pas *a priori* réductibles à des unités discrètes (Ménard 1989 : 468).

Selon Tesitelova (1992 : 135-138), les études de statistique sémantique tentent de quantifier les différents sens des unités lexicales comme le dictionnaire de West (1953) ou d'estimer les relations lexico-sémantiques dans des corpus (Juilland 1985, Ménard 1989).

Quant à la statistique lexicale, elle constitue un domaine de recherche très florissant en linguistique quantitative. Elle est tantôt appelée *lexicostatistique*, *lexicométrie* ou *lexicologie quantitative*. Ce dernier terme est peu usité en linguistique; nous ne le retenons pas. Nous écartons également de la terminologie de cette étude le terme « lexicostatistique » compte tenu de son utilisation en glottochronologie[2]. Nous retenons les termes synonymes de *statistique lexicale* et de *lexicométrie*.

---

[2] La glottochronologie est une technique utilisée pour établir l'époque à laquelle deux ou plusieurs langues apparentées se sont séparées de la langue originaire commune. Elle repose sur l'idée suivante : si à un moment une langue donne naissance à deux langues différentes, celles-ci perdent, en évoluant, des mots de leur fonds primitif commun et les remplacent par des emprunts ou des créations nouvelles. Il suffirait alors de compter le nombre de mots qui dans une série donnée n'appartient pas à la langue d'origine pour déterminer approximativement l'époque où un idiome s'est séparé de la langue-mère (cf. Évrard 1966 : 85-86).

2.1. LA STATISTIQUE LEXICALE

Selon Muller (1977 : 10), on parle de statistique lexicale pour désigner une étude quantitative des mots d'un corpus. Cette étude procède par le rattachement des mots aux vocables. Dugast (1980 : 5) indique quant à lui que la statistique lexicale est « l'étude de l'organisation du vocabulaire dans le corpus » tandis que pour Dubois *et al.* (1994) elle est « l'application des méthodes statistiques à la description du vocabulaire d'un texte ».

Il se dégage de ces définitions que la statistique lexicale a un objet d'étude (le vocabulaire), une méthode d'analyse (la statistique) et des unités d'analyse (le mot et le vocable).

Notre recherche, qui vise à élaborer un vocabulaire de base du kirundi écrit, relève donc de la statistique lexicale. Nous présentons au chapitre 2 le corpus que nous utilisons, les unités d'analyse que nous retenons et les résultats statistiques que nous en attendons.

Les recherches en statistique lexicale sont menées selon deux principaux axes : un axe descriptif et un axe pédagogique. Nous les abordons dans cet ordre.

2.1.1. LES ÉTUDES DE STATISTIQUE LEXICALE DESCRIPTIVES

Les études de statistique lexicale descriptives trouvent de nombreuses applications en lexicographie, en stylistique, en recherches sur l'attribution des textes, en typologie, en dialectologie et en sociolinguistique.

En lexicographie, la statistique lexicale permet de mettre au point des dictionnaires de fréquence. L'on peut citer pour le français les études de Juilland *et al.* (1970), les sept volumes du *Dictionnaire des fréquences* du Trésor de la Langue Française (CNRS 1971). Pour le français québecois, Vikis-Freibergs (1974), Beauchemin & Martel (1979), Baudot (1992) et Beauchemin *et al.* (1992) illustrent ce type de recherches alors que Kucera & Francis (1967) le fait pour l'anglais.

Lorsque le but n'est pas de constituer un dictionnaire mais de caractériser un état de langue, un auteur ou un corpus, la statistique lexicale permet d'élaborer des listes de fréquences des unités lexicales. Entrent dans cette catégorie les travaux de Dubois (1962 b) et Brunet (1981) pour le français, et ceux de Hofland (1992) et Johansson & Hofland (1989) pour l'anglais.

La statistique lexicale est aussi utilisée dans les études stylistiques. Elle permet, à certaines conditions, de caractériser le vocabulaire d'un auteur ou d'un locuteur particulier. C'est dans cette ligne que se situent les travaux de Yule (1944), Muller (1968), Brunet (1983, 1985, 1988), Ménard (1983), Labbé (1990), Cossette (1994) ainsi que De Chantal (1997) qui porte, en outre, sur la typologie des textes.

La statistique lexicale sert également dans l'attribution d'auteurs en cas de conflit pour déterminer l'auteur d'une œuvre littéraire. Ces recherches sont fondées sur des données quantitatives relatives notamment à la longueur du mot et de la phrase. Citons dans cette veine les travaux de Ellegard (1962), Debièvre (1977) et Dubrocard (1985).

Les données de statistique lexicale sont aussi utiles à la dialectologie. On peut en effet quantifier la dispersion des unités lexicales dans un espace géographique et en tirer des conclusions sur l'évolution de la langue. Guillaume *et al.* (1978) ressortit par exemple à ce type de recherche. Les recherches en glottochronologie (cf. Dyen 1975) procédaient de la même logique.

La sociolinguistique met également à contribution la statistique lexicale. Les choix lexicaux et grammaticaux opérés par les locuteurs peuvent être quantifiés et servir d'indicateurs sociolinguistiques. En témoignent les travaux de Sankoff & Cedergren (1971), Labov (1972 : 302-316), Ménard & Santerre (1979), Martel (1986) et Paquot (1988).

La statistique lexicale se prête aussi à l'étude de l'évolution des langues. L'on sait par exemple (cf. Manczak 1966 : 99-104) que les mots courts (qui sont généralement les plus fréquents) se maintiennent mieux que les mots longs dans la langue. De même, entre deux allomorphes, c'est le plus fréquent qui survit (Bybee 1994). Il existerait donc un lien entre la fréquence des unités linguistiques et leur évolution.

L'analyse du discours recourt également à la statistique lexicale. Les données de statistique lexicale permettent d'analyser objectivement certaines notions comme la cohésion lexicale et de dégager des indices pour la caractériser (Ménard 1988). Les travaux du Laboratoire « Lexicométrie et textes politiques » de l'École Normale Supérieure de Fontenay-Saint-Cloud menés notamment par Lebart & Salem (1988, 1994) proposent des analyses quantitatives des réponses à des questions ouvertes posées lors d'enquêtes. Nous reviendrons sur les choix lexicométriques du groupe de Saint-Cloud au chapitre 1 § 5.4.1.

En traduction, les données sur la fréquence des unités lexicales sont essentielles dans la mise au point de programmes informatiques de traduction automatique. Ces programmes se servent de dictionnaires sources et cibles renfermant des unités lexicales sélectionnées selon leurs fréquences et qui permettent de se faire une idée plus ou moins précise de la probabilité du sens d'une unité. Ainsi, selon Brown *et al.* (1990 : 79-86), le mot anglais *the* a une probalité de 0,788 d'être traduit en français par les mots « le » ou « la », les autres possibilités (*l', les, ce*, etc.) n'ayant qu'une probabilité de 0,212.

Signalons ici la thèse de Bélanger (1992). Ce travail fournit une analyse statistique des éléments cohésifs grammaticaux à partir de la comparaison de textes originaux en anglais et de textes traduits en français. Dans la foulée, Bélanger (1992) analyse dans une perspective typologique, la répartition des éléments cohésifs dans des textes de spécialité.

2.1.2. LES ÉTUDES DE STATISTIQUE LEXICALE À VISÉES PÉDAGOGIQUES

La statistique lexicale est mise à contribution dans des études linguistiques menées à des fins pédagogiques. L'on peut dégager deux orientations : la première rassemble des recherches sur les vocabulaires des élèves de l'école primaire ou secondaire et la seconde s'intéresse à l'élaboration des vocabulaires de base. Nous rangeons dans la première catégorie les recherches psycholinguistiques menées sur le vocabulaire des enfants d'âge préscolaire.

2.1.2.1. LES RECHERCHES PSYCHOLINGUISTIQUES

En psycholinguistique, des études de statistique lexicale sur l'acquisition du lexique pemettent de juger du retard ou de la précocité du langage d'un enfant et de diagnostiquer certains troubles comme l'aphasie amnésique[3].

Dans une étude menée sur l'anglais, Clark (1993 : 13) estime que, à un an, un enfant qui a un développement langagier normal utilise entre 50 et 200 mots, et à deux ans entre 500 et 600 mots. Vers 6 ans, ces chiffres atteignent 14 000 mots alors qu'un adulte dispose d'entre 20 000 et 50 000 mots. De nombreuses études dont celles de Moreau & Richelle (1981) et MacNamara (1982) fournissent des données sur le vocabulaire de l'enfant et son acquisition dès la tendre enfance. Elles montrent notamment que les noms y occupent une place de choix (François 1978 : 116). D'autres études portent sur la structure morphologique du vocabulaire identifié pour chaque âge (Anglin 1993) et sur le rôle de la morphologie dans l'apprentissage et l'enseignement des langues (Nation 1994).

Selon Cordier (1994 : 70), une des tâches des recherches en psycholinguistique serait d'arriver à estimer le rôle de la fréquence des unités lexicales (et des morphèmes) dans l'apprentissage du langage. Les recherches de Bradley (1983) et Gordon (1983) montrent par exemple qu'il existe une corrélation négative entre la fréquence d'un mot et le temps que les locuteurs mettent à réagir à ce mot dans une tâche de décision lexicale. Plus le mot est fréquent, plus la réaction est rapide.

D'autres recherches, comme celles de Dhal (1979) et de Howes (1966) appliquent la statistique lexicale à la psychanalyse. Les corpus analysés par ces chercheurs ont été recueillis au cours d'entrevues entre patients et psychanalystes.

Outre ces études psycholinguistiques, d'autres études de statistique lexicale ont été menées sur le vocabulaire des élèves de l'école primaire ou secondaire. Le but de ces études est d'établir des échelles et une progression pour les programmes d'enseignement des langues et le matériel didactique.

---

[3] L'aphasique amnésique manque de mots dans son langage et la dénomination d'objets ou d'images présente des déficits (Dubois *et al.* 1994 : 42).

En Europe, on peut citer les travaux d'Aristizabal (1938), de Dottrens & Massarenti (1963), Ters *et al.* (1964) et Rivière (1979). Relèvent également de cette catégorie les études québécoises de Préfontaine & Préfontaine (1968), Primeau & Labelle (1981), Gratton & Barbaud (1981), Rousseau (1985) et Fortier (1993). Aux États-Unis, on peut citer notamment Risland (1945), Carroll *et al.* (1971) et Johnson *et al.* (1983).

2.1.2.2. LES VOCABULAIRES DE BASE

La statistique lexicale est mise à contribution dans l'élaboration des vocabulaires de base pour l'enseignement des langues maternelles et étrangères. Il s'agit de dégager les mots les plus employés de la langue pour en assurer l'enseignement prioritaire.

Les études de Henmon (1924), Ward (1926), Cheydleur (1929), Vander Beke (1929), Haygood (1936), Tharp *et al.* (1939), Vinette (1943), Thorndike & Lorge (1944), Carrière (1952), Dottrens & Massarenti (1963), Gougenheim *et al.* (1964), Mackey *et al.* (1970), Juilland *et al.* (1970), Tashdjan (1972), Rivière (1979) et Beauchemin & Martel (1979) s'inscrivent dans cette voie.

Notre étude se situe dans ce dernier cadre puisque nous voulons dégager un vocabulaire de base du kirundi écrit susceptible d'être utilisé dans l'enseignement de la langue, dans la vulgarisation scientifique et dans la post-alphabétisation.

Cette recherche a pour objet le niveau lexical. Nous en avons posé les jalons dans Ntirampeba (1988, 1993). Ce sont, à notre connaissance, les seules études abordant les problèmes de statistique lexicale du kirundi et même d'une langue bantoue.

Comme on peut le constater à travers l'inventaire des travaux déjà publiés, de nombreuses recherches ont été réalisées en statistique lexicale. Elles partagent toutes un même fondement - le corpus - et visent l'étude d'un ou de plusieurs aspects du vocabulaire. Ils diffèrent cependant quant à leur portée. Celle-ci dépend notamment des moyens de dépouillement et de la taille des corpus à analyser. Nous abordons ces deux aspects au § 2.1.3. où nous traitons de l'évolution des technologies en statistique lexicale.

2.1.3. L'ÉVOLUTION DES TECHNOLOGIES EN STATISTIQUE LEXICALE

Si l'on considère les moyens de dépouillement utilisés en statistique lexicale, on peut distinguer trois cas : le dépouillement manuel, le dépouillement mécanographique et le dépouillement semi-automatique.

Le dépouillement manuel caractérise les débuts de la statistique lexicale. D'importants travaux ont été réalisés par ce biais, à la fin du 19e siècle et au début du 20e siècle, notamment par F.W. Käding et Jean-Baptiste Estoup tel que les rapporte Dugast (1980 : 9-10). Ces premiers travaux de statistique lexicale étaient orientés vers des applications en sténographie. Les corpus des travaux étaient assez volumineux et mobilisaient plusieurs chercheurs. L'étude de Henmon (1924) portait par exemple sur 400 000 occurrences, celle de Vander Beke (1929) sur 1 547 748 occurrences et celle de Gougenheim *et al.* (1964) sur 312 135 occurrences. Guiraud a également réalisé d'importants dépouillements manuels dont de nombreux index (Guiraud 1955). Nous avons quant à nous expérimenté le dépouillement manuel dans Ntirampeba (1988).

Le dépouillement manuel constitue un travail pénible et fastidieux qui oblige souvent à travailler sur un corpus de taille réduite. Il présente cependant l'avantage de mettre directement le chercheur en contact direct avec le texte et de lui permettre, par un mouvement d'aller et retour, de formuler un jugement sur chaque mot du corpus. Ainsi, chaque mot est lu, désambiguïsé et rattaché à son lemme.

Les années 50 sont marquées, en statistique lexicale par les dépouillements mécanographiques; les mots sont enregistrés sur cartes perforées de 80 colonnes à raison d'un caractère par colonne. En 1959, le fichier mécanographique du laboratoire du Centre d'Étude du Vocabulaire Français de Besançon (France) comptait déjà 4 millions de cartes (Quémada 1962 : 58-63), représentant un dépouillement de textes du XVe - XXe siècle . Selon Quémada (1962 : 90-95), la préparation de l'allemand fondamental et du vocabulaire scientifique général du français s'est faite sur des cartes perforées. Aux États-Unis, l'ouvrage de Thorndike & Lorge (1944) est fondé sur le dépouillement d'un corpus de 2 millions d'occurrences.

Si la technologie des cartes perforées permettaient de travailler sur des corpus plus volumineux que ceux dépouillés manuellement, les opération de saisie de mots (perforation des cartes), de vérification et de triage de cartes étaient longues

et rebutantes. De plus, la désambiguïsation était faite manuellement lors de la saisie, ce qui allongeait le temps de traitement.

Dès les années 60, les ordinateurs font faire des pas de géant à la statistique lexicale. La vitesse de traitement s'accroît ainsi que la taille des corpus. Le *Dictionnaire des fréquences* du T.L.F., qui porte sur près de 71 millions d'occurrences, en constitue la meilleure illustration.

L'ordinateur permet également de travailler sur des tranches du corpus et de procéder à des calculs impossibles à réaliser autrement (cf. Brunet 1981).

Le travail va de l'enregistrement informatique des textes (Lafon *et al.* 1985) à l'interprétation des résultats en passant par des analyses spécifiques réalisées grâce à des logiciels de traitement de données linguistiques.

Reste la question de la désambiguïsation sous-jacente à tout ce processus. D'une part, on ne peut pas vérifier par exemple chacune des 70 273 552 millions d'occurrences[4] qui ont servi au *Dictionnaire des fréquences* du T.L.F.; de l'autre, on n'a pas les moyens d'une désambiguïsation automatique qui permettrait de se passer totalement de toute intervention humaine. Une telle entreprise exigerait beaucoup de sous-programmes pour traiter chaque cas d'ambiguïté.

De plus, comme aucun corpus ne peut épuiser toutes les occurrences des mots d'une langue, ne pas recourir à l'intervention du chercheur entraînerait des risques d'erreur. C'est ce constat qui amène encore certains chercheurs à désambiguïser manuellement des corpus assez volumineux[5] ou à mettre au point des logiciels qui, quoique très performants, laissent quand même la place à l'intervention humaine (Beauchemin & Théoret 1984). De tels logiciels permettent de réaliser des dépouillements semi-automatiques.

Pour ce qui est spécifiquement de la désambiguïsation, on confie la levée de certaines ambiguïtés à l'ordinateur et les ambiguïtés restantes, assorties de leurs contextes, sont traitées manuellement. On réduit ainsi les interventions manuelles tout en gagnant sur la taille des corpus.

---

[4] Citons également l'étude de Beauchemin *et al.* (1992) sur le français québécois (1 000 000 d'occurrences), celle de Kucera & Francis (1967) sur l'anglais américain (1 014 232 occurrences) et celle de Brunet (1988) sur le vocabulaire de Victor Hugo (2 074 228 occurrences).

[5] Baudot (1992) a par exemple désambiguïsé manuellement son corpus (1 040 150 occurrences).

Nous indiquons au chapitre 2 les moyens technologiques utilisés et les décisions que nous avons dû prendre pour mener à bien les dépouillements lexical et morphologique de notre corpus.

Rappelons au passage que notre étude comporte un volet de morphologie quantitative. Nous estimons nécessaire d'inclure dans notre recherche, l'établissement d'index de fréquences de morphèmes. Notre étude s'inscrit de ce fait dans les études de « statistique morphologique » appelée aussi « statistique grammaticale ».

2.2. LA STATISTIQUE MORPHOLOGIQUE

La morphologie d'une langue peut faire l'objet de quantification. Nous utilisons le terme « morphologie » dans le sens de « morphologie du morphème » (Corbin 1987 : 182-183). La « morphologie du morphème » (par opposition à la morphologie du mot et du mot-morphème)[6] postule que la formation des mots opère à partir de morphèmes (radicaux et affixes) qui peuvent, dès lors qu'ils sont identifiés, faire l'objet d'analyses statistiques.

Dubois (1962 a) a, par exemple, analysé les mouvements des suffixes du français entre 1906 et 1960 sur base de calculs statistiques. Muller (1979 a : 57-72) a travaillé sur l'utilisation des personnes verbales en français et sur les tempa verbaux du français (Muller 1979 b). Vaneste (1988) a analysé, à partir de listes de fréquences de mots, les règles morphologiques les plus appliquées (rendement d'une règle) et l'ordre dans lequel elles le sont. Faitelson-Weiser & Gingras (1992) ont étudié la disponibilité suffixale en espagnol.

Notre étude porte sur l'ensemble de la morphologie du kirundi. Nous étudions aussi bien la fréquence des affixes (préfixes et suffixes) que celle des radicaux, qu'ils soient verbaux, nominaux ou adjectivaux.

Une telle étude ne pouvait être menée manuellement. Le traitement du corpus a nécessité le recours à l'ordinateur et la mise au point d'un analyseur morphologique qui sert à la reconnaissance des morphes afin de fournir les morphèmes correspondants. L'usage de ces technologies situe notre étude en

---

[6] Cf. Corbin (1987 : 182-183).

linguistique informatique et plus spécifiquement dans un nouveau domaine de recherche appelé « morphologie computationnelle » (Sproat 1992).

2.3. LA MORPHOLOGIE COMPUTATIONNELLE

2.3.1. DATATION DU SYNTAGME « X *COMPUTATIONNEL* »

Le terme « computationnel » date du début des années soixante[7]. Il a été introduit en français en même temps qu'en anglais par Bernard Vauquois du CETA (Centre d'études sur la traduction automatique - aujourd'hui GETA - à Grenoble). Il parlait alors de « linguistique computationnelle ».

Plus tard, des puristes, y voyant un anglicisme (alors que *computationnel* est un dérivé d'un vieux mot français *computation*, lui-même emprunté au latin médiéval *computatio* 'calcul' et tout à fait conforme à la morphologie du français actuel) ont voulu le remplacer par « linguistique informatique » ou « linguistique calculatoire », termes beaucoup trop restrictifs[8]. Rey *et al.* (1992 : 463) émettent aussi des doutes sur le statut d'emprunt du terme « computationnel » à l'anglais même si des puristes l'ont considéré comme tel.

C'est à la fin des années 80 qu'on a commencé à appliquer l'adjectif « computationnel » aux sous-disciplines de la linguistique : phonologie, morphologie, syntaxe, sémantique, etc.

2.3.2. LA MORPHOLOGIE COMPUTATIONNELLE EN TRAITEMENT AUTOMATIQUE DES LANGUES NATURELLES

La morphologie « computationnelle » est une branche de la linguistique qui utilise l'ordinateur aux fins du traitement automatique de la morphologie des langues naturelles. Le traitement automatique des langues naturelles (désormais TALN) requiert la mise au point de programmes informatiques capables de traiter

---

[7] Communication personnelle du 11 mars 1997 de M. Yves-Charles Morin, professeur titulaire au département de linguistique et de traduction de l'université de Montréal. Nous le remercions de son aide.

[8] Le nom *computation* est lui-même formé sur le nom « comput » (1854) emprunté au latin « computus » 'calcul' / 'décompte'.

les données linguistiques (King 1995 : 7) à tous les niveaux de l'analyse (phonologie, morphologie, syntaxe, etc.).

Le TALN constitue un volet important des recherches en intelligence artificielle (désormais IA). Le but de l'IA est d'arriver à simuler l'activité mentale et de fournir des instructions à la machine de manière à ce qu'elle exécute certaines tâches selon une certaine logique, comparable au cheminement de la pensée humaine. En témoignent les recherches sur les jeux (échecs par exemple), les systèmes-experts[9], les robots, etc. Dans ces recherches, les questions théoriques sont intimément liées aux applications.

De façon spécifique, il s'agit en TALN de modéliser d'abord le fonctionnement du langage humain (King 1995 : 7) et de mettre ensuite au point des produits commerciaux. L'on trouvera dans OFIL (1994) un répertoire des produits disponibles en rapport avec le traitement automatique de nombreuses langues avec mention des composants linguistiques (morphologie, lexique, etc.) sur lesquels ils sont construits[10].

---

[9] Un système-expert est un logiciel qui, à partir d'une base de connaissances, formalise l'acquis intellectuel et les modes de raisonnement d'experts dans un domaine technique, porte des diagnostics sur les disfonctionnements ou propose des solutions à des problèmes qui lui sont fournis (Wijnands 1989 : 502).

[10] Des contacts personnels nous ont permis d'apprendre que quelques logiciels ont été développés pour le traitement de certaines langues bantoues.

À l'Université de Leiden (Pays-Bas), au département des langues africaines, M. Thilo C. Schadeberg a développé le logiciel *AINI*, un analyseur morphologique pour le swahili standard.

À Naples, Maddalena Toscano (*Istituto Universitario Orientale, dipartimento di studi e ricerche su Africa e Paesi Arabi*) travaille à la mise au point d'un lemmatiseur du zulu (langue bantoue parlée en République Sud-Africaine).

Doris Payne (*University of Oregon*) utilise le logiciel *Shoebox* 4.0 pour traiter le lunda (langue bantoue de Zambie) et le Massai (Kenya, Tanzanie).

Mais le traitement automatique du langage ne se fait pas sans heurts. Nous passons ici en revue les principales difficultées rencontrées dans le traitement des langues à morphologie complexe dont fait partie le kirundi. Nous indiquons aussi quelques-unes de ses applications.

2.3.2.1. LES PRINCIPALES DIFFICULTÉS

L'on admet généralement que les langues du monde se répartissent, eu égard à leur morphologie, en langues isolantes, flexionnelles (ou fusionnelles), agglutinantes et polysynthétiques (cf. Sapir 1921, Meillet & Cohen 1924, Spencer 1991 : 37-38).

Les deux premières catégories (langues isolantes et flexionnelles) regroupent les langues à morphologie pauvre. Parmi les langues isolantes, on peut citer le vietnamien et le chinois tandis que les langues flexionnelles sont par exemple le latin et le russe.

Les langues à morphologie riche sont les langues agglutinantes et les langues polysynthétiques. Les premières sont caractérisées par une tendance à avoir de longs mots-formes formés de morphèmes agglutinés autour du radical. Le hongrois, le turc, le swahili et le kirundi en sont des illustrations. Les secondes sont caractérisées également par de longs mots-formes mais formés par la concaténation d'unités lexicales. Les langues amérindiennes, tel le makah (Gill & Renker 1992 : 185-205), en sont des exemples.

Signalons cependant que les langues présentent souvent des caractéristiques de plusieurs de ces types linguistiques. Ainsi, quoique flexionnels, le français et l'allemand ont par exemple une composition qui les rapproche des langues polysynthétiques[11]. De même, des langues polysynthétiques comme le chukchee (Spencer 1991 : 39) ont un système dérivationnel agglutinant.

Les problèmes cruciaux de traitement automatique posés par les langues à morphologie complexe sont de deux types : le choix du lemme et la complexité de l'analyse morphophonologique.

---

[11] Ainsi par exemple, tout comme dans les langues polysynthétiques, les composés français *cessez-le-feu, porte-parole* résultent d'une concaténation d'unités lexicales.

Pour la famille des langues salish caractérisées par une forte suffixation (par exemple en makah, le mot-forme peut incorporer jusqu'à douze morphèmes différents), la complexité du mot-forme a contraint Thompson, Thomson & Hsu (1992 : 3-31) et Kinkande (1992 : 31-47) à déterminer le lemme en faisant abstraction de certains faits comme la réduplication et les infixes prévisibles. L'unité en vedette de leur dictionnaire électronique est le radical, suivi dans l'entrée par les dérivés non prévisibles (Thompson, Thomson & Hsu 1992 : 8). Nous présentons pour le kirundi, au chapitre 2 la norme de lemmatisation que nous adoptons.

Quant à la complexité de l'analyse morphophonologique, elle est telle, pour les langues salish par exemple, que la lemmatisation constitue à elle seule une entreprise de longue haleine :

> « *A complete lemmatization procedure for a salishan language amounts to having the computer do a complex morphophonemic analysis. Such a procedure may be possible to explicitly describe, but it would certainly be a major project in itself* » (Montler 1992 : 92).

Nous traitons au chapitre 3 de la complexité morphophonologique du kirundi et de l'analyseur morphologique mis au point pour l'aborder. Voyons maintenant les principales applications du traitement automatique de la morphologie des langues naturelles.

2.3.2.2. LES APPLICATIONS

La morphologie computationnelle trouve de nombreuses applications, notamment dans l'étude de la parole et du langage écrit, en recherche documentaire et en traitement de textes (Sproat 1992 et Haton & Haton 1989). L'on comprendra que toutes ces recherches sont susceptibles de nombreuses exploitations dans les milieux éducatifs et professionnels.

2.3.2.2.1. Les recherches sur la parole

Dans l'étude de la parole, les recherches visent la reconnaissance et la synthèse. L'on trouvera dans Tubach (1989) un historique de la recherche dans le domaine.

En mode reconnaissance (parler à une machine et se faire comprendre), les recherches s'orientent aujourd'hui principalement vers la reconnaissance de la parole indépendamment du locuteur et vers la dictée automatique (Tubach 1995 : 42-47). Les données morphologiques de la langue deviennent indispensables : la machine a besoin de connaître le système de formation des mots afin de les prendre en compte quand ils apparaissent dans les textes. Peut-on imaginer en effet une dictée automatique en français si la machine en ignore la flexion verbale et ne différencie pas le futur du conditionnel ?

En mode synthèse (faire parler une machine), il s'agit de passer du texte écrit au son grâce à la machine. Il devient dès lors nécessaire de connaître les particularités morphologiques d'une langue pour prévoir, par exemple, la place des accents, des tons, etc.

2.3.2.2.2. Les recherches sur la langue écrite

Les recherches sur le langage écrit portent essentiellement sur la traduction automatique, la génération et l'analyse de textes, le résumé automatique, la génération de textes à partir de données non linguistiques (tableaux, graphiques, etc.), la production de dictionnaires électroniques et de lemmatiseurs.

Les informations morphologiques sont indispensables dans la plupart de ces domaines. C'est le cas en traduction automatique de l'anglais au français (King 1995 : 14) ou du japonais écrit au chinois écrit (Sproat 1992 : 9-10). L'on ne peut

pas en effet envisager une génération de texte sans module morphologique à moins de faire croître exponentiellement le lexique. La mise au point de lemmatiseurs requiert également des informations morphologiques.

2.3.2.2.3. La recherche documentaire

La recherche documentaire recourt à la morphologie de la langue notamment dans l'accès à des bases de données textuelles et en indexation automatique.

Il s'agit par exemple, à partir d'un mot-clé, d'accéder à différents textes qui contiennent ce mot dans une base de données. Si la tâche est aisée pour les langues à morphologie pauvre, elle est ardue pour des langues à morphologie complexe. Le système doit générer tous les mots possibles à partir d'une racine afin que l'usager puisse retrouver tous les mots apparentés au mot-clé mais qui en diffèrent par la flexion. Cette génération se fait grâce aux données morphologiques. Attar *et al.* (1978) illustrent ce type de recherche.

Les données morphologiques sont également utiles en indexation automatique. Celle-ci recourt en effet à un analyseur automatique de textes qui dispose d'un module morphologique. Le travail de l'analyseur (Deweze 1989 : 194) consiste à lire le texte (titre ou résumé d'un document par ex.), à en comparer les mots avec un lexique de son dictionnaire et à réduire les mots fléchis du texte à des formes lexicales canoniques. Les données morphologiques sont fondamentales dans cette analyse.

2.3.2.2.4. Le traitement de textes

La morphologie intervient dans la mise au point de correcteurs d'orthographe, de correcteurs grammaticaux et de conjugueurs. Les informations sur le pluriel ou le féminin des noms doivent être fournies à la machine afin qu'elle s'en serve quand elle fait des suggestions aux usagers. Ces informations relèvent de la morphologie.

Nous pouvons donc conclure que la morphologie est au centre de nombreuses applications dans le traitement automatique des langues naturelles. Ces applications sont fondées sur la mise au point de programmes d'analyse morphologique appelés « analyseurs morphologiques ».

2.3.2.3. LES ANALYSEURS MORPHOLOGIQUES

Un analyseur morphologique est un logiciel développé pour des fins de reconnaissance ou de synthèse des mots-formes d'une langue naturelle (Karlsson 1994 : 2570). Il s'agit dans le premier cas de reconnaître des mots-formes, les morphes et les morphèmes qui les composent, d'attribuer des traits morphosyntaxiques aux mots-formes comme la catégorie grammaticale, etc.

Dans le second cas, c'est-à-dire en synthèse, il s'agit de faire le chemin inverse; à partir de morphèmes et d'un ensemble de règles, le programme informatique doit arriver à construire des mots-formes de la langue.

Nous présentons au chapitre 2 l'analyseur morphologique utilisé pour traiter nos données. Il nous suffit de dire pour l'instant que c'est un logiciel de reconnaissance. On peut distinguer deux grands types de logiciels de reconnaissance : les catégoriseurs et les logiciels de segmentation.

Certains catégoriseurs sont des analyseurs construits à partir de données statistiques probabilistes sur les mots-formes tel que celui de Brill (cf. Martin 1997) utilisé pour le traitement de corpus à l'INALF. On recourt à la fréquence et à des règles contextuelles pour attribuer une catégorie grammaticale à un mot-forme et séparer des homographes. Cette opération s'appelle l'étiquetage (*tagging*).

Quant aux logiciels de segmentation, ils sont construits à partir d'une description de la correspondance formelle entre une représentation lexicale et les formes de surface des mots-formes. Il s'agit de fournir des mots-formes à un analyseur qui en propose une structure morphologique.

La majorité des analyseurs de reconnaissance morphologique se situent dans cette catégorie. Ils sont élaborés à partir d'un modèle morphologique à deux niveaux (représentations lexicales et formes de surface), niveaux reliés par des règles qui rendent compte des différentes modifications morphophonologiques à l'intérieur du mot-forme (Koskenniemi 1983). Ils privilégient le modèle morphologique IA (*Items et Arrangements*).

L'on a construit sur ce modèle des analyseurs morphologiques pour plus d'une vingtaine de langues dont le français (analyseur Flemm)[12], l'anglais (ENGTWOL[13]), l'estonien (Kaalep 1997), l'espagnol (Tzoukermann & Liberman 1990), le suédois (Karlsson 1992), le turc (Oflazer 1994), le basque (Alegria *et al.* 1996), le grec (Sgarbas *et al.* 1995), etc. L'on trouvera dans Frandin (1994) une synthèse de ce modèle qui se fonde sur un formalisme rigoureux de règles phonologiques.

L'analyseur du kirundi est construit selon le modèle IA et sur douze règles morphophonologiques qui rendent compte des écarts entre les mots-formes en surface et leur structure morphologique. Nous le décrirons en détail au chapitre 2.

Disons pour l'instant que l'analyseur morphologique du kirundi - désormais AMK - permet d'attribuer une catégorie grammaticale à chaque mot-forme du corpus : 0 pour les noms, 1 pour les verbes et 2 pour les adjectifs. Il permet également la segmentation des mots-formes de la langue. Il combine ainsi les fonctions des deux types d'analyseurs décrits, qui rappelons-le, sont des catégoriseurs et des outils de segmentation de mots-formes en entrée.

L'analyse automatique de la morphologie des langues débouche entre autres choses sur des travaux de morphologie quantitative qu'il aurait été impossible de réaliser manuellement. Ces données quantitatives sont utilisées en didactique des langues (Nation 1994 : 2582-2584). En effet, en mettant l'accent sur les morphèmes fréquents et réguliers, l'enseignement du vocabulaire peut s'en trouver facilité suite à la réduction du nombre d'items à apprendre.

Que ce soit au niveau lexical ou morphologique, notre étude est articulée sur un ensemble d'hypothèses de recherche, soit une hypothèse générale et trois spécifiques. Nous les présentons dans cet ordre.

---

[12] cf. « Franlex. Liens en morphologie computationnelle », http://m17.limsi.fr/individu/jacquemi/ FRANLEX/Liens.html.

[13] Cf. « Indexing and retrieval. Morphological analysis », http://www.lingsoft.fi/en/indexing/.

3. HYPOTHÈSES DE RECHERCHE

3.1. HYPOTHÈSE GÉNÉRALE

Le kirundi est une langue agglutinante, à classes nominales et à tons. Malgré des caractéristiques particulières, le kirundi, en tant que langue naturelle, doit partager sur les plans lexical et morphologique certaines caractéristiques quantitatives universelles avec les autres langues du monde.

Sur le plan lexical, ces traits sont les suivants (Guiraud 1960 : 19) :

- un petit nombre de mots grammaticaux ont les plus hautes fréquences;
- les mots les plus courts sont les plus fréquents;
- les formes simples sont plus fréquentes que les formes complexes.

Sur le plan morphologique, certaines recherches comme celles de Carstairs-McCarty (1994) et Bauer (1994) ont identifié des catégories morpho-sémantiques universelles (temps, aspect, causation, possession, etc.) et des recherches typologiques, telle celle de Greenberg (1966 : 25-55), ont permis d'étudier sur la base des fréquences la hiérarchie de ces catégories. Les fréquences des morphèmes du kirundi respectent-elles les tendances universelles identifiées ?

Notre hypothèse générale est que la fréquence des unités lexicales et des morphèmes du kirundi respecte les faits généralement observés pour d'autres langues. Pour vérifier cette hypothèse, nous exploiterons les données disponibles dans certaines études typologiques, notamment celle de Greenberg (1966), et les ouvrages de morphologie de Mel'cuk (1993, 1994, 1996).

3.2. HYPOTHÈSES SPÉCIFIQUES

Nous ne prétendons pas disposer de données couvrant tous les aspects morphologiques et lexicaux des langues du monde. Les études typologiques quoique développées couvrent encore un petit nombre de langues. Nous proposons donc de centrer nos comparaisons sur le français et l'anglais pour les aspects où les données sont disponibles.

3.2.1. HYPOTHÈSE SUR LA STABILITÉ DES FRÉQUENCES DES CATÉGORIES GRAMMATICALES

Gougenheim *et al.* (1964 : 117) a montré, pour le français, la stabilité de la fréquence des catégories grammaticales. D'un corpus à l'autre, le pourcentage des verbes varie moins que celui des mots grammaticaux, qui à son tour varie moins que celui des adjectifs. Les noms subissent la plus grande variation. Cette hiérarchie vaut-elle pour le kirundi?

3.2.2. HYPOTHÈSE SUR LA HIÉRARCHIE DES FRÉQUENCES DES CATÉGORIES GRAMMATICALES

Il est établi, pour le français (Brunet 1981 : 298) et pour l'anglais (Johansson & Hofland (1989 : 5)[14], que les catégories grammaticales les plus fréquentes sont les mots grammaticaux, puis les substantifs; viennent ensuite les verbes et enfin les adjectifs. Il s'agira de voir si le même ordre se vérifie en kirundi.

3.2.3. HYPOTHÈSE MÉTHODOLOGIQUE

Comme nous l'avons souligné plus haut, nous ne disposons pas d'autres études en statistique lexicale du kirundi ou de langues bantoues. Nous faisons donc figure de pionnier. Dans ce cas, le cheminement méthodologique constitue en lui-même un objet de recherche. Nous voulons analyser les problèmes que pose le kirundi lors de l'élaboration de vocabulaires fondamentaux. Les problèmes généraux relevés dans l'élaboration des vocabulaires fondamentaux du français par exemple, se retrouvent-ils en kirundi ?

Nous ne disposons pas non plus d'étude quantitative sur la morphologie des langues agglutinantes. Nous avons évoqué plus haut les problèmes rencontrés dans le traitement automatique de la morphologie de ces langues à savoir : le choix du lemme et la complexité morphophonologique.

---

[14] Voir le tableau - 16.

Tout en partageant ces difficultés, le traitement automatique de la morphologie du kirundi en pose d'autres, spécifiques à la langue, et qui sont liés notamment au fait que le kirundi est une langue à classes et à tons.

Ainsi, l'analyse des problèmes que pose le traitement automatique de la morphologie du kirundi constitue un aspect important de notre recherche.

Enfin, le kirundi étant une langue agglutinante, on y observe la concaténation d'affixes autour d'un radical. Cette particularité suscite plusieurs interrogations. Quels sont, par exemple, les radicaux les plus utilisés dans le discours écrit du kirundi ? Quels sont les préfixes de classe les plus exploités de la classification nominale du kirundi ? Quels suffixes de dérivation sont les plus productifs ?

Les réponses à ces questions jetteront un premier éclairage quantitatif sur la morphologie du kirundi et aideront ultimement à faire des choix éclairés en ce qui a trait à la didactique de la langue.

4. DIFFICULTÉS ET LIMITES DE L'ÉTUDE

Une première difficulté dans notre travail tient à l'absence d'études similaires antérieures sur le kirundi ou sur une langue structurellement apparentée. Il s'agit donc d'un travail de pionnier.

À cela s'ajoute l'absence d'un dictionnaire de référence monolingue. Nous avons été de ce fait contraint de travailler avec le dictionnaire bilingue de Rodegem (1970).

Une autre difficulté est liée au manque d'outils informatiques. Nous ne disposions ni d'un lemmatiseur, ni d'un vérificateur d'orthographe, ni d'un analyseur morphologique qui auraient permis de gagner un temps appréciable.

Non moins négligeable fut le travail fastidieux sur le corpus que nous présentons au chapitre 2. Il a fallu modifier l'orthographe pour certains mots du corpus, notamment en marquant la tonalité et la quantité vocalique, pour que le travail aboutisse à des index de mots-formes interprétables.

Une étude exploratoire comme la nôtre ne peut aborder tous les aspects numériques de notre corpus tant au niveau lexical que morphologique. Nous n'y procéderons qu'à des calculs strictement nécessaires à la détermination du vocabulaire fondamental du kirundi écrit.

5. PLAN

Notre étude s'organise en cinq chapitres. Le premier présente un état de la question sur les vocabulaires de base. Nous expliquons notamment pourquoi notre étude est fondée sur un corpus écrit. Nous passons en revue les critères de sélection des vocabulaires de base, les matériaux et les unités d'analyse utilisés.

Le second chapitre décrit notre cheminement méthodologique depuis l'échantillonnage jusqu'à l'analyse des résultats. Y sont abordés les aspects relatifs à la saisie des données (désambiguïsation des homographes et modifications diverses au corpus), à la norme lexicologique (détermination du nombre de mots et de vocables), au dictionnaire de référence, à l'indexation et à l'analyse des résultats.

Le chapitre 3 est consacré à la description de la langue. Il fournit les règles morphophonologiques sur lesquelles fonctionne l'analyseur morphologique et une typologie des mots-formes et des morphèmes de la langue.

Le chapitre 4 est consacré à la présentation et à l'analyse des résultats de notre recherche. Ces derniers portent d'abord sur le plan lexical puis sur le plan morphologique.

Le chapitre 5 fournit quelques applications du vocabulaire de base en didactique du kirundi. Il aborde l'utilisation du vocabulaire de base sélectionné à l'école primaire, à l'école secondaire, en alphabétisation des adultes et en enseignement du kirundi L2.

CHAPITRE 1

ÉLABORATION DES VOCABULAIRES DE BASE : ÉTAT DE LA QUESTION

1. QU'EST-CE QU'UN VOCABULAIRE DE BASE ?

Un vocabulaire de base - appelé aussi vocabulaire fondamental - est un ensemble d'items lexicaux d'une langue sélectionnés à des fins pédagogiques. Ces items sont les plus fréquemment utilisés par les locuteurs et appartiennent à l'usage le plus courant (Dubois *et al.* 1994). Cela signifie une limitation au vocabulaire commun et la mise à l'écart des vocabulaires spécialisés.

La notion de gradation est également au cœur de la conception des vocabulaires de base. Le vocabulaire de base est conçu comme une première étape dans l'apprentissage d'une langue. Il est mis au point pour servir à un enseignement progressif.

Un vocabulaire de base comprend aussi bien des mots pleins que des mots grammaticaux. Gougenheim *et al.* (1964) pour le français et Beauchemin & Martel (1979) pour le français québécois en constituent des illustrations. Les vocabulaires de base ne doivent cependant pas être confondus avec les dictionnaires de fréquence ou les listes de fréquences. L'on peut citer parmi les premiers, Beauchemin *et al.* (1992), les sept volumes du *Dictionnaire des fréquences* du CNRS (1971) et bien d'autres.

Quant aux listes de fréquences, elles servent généralement à caractériser un état de langue, un auteur ou un corpus. L'on peut citer dans cette veine Dubois (1962 b) et Brunet (1981, 1983, 1985, 1988).

Ces dictionnaires de fréquences et ces listes ne constituent pas des vocabulaires de base; elles peuvent néanmoins être des outils pour les constituer comme le montrent les travaux de Dottrens & Massarenti (1963), Carrière (1952), Tharp *et al.* (1939) et Haygood (1936) fondés sur une compilation de dictionnaires

et / ou de listes de fréquences. Selon Vaneste (1988 : 124), les listes de fréquences (et les dictionnaires de fréquences) constituent des descriptions tandis que les vocabulaires fondamentaux sont des outils fonctionnels :

> *« un vocabulaire fondamental s'emploie dans l'enseignement des unités essentielles et indispensables dans la maîtrise de la langue. La liste de fréquences est un objet de description linguistique, le vocabulaire fondamental est un objet fonctionnel ».*

2. UN VOCABULAIRE DE BASE FONDÉ SUR L'ÉCRIT : JUSTIFICATION

L'analyse de l'évolution de la linguistique permet de constater que les grammairiens de Port-Royal (17e siècle) étudiaient prioritairement la langue écrite. Les études linguistiques du 19e siècle, qui relèvent essentiellement de la linguistique historico-comparative, sont également fondées sur l'écrit.

Depuis De Saussure (1916), la plupart des linguistes du 20e siècle, en réaction contre leurs prédécesseurs, accordent une place importante à l'oral (Goody 1994 : 250) traitant l'écrit comme un phénomène dérivé.

C'est notamment le cas de Jakobson, qui, autour des années 30, consacre dans les *Six leçons sur le son et les sens* une conception de l'écriture comme représentation de l'oral (Anis *et al.* 1988 : 43).

Les années 50 sont marquées par le behaviorisme, un courant psychologique qui accorde la primauté à l'observable, au visible. Le distributionnalisme se développera à la faveur du behaviorisme et contribuera à la description de nombreuses langues non écrites d'Amérique (Bloomfield 1933). Dans ce contexte, l'oral constitue le « discours normal » et l'écrit un simple moyen d'enregistrement de la langue (Anis *et al.* 1988 : 53).

La linguistique générative, qui supplante la linguistique distributionnelle vers les années 1965-1970, cherche notamment à rendre compte de la créativité du sujet parlant, de sa capacité à produire et à comprendre des phrases inédites. Une telle préoccupation relègue l'écrit à l'arrière-plan.

C'est dans ce contexte favorable à l'oral que de nombreux travaux ont vu le jour et ont porté sur l'analyse conversationnelle (Auchlin 1981), la variation linguistique et le contact des langues (Labov 1970, 1972). Des études psycholinguistiques ont porté sur l'analyse des erreurs et des hésitations (Richards 1974).

Certains chercheurs, comme Vachek (1989), ont cependant souligné que ces études ont tendance à négliger l'influence de l'écrit sur les processus cognitifs et insisté sur l'écriture comme voie séparée et registre distinct, affirmant ainsi que l'écrit a une spécificité et n'est pas un produit dérivé de l'oral.

L'oral et l'écrit peuvent être distingués sous au moins trois aspects : dans leur psychogénèse, leurs ressources et leur structure lexicale.

Dans leur psychogénèse (contexte de création et de réception), les deux registres sont différents : on envisage une lettre avant de l'écrire, on en modifie la formulation, on peut même la déchirer. Il s'agit d'un geste plus réfléchi, et en même temps, cela permet de s'exprimer de façon plus rigoureuse.

L'écrit se distingue également de l'oral par les ressources mises en activité. L'oral exploite la présence fréquente du référent, les gestes, les mimiques, l'intonation et les répétitions, qui constituent ses éléments de redondance. À l'écrit, cette redondance est suppléée par d'autres procédés spécifiques, car, comme le soulignent Anis *et al.* (1988 : 146) :

> *« La distance spatio-temporelle exige du scripteur une plus grande explicitation. D'où l'idée d'une plus grande redondance de la langue écrite, qui est très évidente si l'on considère (...) la présence massive de marques grammaticales (pluriel des noms et adjectifs, temps et personnes des verbes), (...) la fréquence des connecteurs. Le temps permet de corriger les erreurs, ce qui fait que les messages écrits sont plus élaborés, (...) que les oraux qui sont plus elliptiques, qui se permettent d'omettre certains éléments à cause de la présence de l'interlocuteurs.»*

Vachek (1989 : 198) fait le même constat :

« (...) Written utterances enable the language user to react (...) in a documentary and easily surveyable manner, the spoken utterances serve the purpose of reacting (...) in a manner which can be characterized as ready and immediate. Beside, spoken utterances can provide, with primary means, also for the emotive aspects of the reaction of the language user, whereas written utterances are primarily concerned with the notional content of the extralingual situations refered by them ».

Pour un même message, l'oral exige moins d'informations linguistiques. Par contre, à l'écrit (cf. Hazael-Massieux 1993 : 195), les mots de liaison remplacent l'intonation, le lexique permet de décrire des sentiments exprimés à l'oral par la mimique.

Du point de vue de leur structure lexicale, Goody (1994 : 270) inventorie les traits lexicaux spécifiques au français écrit : tendance à se servir de mots plus longs, vocabulaire plus varié (par exemple dans le choix des adjectifs), nominalisations plus fréquentes par opposition à une préférence pour les verbes à l'oral, moins de pronoms personnels, plus d'épithètes et de mots dérivés du latin.

Du point de vue morphologique, Genouvrier & Peytard (1970 : 26-27), reprenant les données de Imbs (1960), indiquent pour le français, que l'écrit exploite plus de temps verbaux que l'oral.

Dès lors, l'écrit nous semble offrir un éventail plus complet d'outils que l'oral pour l'élaboration d'un vocabulaire fondamental. Il est décontextualisé. Il ne met en jeu que le signe linguistique avec ses trois aspects fondamentaux : le signifiant, le signifié et le syntactique. L'écrit marque une distance par rapport à la réalité extralinguistique. Goody (1994) a montré que la langue écrite utilise un vocabulaire plus varié et plus abstrait que la langue orale; elle offre donc au chercheur plus de matériaux et une image différente mais plus complète du système de la langue.

Nous fondons donc notre étude sur l'écrit en souscrivant à l'hypothèse d'Anis *et al.* (1988) que la langue écrite est lexicalement plus riche que la langue orale, du fait de la nécessité de l'élaboration et de l'explicitation des messages écrits qui permettent de pallier l'absence de contexte extralinguistique. Cette position est partagée par Hazael-Massieux (1993 : 195) et Rader (1982).

Mais la presse écrite, qui sert de corpus à notre recherche, reflète-t-elle suffisamment le langage du commun des locuteurs du kirundi, autre impératif de la constitution des vocabulaires fondamentaux ?

Selon Bagnal (1993 : 1-15), la presse écrite n'est pas trop éloignée du langage commun. Elle reflète en général « *the common language* ». Selon le même auteur, le lectorat espère trouver dans les journaux écrits son propre langage[1], ce qui pousse les journaliste à utiliser « *the common language of men (...) and (...) women* ».

De plus, pour notre recherche, les thèmes retenus relèvent des préoccupations quotidiennes des Burundais et devraient fournir un vocabulaire synchroniquement valable.

Du fait de la censure et de l'auto-censure, la presse écrite africaine en général, celle en kirundi en particulier, est contrainte à un niveau d'explicitation élevé qui permet d'éviter l'ambiguïté du message. Comme le souligne Hachten (1993 : 10), la presse africaine, après avoir été un instrument au service des politiciens sous la colonisation, est devenue la victime des politiques et des répressions après les indépendances. De ce fait, elle est prudente et use d'un vocabulaire accessible au peuple, pour faire passer les messages des instances du pouvoir.

C'est à partir de textes de la presse écrite burundaise que nous menons la présente recherche. Nous passons en revue les critères qui ont généralement cours dans la sélection des vocabulaires de base et indiquons ceux que nous avons retenus.

---

[1] En un peu plus soigné, peut-être, puisqu'il s'agit d'écriture.

3. CRITÈRES DE SÉLECTION DES VOCABULAIRES DE BASE

Les études consultées relatives à l'élaboration des vocabulaires de base telles celles de Gougenheim *et al.* (1964) et Juilland *et al.* (1970) permettent de dégager des critères sur lesquels se fonde la sélection des vocabulaires de base. Ces critères sont : la fréquence, la répartition, la disponibilité, la dispersion et l'usage.

3.1. LA FRÉQUENCE

La fréquence est le nombre d'occurrences d'une unité (ici mot-forme ou vocable) dans un corpus. Selon Guiraud (1960 : 17-18), la fréquence du signe linguistique n'est pas un accident de la parole : *« [Elle] serait (...) un attribut objectif de la langue tout aussi important que sa forme ou sa signification »*. Les locuteurs retiendraient autant la forme et la signification d'un mot que sa fréquence. Certaines recherches sur la fréquence suggestive démontrent en effet une forte corrélation entre la fréquence calculée statistiquement à partir de données linguistiques et celle codifiée dans la mémoire des usagers (Fortier 1993 : 43).

D'un point de vue psycholinguistique, il existe donc un lien entre la fréquence et la reconnaissance d'un item. Selon Sproat (1992 : 113), les mots les moins fréquents prennent plus de temps à être reconnus que les mots les plus fréquents.

La fréquence reste toutefois une mesure relative et elle est en outre fonction de l'échantillon. Elle est soumise à la variabilité de plusieurs facteurs dont le genre, l'auteur, l'époque, le thème, le style, etc. Pour pallier les limites, elle est généralement corrigée par la répartition (cf. Gougenheim *et al.* 1964 par exemple).

3.2. LA RÉPARTITION

La répartition étudie l'occurrence d'un mot-forme ou vocable dans les différentes parties du corpus (Muller 1977 : 55). Pour un corpus divisé en 16 tranches comme le nôtre, la répartition d'un vocable varie de 1 à 16.

La répartition met en relief les unités à faible fréquence mais distribuées dans les différentes parties du corpus. Gougenheim *et al.* (1964) ont par exemple retenu les mots figurant dans au moins cinq textes, quelle que soit leur fréquence.

La répartition dépend du nombre de tranches établies pour un corpus; elle est donc plus tributaire des décisions du chercheur qu'un attribut de l'unité lexicale. C'est pourquoi on remplace souvent la répartition par un indice de dispersion qui exprime de façon plus adéquate la plus ou moins grande stabilité d'un item lexical dans les tranches du corpus.

3.3. LA DISPERSION

La dispersion rend compte de la plus ou moins grande stabilité d'un mot-forme ou d'un vocable dans les tranches du corpus. Quand les occurrences sont réparties de façon très régulière dans les différentes tranches, l'indice de dispersion tend vers 1; à l'opposé, il tend vers 0.

Pour calculer l'indice de dispersion, il faut que le corpus soit divisé dans des tranches égales; ce qui n'est pas notre cas. Il nous a fallu donc recourir à un calcul des fréquences théoriques des vocables en supposant leur répartition aléatoire sur le total du corpus. Nous y reviendrons au chapitre 2 § 2.5.2.

3.4. LA DISPONIBILITÉ

On appelle *vocabulaire disponible*, l'ensemble des mots de fréquence faible et peu stable mais usuels et utiles, qui sont à la disposition du locuteur. Les recherches sur le vocabulaire disponible utilisent la méthode des centres d'intéret. Gougenheim *et al.* (1964) identifient seize centres d'intérêt dont les parties du corps, les vêtements, la maison, etc.

En procédant à une enquête de disponibilité, Gougenheim et son équipe ont demandé aux enquêtés de fournir sur un centre d'intérêt déterminé les mots qui leur venaient immédiatement et naturellement à l'esprit. C'est ainsi que, dans leur liste, sont apparus des mots concrets comme *fourchette, coude, dent* qui sont rares dans les corpus.

Dans les limites de notre recherche, nous évacuons l'étude de la disponibilité des vocables. Elle suppose un inventaire de thèmes (centres d'intérêt) et des enquêtes auprès de locuteurs du kirundi pour fournir un vocabulaire thématique concret.

Il faut indiquer ici que l'échantillonnage du corpus sur lequel se fonde notre étude a été réalisé sur une base thématique (cf. chapitre 2) dans le but de dégager un vocabulaire thématique et non thématique. Il serait donc superflu d'opérer un autre choix de thèmes et de mener des enquêtes de disponibilité. L'on comprendra donc que nous ne nous servions donc pas de ce critère dans cette étude.

3.5. L'USAGE

L'*usage* est un critère utilisé notamment par Juilland *et al.* (1970), Beauchemin *et al.* (1983) et Beauchemin *et al.* (1992). Il est le résultat du croisement entre la fréquence et la dispersion. Il est obtenu par la formule suivante où U = l'usage, D = la dispersion et F = la fréquence :

$$U = F \times D$$

Ce critère est crucial dans l'élaboration des vocables de base et les applications qui en découlent. Nous y revenons au chapitre 2 § 5.2.3 et au chapitre 5.

Avant de présenter les matériaux sur lesquels se fonde cette étude et qui servent à la sélection des vocabulaires de base, nous reconnaissons avec Vaneste (1988 : 127) une limite inhérente à toutes les recherches fondées sur les données statistiques :

> *« Quelle que soit la technique de sélection (...) adoptée, il faudra toujours se garder d'identifier les mesures objectives (faites sur le corpus) à la réalité intrinsèque du langage. Le statisticien doit être suffisamment réaliste pour admettre que si ses calculs sont utiles, voire (...) indispensables à l'enseignement de la langue, il n'en demeure pas moins vrai qu'ils constituent un mode très spécifique de description d'une infime portion de la langue à un moment donné ».*

4. LES SOURCES DES VOCABULAIRES FONDAMENTAUX

Nous distinguons trois principales sources pour la constitution des vocabulaires de base : les corpus produits par des répondants au cours d'enquêtes sur la disponibilité, la compilation de listes de mots ou de recueils lexicographiques et les textes.

4.1. LES CORPUS PRODUITS PAR DES RÉPONDANTS AU COURS D'ENQUÊTES SUR LA DISPONIBILITÉ

Les enquêtes de disponibilité dont il est ici question portent sur le vocabulaire écrit ou oral des enfants, des adolescents et / ou des adultes. Citons notamment celles de Vinette (1943), Préfontaine & Préfontaine (1968), et Mackey *et al.* (1970). Vikis-Freibergs (1974) et Fortier (1993) s'en écartent partiellement dans la mesure où la première ajoute la fréquence pour sélectionner son vocabulaire alors que le second ajoute la fréquence et la répartition.

4.2. LES COMPILATIONS DE LISTES DE MOTS OU DE RECUEILS LEXICOGRAPHIQUES

Certaines listes résultent de la compilation de listes de mots déjà établies ou de recueils lexicographiques. Entrent dans cette catégories les travaux de Dottrens & Massarenti (1963), Carrière (1952), Tharp *et al.* (1939), Haygood (1936) et Henmon (1924).

Signalons ici le cas de Matoré (1963). Les 5 000 vocables du *Dictionnaire du vocabulaire essentiel* ont été sélectionnés à partir du dictionnaire Larousse par Matoré qui se référait, dit-il, à sa « conscience linguistique » et à son « expérience pédagogique ».

4.3. LES TEXTES

De nombreuses listes sont fondées sur le dépouillement de textes. Ces textes sont soit oraux, comme dans les études de Beauchemin & Martel (1979) et Gougenheim *et al.* (1964), soit écrits comme chez Juilland *et al.* (1970) et

Tashdjan (1972) ou encore mixtes (oral et écrit) comme dans l'ouvrage de Vander Beke (1929).

La taille de l'échantillon est un caractère essentiel pour la fiabilité des résultats. Selon Fortier (1993 : 43) :

> *« il existe une taille minimum en deçà de laquelle la valeur des résultats deviendrait discutable. Il semblerait qu'un échantillon de 100.000 mots confirme la validité des résultats ».*

Quant à la décision sur la taille du vocabulaire fondamental retenu, elle est fonction des critères que nous avons présentés au § 3 à savoir la fréquence, la répartition, la disponibilité, la dispersion ou l'usage. Les décisions sur les seuils minimaux pour tous ces critères dépendent des intuitions des chercheurs.

Nous présentons dans la tableau 1 la taille des corpus retenus par les auteurs cités précédemment et celle des vocabulaires sélectionnés où N représente le nombre de mots-formes du corpus, V la taille du vocabulaire, F la fréquence, R la répartition et U l'usage. Notons que Beauchemin & Martel (1979) présentent exhaustivement tous les vocables de leur corpus.

Auteur	*N*	*V*	*Critères*
Henmon (1924)	400 000	3 905	$F \geq 4$
Vander Beke (1929)	1 547 748	6 067	$F \geq 4$, $R = 5$
Gougenheim *et al.* (1964)	312 135	1 475	$F \geq 29$, $R = 5$
Juilland *et al.* (1970)	500 000	5 083	$F = 4$; $D = 5{,}83$; $U = 3$
Tashdjan (1972)	40 000	1 750	$U = 50$
Beauchemin & Martel (1979)	175 000	3 901	inventaire exhaustif

Tableau 1 - *Taille des corpus et des vocabulaires fondamentaux*

Bien qu'il ne s'agisse pas de vocabulaire de base, il faut signaler ici le *Dictionnaire des fréquences* du Trésor de la Langue Française (CNRS 1971). Celui-ci se distingue par la taille de son vocabulaire (71 415 vocables) et du corpus dépouillé (près de 71 millions d'occurrences).

Comme on peut le constater au tableau 1, la taille des corpus et des vocabulaires est très variable. L'explication des écarts au niveau de la taille du vocabulaire est à rechercher notamment dans les décisions des auteurs. Chaque auteur décide en effet, selon ses intuitions, des critères d'admissibilité des unités lexicales dans le vocabulaire de base. Prenons par exemple Henmon, Vander Beke et Tashdjan.

À l'aide d'un dictionnaire de français, Henmon a dressé une liste de 5 000 vocables qu'il estimait très fréquents (Gougenheim *et al.* 1964 : 32-33). Pour faire le dépouillement, il a confié cette liste à un certain nombre d'enseignants dans des écoles secondaires en même temps que des tranches de textes littéraires français d'une longueur de 5 000 mots. Les enseignants en ont fait le dépouillement en relevant la fréquence des 5 000 vocables pré-selectionnées dans les tranches; Henmon n'a retenu que les vocables ayant une fréquence supérieure ou égale à 4, soit 3 905 vocables.

Vander Beke a amélioré le dictionnaire d'Henmon notamment en se servant du critère de la répartition (R = 5) en plus de celui de la fréquence; ce qui a réduit considérablement la taille du vocabulaire sélectionné. L'on se serait en effet attendu, considérant la taille du corpus, à plus d'unités lexicales de fréquence supérieure ou égale à 4.

Pour sa part, Gougenheim et ses collègues, qui ont travaillé sur un corpus oral, ont sélectionné des unités lexicales dont la fréquence est supérieure ou égale à 29 et dont la répartition est égale ou supérieure à 5. Par rapport à Henmon (1924), dont la taille du corpus est proche de celle de Gougenheim *et al.* (1964), c'est essentiellement l'exigence d'une haute fréquence combinée avec une répartition de 5 qui réduit la taille du vocabulaire sélectionné.

Quant à Tashdjan (1972), il travaille sur un corpus écrit de 40 000 mots, répartis en 5 sous-ensembles thématiques, eux-mêmes subdivisés chacun en tranches de 2 000 mots, ce qui fournit 20 tranches. Le *Vocabulaire d'accès à l'information* (1 750 unités lexicales) a été sélectionné en se servant d'un indice d'usage U = F x (10R + r) où F = la fréquence, R = la répartition dans les 5

sous-ensembles thématiques (R est donc inférieur ou égal à 5) et r = la répartition dans les 20 tranches (r est donc inférieur ou égal à 20). Les unités sélectionnées avaient un indice d'usage supérieur ou égal à 50. Une unité dont F = 3, R = 2 et r = 3 obtient un indice d'usage U = 69 soit 3 (10 x 2 + 3). L'indice d'usage retenu (U = 50) est donc petit; il ne permet pas de discriminer aisément les unités lexicales, ce qui justifie la grande taille du vocabulaire sélectionné si on le met en rapport avec le corpus de départ.

Quant à nous, nous avons obtenu 4 025 vocables à partir d'un corpus de 103 561 mots-formes. Notre inventaire étant exhaustif, nous n'avons éliminé aucune unité lexicale. Nous avons cependant réparti les 4 025 vocables en trois groupes. Le premier groupe les vocables dont U ≥ 3, le second ceux dont U est compris entre 3 et 0; le troisième groupe comprend les vocables dont U est inférieur à 0. Nous revenons en détail sur ces trois groupes au chapitre 4, où nous présentons les résultats de l'étude.

Il existe des vocabulaires fondamentaux qui ne s'inscrivent pas dans les trois catégories inventoriées aux §§ 4.1, 4.2 et 4.3. Il s'agit des listes mises au point à partir d'items lexicaux appartenant à des langues différentes, de la liste du *Basic* et de *L'anglais simplifié*. Nous les présentons dans la section qui suit.

4.4. LES LISTES ÉLABORÉES À PARTIR DE LANGUES DIFFÉRENTES

Ward (1926) fonde sa liste du français fondamental sur les listes anglaises de Thorndike (1921) et Horn (1926). Elle en tire une liste de 2 000 mots qu'elle traduit en français et qu'elle compare avec celle de Henmon (1924).

4.5. LA LISTE DU BASIC

Le Basic English (***B**ritish **A**merican **S**cientific **I**nternational **C**ommercial*) est une liste mise au point par Richards & Ogden en 1929 (Richards *et al.* 1985 : 26). Ils avaient constaté que si l'on compare les définitions des dictionnaires entre elles, on observe que certains mots y reviennent fréquemment.

Ils en ont conclu que l'on pouvait définir les mots à l'aide d'un nombre très restreint de termes et qu'il était possible de concevoir un vocabulaire très limité mais dont les éléments avaient un pouvoir d'expression étendu (Coste *et al.* 1976 : 35-37).

Le *Basic* ne comprend que 850 mots. Il exploite le principe de la paraphrase, ce qui permet de ne pas multiplier le nombre d'unités lexicales.

Pour que les unités lexicales en (1) par exemple ne figurent pas dans la liste des unités retenues, le *Basic* les définit ainsi :

(1) *to ask* : to make a request, to put a question
to count : to get the number of
husband : married man
wife : married woman

Contrairement au vocabulaire fondamental, le *Basic* est un ensemble fermé, définitif, qui se suffit à lui-même, sans préoccupation de progression.

Dans la même veine se situe *l'anglais simplifié* (Humphreys 1992). Il s'agit d'un vocabulaire sélectionné aux fins de la rédaction technique pour des locuteurs non natifs de l'anglais. Il repose sur le principe qu'une famille de synonymes ou de quasi-synonymes n'est représentée que par une seule unité lexicale.

Ainsi, les verbes *to begin, to commence, to initiate*, etc. sont absents de la liste de l'*anglais simplifié* où on ne retrouve que l'unité *to start* pour représenter le sémantisme de ces verbes. *L'anglais simplifié* contient aussi bien des unités lexicales que des unités grammaticales.

Telles sont, à notre connaissance, les principales sources qui servent à l'élaboration des vocabulaires fondamentaux. Ils s'agit, rappelons-le, de corpus oraux et / ou écrits, de listes de mots ou de recueils lexicographiques, de dictionnaires appartenant à des langues différentes et même de recherches lexico-sémantiques (le *Basic* et *L'anglais simplifié*).

L'élaboration des vocabulaires fondamentaux recourt à un appareillage terminologique particulier à ce champ de recherche. Nous l'abordons dans les lignes qui suivent.

5. TERMINOLOGIE UTILISÉE

5.1. LE MOT, LE LEXÈME ET LE VOCABLE

Nous allons examiner ici trois termes très présents dans les études lexicométriques et lexicographiques. Pour bien suivre la suite de notre propos, nous synthétisons dans le tableau ci-dessous les principaux termes retenus par ces études. Trois ouvrages (Muller 1977, Mel'cuk 1993 et Mel'cuk *et al.* 1995) nous servent d'échantillons. Nous indiquons dans la dernière colonne nos choix terminologiques (Ntirampeba).

	Muller (1977 : 6-7)	*Mel'cuk (1993), Mel'cuk et al. (1995)*	*Ntirampeba*
1. Niveau du discours			
Unités 1	occurrence, mot, forme	mot-forme, phrasème	mot-forme
Unités 2	vocable	- - - - - - - - - - - - - -	vocable
Ensemble des unités 1	N : nombre de mots	nombre de mots-formes	N : nombre de mots-formes
Ensemble des unités 2	V : nombre de vocables	a. vocabulaire 1[1] : lexies qu'un individu x utilise ou qu'on trouve dans un texte x. b. vocabulaire 2 : lexies utilisées uniquement pour parler d'un domaine x	V : nombre de vocables
Nature des unités 1	unité de longueur textuelle	unité de longueur textuelle	unité de longueur textuelle
Nature des unités 2	- plusieurs acceptions - lexème employé dans un texte	- -	- plusieurs acceptions - lexème employé dans un texte
2. Niveau de la langue			
Unités 1	lexème	lexie : lexème ou phrasème	lexème
Unité 2	lemme	vocable	- - - - - - - - - - - - - - - - - -
Ensemble des unités 1	Lexique (L)	Lexique (L)	Lexique (L)
Ensemble des unités 2	- -	- -	- - - - - - - - - - - - - - - - - -
Nature des unités 1	plusieurs acceptions	une seule acception	plusieurs acceptions
Nature de l'unité 2	entrée de dictionnaire correspondant à un lexème actualisé par un vocable.	lexies à signifiants identiques partageant une composante sémantique non triviale.	entrée de dictionnaire correspondant à un lexème actualisé par un vocable.

Tableau 2 - *Terminologie de Muller (1977 : 6-7), Mel'cuk (1993), Mel'cuk* et al. *(1995) et Ntirampeba*

[1] La numérotation est nôtre.

Comme on peut le voir, le tableau 2 est construit sur la distinction traditionnelle en linguistique entre langue et discours. Il n'est pas dans notre intention d'entreprendre ici l'analyse de cette dichotomie. Il nous suffit de dire qu'elle est généralement admise par les linguistes. L'on admet généralement aussi que le discours constitue une réalisation de la langue. Nous analysons le tableau sur les deux plans. Comme notre étude s'appuie sur un corpus de textes écrits, notre analyse ira du concret à l'abstrait, de l'actualisé à la virtualité, du discours à la langue.

5.1.1. AU NIVEAU DU DISCOURS : MOT, FORME, OCCURRENCE ET MOT-FORME

Notre étude se fonde sur un corpus de textes écrits. Elle porte donc avant tout sur le discours. Ce n'est que par inférence que l'on peut se permettre de projeter les faits de discours observés sur la langue. Les textes sont composés de « mots »; le nombre total de « mots » constitue, en lexicométrie, une mesure de la longueur des textes.

Tous les linguistes s'accordent cependant pour reconnaître l'ambiguïté du « mot ». Ses synonymes sont les termes suivants : occurrence, forme et mot-forme. L'équivalent anglais est *token* (par opposition à *type*).

Il convient ici de signaler que les termes *occurrence* et *forme* sont relationnels, ce qui leur donne une incomplétude sémantique : l'on fait en effet référence à un autre élément, soit une *occurrence* de *x* ou une *forme* de *x*.

Affirmer que « mot » est ambigu, c'est dire qu'il correspond à deux sens : le mot-forme et le mot-lexème. Seul le premier relève du discours; le second relève de la langue.

Dans le sens de « mot-forme », le mot constitue une entité concrète : on dit par exemple que *suis, es, est, sommes, êtes,* etc. sont des mots différents du français. Et effectivement, par leur forme, ce sont des mots différents. Ce sens correspond à la définition orthographique du mot qui, pour des raisons pratiques, est utilisée en lexicométrie. Beauchemin *et al.* (1992 : XIX)) définissent par exemple le mot comme *« un groupe de lettres séparé d'un autre par un espace. Le nombre de mots d'un texte est représenté par N dont la valeur est une mesure (...) de l'étendue d'un texte ».*

Cette définition du mot souffre quelques exceptions portant notamment sur les groupes de mots lexicalisés appelés aussi locutions, syntagmes figés, ou, chez Mel'cuk (1993 : 362), phrasèmes (comme par exemple *casser sa pipe* 'mourir').

Muller (1977) n'utilise pas le terme « phrasème » ni ne fournit d'unité équivalente. Nous ne l'utilisons pas non plus. Tout comme Muller (1977), nous considérons les locutions comme composées de plusieurs mots-formes. Nous justifions cette décision au chapitre 2 lorsque nous présentons la norme lexicologique sur laquelle se fonde le dépouillement du corpus.

Bref, là où Muller (1977) parle de mot, de forme ou d'occurrence, Mel'cuk (1993) parle de mot-forme. Nous retenons ce dernier terme parce qu'il n'est ni ambigu ni relationnel.

5.1.2. AU NIVEAU DE LA LANGUE : LEXÈME ET LEXIE

Si le mot-forme est une unité du texte ou du discours, son correspondant en langue est le lexème, dénommé également item lexical ou unité lexicale (Muller 1985 : 451). Nous retenons le terme de lexème parce qu'il est plus court et plus cohérent à l'intérieur de la terminologie linguistique (phonème, morphème, sème, etc.). Son sens n'est cependant pas univoque.

Selon Muller (1977 : 9), *lexème* s'applique *« à toutes les unités de la langue qui peuvent avoir des occurrences dans un texte et qui, dans la tradition et la pratique lexicographique, constituent ou pourraient constituer une entrée ».*

Pour Mel'cuk *et al.* (1995), contrairement à Muller (1977), le lexème n'est pas l'unité fondamentale du lexique. C'est plutôt la lexie qui est soit un mot pris dans une seule acception (il s'agit dans ce cas de lexème) soit une locution considérée également dans une acception spécifique (il s'agit dans ce cas de phrasème). La lexie est considérée simultanément sur les plans du signifiant, du signifié et de la combinatoire (syntactique). La moindre différence sur l'un de ces plans entre deux unités autorise à poser deux lexies différentes. Ainsi Mel'cuk *et al.* (1984 : 68) inventorient une quinzaine de lexies (soit quinze sens) pour le mot-forme « cœur » tel qu'illustré ci-dessous :

COEUR, nom, masc.

I.1a. Organe principal de la circulation sanguine d'une personne... [*le cœur de Jean*]

1b. Organe principal de la circulation sanguine d'un animal [*le cœur du lion*]

2. Produit alimentaire ... [*le cœur de veau*]

3. Partie de la poitrine d'une personne ... [*Il a serré son fils sur son cœur*]

4a. Organe imaginaire des sentiments ... [*Le cœur espère toujours*]

4b. Organe imaginaire de l'intuition ... [*Son cœur le lui dit*]

5a. ... propriété de la personnalité ... [*un cœur de glace*]

5b. Personne possédant le cœur I.5a [*Vous devez la vie à un noble cœur, à un homme vaillant*]

II.1a. Partie principale d'une unité fonctionnelle... [*le cœur du bateau*]

1b. Élément principal [*le cœur du problème*]

2a. Partie centrale d'un espace... [*le cœur du royaume*]

2b. Partie centrale ... d'une plante ... [*le cœur de la salade*]

3. Objet... ayant la forme du cœur I.1a [*un cœur en papier*]

4. Une des quatre couleurs 2 des cartes à jouer... [*l'as de cœur*]

III. Organe imaginaire des nausées ... [*Cette senteur lui tournait le cœur*]

Par contre, pour le même mot-forme « cœur », Muller (1964 : 163) groupe ses 36 occurrences sous une même entrée d'index, posant ainsi un vocable CŒUR[2] actualisation du lexème « cœur » dans le corpus de *L'illusion comique* de Pierre Corneille.

Les recherches lexicométriques qui portent sur des corpus très étendus ne peuvent se permettre un raffinement dans les analyses syntactico-sémantiques comme celui adopté par le modèle sens-texte (cf. notamment Mel'čuk 1984, Mel'čuk *et al.* 1995).

Nous optons, aux fins de notre étude, pour l'unité « lexème » au sens de Muller (1977 : 9). Ainsi, au lieu de séparer des lexèmes ayant des signifiants identiques et une certaine parenté sémantique, nous les traiterons, en suivant Muller (1977 : 9) comme un même lexème.

L'ensemble des lexèmes d'une langue constitue son lexique tandis que l'ensemble des lexèmes qui ont des occurrences dans un texte constitue le vocabulaire de ce texte, dont l'unité est le vocable.

---

[2] Tout au long de ce travail, les vocables sont notés en majuscules et les occurrences en minuscules.

5.1.3. LE VOCABLE ET LE LEMME

Dans la tradition de Mulller, « *on appelle vocable un lexème actualisé en parole, dont le nombre est représenté par V, qui est la mesure de l'étendue du vocabulaire d'un texte* » (Beauchemin *et al.* 1992 : XIX). Le vocable a pour équivalent anglais *type*.

Ainsi, si les mots-formes *suis, es, est, sommes, êtes* se retrouvent dans un texte, nous disons, selon Muller (1977), que le texte a cinq occurrrences du vocable ETRE. Le vocable est donc une unité de discours.

De façon concrète, à un vocable correspond un lemme, c'est-à-dire une étiquette sous laquelle le lexème correspondant est habituellement enregistré dans les dictionnaires.

Il est important de signaler ici que tout en usant du terme « vocabulaire » Mel'cuk (1993) de même que Mel'cuk *et al.* (1995) ne distinguent pas une unité de discours appelée « vocable ». Le vocable constitue dans leur optique une unité lexicographique, plus précisément une superunité regroupant des lexies (lexèmes et phrasèmes) formellement identiques qui partagent une composante sémantique substantielle.

Mel'cuk *et al.* (1995 : 19) distinguent néanmoins pour le terme *vocabulaire* deux sens :

a. Le vocabulaire d'un individu x ou de l'ensemble de textes particuliers x : c'est l'ensemble de toutes les lexies que x utilise ou qu'on trouve dans x.

b. Le vocabulaire du domaine x : c'est l'ensemble des lexies qu'on utilise uniquement pour parler de x (ex. vocabulaire de la pêche en français). Ce vocabulaire ne comprend pas de mots-formes grammaticaux. Ce sens de « vocabulaire » a cours en terminologie.

Nous ne retenons pas les définitions fournies par Mel'cuk *et al.* (1995 : 19) pour le vocable. Elles sont fondées sur la notion de lexie dont nous avons montré qu'elle est difficilement opératoire dans le cadre de la statistique lexicale menée sur de gros corpus étant donné qu'elle fait appel à une grande finesse des analyses syntactico-sémantiques.

Le vocable est une unité importante pour notre recherche. Le vocabulaire de base du kirundi écrit que nous cherchons à constituer correspond en effet à une liste de vocables sélectionnés à l'aide de critères de fréquence et de répartition.

Nous retenons donc de la terminologie présentée au tableau 1 les termes suivants :

- le mot-forme entendu comme unité concrète d'un texte
- le vocable entendu comme réalisation d'un lexème (dans le sens de Muller 1977 : 9) dans le corpus.

Ces décisions terminologiques sont importantes. Elles permettent de comprendre les décisions pratiques prises notamment en lemmatisation.

5.1.4. LA LEMMATISATION

La lemmatisation est, dans la tradition de Muller, *« l'opération par laquelle on découpe une suite naturelle d'un texte en mots (...) pour ensuite regrouper les formes fléchies sous leur lemme »* (Beauchemin *et al.* 1992 : XIX).

Ainsi, dans un premier temps, le chercheur qui lemmatise son corpus s'occupe de la détermination du nombre de mots (N), soit une opération de découpage, et dans un second temps de celui des vocables (V), par une opération de groupement. Ces deux opérations supposent l'adoption d'une norme lexicologique que Muller (1977 : 14) définit comme suit :

> *« un ensemble de règles qui, dans la lemmatisation d'un texte, décident de la délimitation des mots et des vocables, afin de soustraire le plus possible de cas douteux à l'appréciation momentanée et subjective de l'opérateur, et de garantir ainsi au mieux la constance des traitements ».*

En lexicométrie, les avis sur la lemmatisation sont partagés. Certains optent pour la lemmatisation (Muller 1979 c, Brunet 1981, 1988), d'autres la refusent et préfèrent travailler avec les mots tels qu'ils se présentent dans le discours (Lafon 1984). L'on trouvera les arguments des lemmatiseurs dans les nombreux

travaux de Muller. Ils sont repris dans sa préface *De la lemmatisation* à l'ouvrage de Lafon (1984). Les arguments des non-lemmatiseurs sont synthétisés notamment dans Lebart & Salem (1981 : 21-24). Le débat achoppe sur trois principaux points (Lebart & Salem 1988 : 21-24, Lafon 1984 : 19) : la difficulté d'établir une norme de dépouillement transparente, les limites imposées par la lemmatisation à la taille des corpus et une perte d'informations utiles.

Pour les non-lemmatiseurs, la difficulté d'établir une norme de dépouillement qui ne laisse pas de zone d'incertitude (Lafon 1984 : IV, Lebart et Salem 1988 : 21-23) et qui, appliquée par des chercheurs différents fournirait des résultats reproductibles, commande d'admette de fonder le dépouillement sur les formes, unités faciles à appréhender.

Mais pour Muller (cf. Lafon 1984 : 4-VI), l'impossibilité d'établir une norme de dépouillement transparente constitue un phénomène normal, dont on ne doit pas s'étonner. Elle tient notamment de la nature même des unités lexicales (dont le sens ne constitue pas une donnée discrète) et de la nature des études statistiques qui exigent une classification des unités, classification qui exige à son tour de sacrifier des nuances.

Il faut signaler ici que la difficulté d'établir une norme transparente n'empêche pas de mener des études dont les résultats se révèlent assez comparables comme celle de Martel (1984) qui porte sur une comparaison entre les 54 vocables les plus fréquents du français fondamental et du québécois fondamental.

La seconde objection à la lemmatisation est qu'elle oblige le chercheur à travailler sur des corpus peu volumineux (Muller dans Lafon 1984 : XI, Lebart et Salem 1988 : 23) et requiert de nombreuses interventions humaines. À l'opposé, la non-lemmatisation permet d'automatiser le dépouillement de gros corpus; les problèmes de désambiguïsation et de lemmatisation étant relégués aux phases ultérieures de l'analyse[3].

---

[3] Entre ces deux extrêmes se retrouve une méthode intermédiaire, utilisée notamment par les chercheurs du T.L.F. pour traiter les 71 millions d'occurrences. Ne pouvant les lemmatiser une à une, ils en ont prélevé des échantillons à partir desquels ils ont tiré des estimations, estimations qu'ils ont ensuite appliquées à l'ensemble du corpus.

La troisième objection est que la lemmatisation occasionne une perte de l'information; cette perte se situe essentiellement au niveau de la grammaire du texte. Elle porte notamment sur le temps des verbes, le genre et le nombre des noms et des adjectifs (Lafon 1984 : 19).

Muller porte cette perte d'information au compte de l'indexation (Muller 1985 : 148) qui « *fait perdre une information de peu de prix pour en créer une autre certainement plus utile dans les exploitations ultérieures de l'index* ».

Les programmes informatiques disponibles permettent de contourner cette difficulté. Ils permettent de travailler, selon les finalités de la recherche sur les occurrences, les vocables ou même les deux.

À notre avis, la lemmatisation ou la non-lemmatisation trouvent chacune leurs justifications et le choix de l'une ou de l'autre repose en définitive sur les objectifs de la recherche.

Nous optons pour la lemmatisation pour deux raisons. D'un côté, la plupart des recherches sur les vocabulaires de base utilisés dans l'enseignement / apprentissage des langues maternelles ou étrangères - et dans lesquels s'inscrit notre recherche - offrent des listes lemmatisées et nous ouvrent ainsi des perspectives de comparaison. De l'autre, la structure agglutinante du mot du kirundi, que nous présentons au chapitre 3, paraît justifier, si ce n'est l'exiger, un tel choix.

5.2. LE MORPHE ET LE MORPHÈME

La statistique lexicale a besoin dans ses analyses de concepts de la morphologie. Parmi les notions de morphologie qui interviennent en statistique lexicale, citons celles de morphe et de morphème (Mel'cuk 1993 : 151).

Le morphe est le signe linguistique élémentaire. Dans le mot-forme *chantons, -ons* '1re pers. pl.' est un morphe. Le morphe se situe au niveau du discours; en langue, il correspond au morphème. Quand plusieurs morphes correspondent à un même morphème, on parle d'allomorphie.

Nous nous servons ici de ces deux concepts pour discuter du statut de mot-forme attribué à certaines séquences que d'autres recherches ont considérées comme des parties de mots-formes (morphes). Nous nous en servons également à des fins statistiques (cf. chapitre 4). Notre objectif est d'aller au-delà de la liste des

mots-formes les plus fréquents pour aboutir aux unités morphologiques simples les plus récurrentes, caractéristiques, croyons-nous, d'un vocabulaire de base de langue agglutinante.

L'analyseur morphologique du kirundi dont nous nous servons, et que nous présentons au chapitre 2, découpe le mot-forme du kirundi en morphes. À l'aide de règles morphologiques et morphophonologiques, il fournit une analyse en morphèmes.

Pour finir, situons-nous dans la perspective de tout le travail. Après avoir défini ce que l'on entend par vocabulaire de base, nous avons justifié notre choix de fonder notre recherche sur un corpus écrit. Nous avons fait observer que la langue écrite est lexicalement plus riche que l'orale du fait de la nécessité d'élaboration et d'explicitation.

Nous avons ensuite passé en revue les critères de sélection des vocabulaires fondamentaux (fréquence, répartition, disponibilité, usage) et fourni une typologie de leurs sources, qui comprennent des corpus fournis par des répondants au cours d'enquêtes sur la disponibilité, des compilations de listes ou de recueils lexicographiques, des textes et des listes élaborées pour d'autres langues. Nous avons signalé au passage les listes du *Basic* et de *L'anglais simplifié* .

Comme notre étude utilise un corpus écrit, il nous faut aborder le passage du corpus au vocabulaire fondamental du kirundi écrit. Nous présentons dans le prochain chapitre notre cheminement méthodologique depuis l'échantillonnage jusqu'à l'analyse des résultats.

CHAPITRE 2

MÉTHODOLOGIE DE L'ÉTUDE

1. LE CORPUS ET L'ÉCHANTILLONNAGE

L'étude que nous menons sur l'élaboration du vocabulaire fondamental du kirundi se fonde sur un corpus écrit. Ce genre de corpus a été utilisé à des fins lexicologiques notamment par Juilland *et al.* (1970) et Baudot (1992). Le premier a sélectionné 5 085 vocables pour son *Frequency dictionary of French words* et le second s'est consacré à l'élaboration d'un dictionnaire de fréquence des mots du français parlé au Québec.

Notre corpus est extrait de deux journaux burundais publiés en kirundi. Il s'agit du bimensuel *Ndongozi* et de l'hebdomadaire *Ubumwe*. Le premier est propriété du clergé catholique, le second du gouvernement du Burundi. Les périodes couvertes vont de 1990 à 1994 et de 1974 à 1980[1] inclusivement pour les deux journaux. Le corpus totalise environ 100 000 mots-formes.

Nous synthétisons dans le tableau ci-dessous le nombre de numéros dont nous nous sommes servi.

---

[1] Initialement, nous comptions travailler sur la période allant de 1975 à 1979 inclusivement. Mais comme nous ne disposions pas d'assez de textes pour les années 1975 et 1979, nous avons fusionné les corpus disponibles pour les années 1974/1975 et 1979/1980. Par ailleurs, à notre connaissance, aucun changement politique ou social ne permet, pour le Burundi, de distinguer l'année 1974 de l'année 1975 et 1979 de 1980.

Années	*Ndongozi*	*Ubumwe*	***Total***
1994	28	14	**42**
1993	30	36	**66**
1992	34	26	**60**
1991	29	35	**64**
1990	19	29	**48**
Sous-total 1	**140**	**140**	**280**
1979 - 80	9	27	**36**
1978	9	17	**26**
1977	15	1	**16**
1976	17	2	**19**
1974-1975	10	13	**23**
Sous-total 2	**60**	**60**	**120**
Total	**200**	**200**	**400**

Tableau 3 - *Nombre de numéros utilisés par année*

Sous réserve d'approximation, les 100 000 mots[2] sont répartis comme suit :

1990 à 1994	*Ndongozi* : 50 000 mots		*Ubumwe* : 50 000 mots	
70 000 mots	échantillon aléatoire simple	échantillon aléatoire stratifié proportionnel	échantillon aléatoire simple	échantillon aléatoire stratifié proportionnel
	5 000 mots	30 000 mots	5 000 mots	30 000 mots
1974 à 1980				
30 000 mots	2 500 mots	12 500 mots	2 500 mots	12 500 mots
Total 100 000 mots	**7 500 mots**	**42 500 mots**	**7 500 mots**	**42 500 mots**

Tableau 4 - *Échantillonnage*

---

[2] Nous nous servons de ce terme dans les tableaux parce qu'il est plus court, mais il faut comprendre qu'il s'agit toujours, en fait, de *mot-forme*.

On comprendra donc que pour chacun des journaux, la période 1990-1994 est représentée par 35 000 mots de notre échantillon dont 5 000 sont fournis par un échantillon aléatoire simple et 30 000 par un échantillon aléatoire stratifié proportionnel.

Quant à la période 1974-1980, elle est représentée pour chacun des journaux par 15 000 mots, soit 2 500 fournis par un échantillon aléatoire simple et 12 500 par un échantillon aléatoire stratifié proportionnel.

1.1. L'ÉCHANTILLONNAGE ALÉATOIRE SIMPLE

L'échantillonnage aléatoire simple se fait en choisissant des individus à l'aide d'un procédé de hasard, de manière que les individus aient la même chance de faire partie de l'échantillon. Nous procédons comme suit :

a. Prendre une table des nombres au hasard (Powell 1982 : 92). Soit l'extrait ci-dessous de la table :

(3)

25	19	64	82	84	62	74	29	92	24	01	03	91	22	48	64	94	63	15	07
23	02	41	46	04	44	31	52	43	07	44	06	03	09	34	19	83	94	62	94
55	85	66	96	28	28	30	62	58	83	65	68	62	42	45	13	08	60	46	28
68	45	19	69	59	35	14	82	56	80	22	06	52	26	39	59	78	98	76	14
69	31	46	29	85	18	88	26	95	54	01	02	14	03	05	48	00	26	43	85
37	31	61	28	98	94	61	47	03	10	67	80	84	41	26	88	84	59	69	14
66	42	19	24	94	13	13	38	69	96	76	69	76	24	13	43	83	10	13	24
33	65	78	12	35	91	59	11	38	44	23	31	48	75	74	05	30	08	46	32
76	32	06	19	35	22	95	30	19	29	57	74	43	20	90	20	25	36	70	69
43	33	42	02	59	20	39	84	95	61	58	22	04	02	99	99	78	78	83	82

b. Lire de gauche à droite dès la première ligne, trois chiffres à trois à trois (car la population pour la période est de 116 numéros). Nous retenons le premier numéro de journal compris entre 1 et 116 inclusivement, soit le numéro 103.

c. Tirer ensuite la page à saisir pour le numéro précédemment tiré. Le journal n'ayant que 16 pages, nous lisons les chiffres de la table 2 à 2 à partir de la première ligne de gauche à droite. La page retenue serait la page 03.

(4)

25	19	64	82	84	62	74	29	92	24	61	03	91	22	48	64	94	63
23	02	41	46	04	44	31	52	43	07	44	06	03	09	34	19	83	94
55	85	66	96	28	28	30	62	58	83	65	68	62	42	45	13	08	60
68	45	19	69	59	35	14	82	56	80	22	06	52	26	39	59	78	98
69	31	46	29	85	18	88	26	95	54	01	02	14	03	05	48	00	26

d. Tirer enfin la colonne à saisir. Comme le journal est organisé en quatre colonnes, nous prenons les chiffres de la table 1 à 1 à partir de la première ligne. La première colonne retenue serait la colonne 2. Si la colonne retenue n'arrive pas à fournir les 250 mots-formes (une page de format 8 sur 11 en police Times 12), on tire encore aléatoirement une colonne qui complète les 250 mots-formes. La colonne complémentaire serait la colonne 1.

L' échantillon aléatoire simple a été tiré avec remise, c'est-à-dire que les numéros tirés une première fois pouvaient l'être une seconde fois. La probabilité de tirer deux fois la même colonne était très faible[3].

1.2. L'ÉCHANTILLONNAGE ALÉATOIRE STRATIFIÉ PROPORTIONNEL

Le préalable à un échantillonnage stratifié proportionnel est l'identification de strates à l'intérieur desquelles on prélève des échantillons aléatoires simples (Trudel & Antonius 1991 : 233 - 234).

Les strates délimitées pour notre échantillon correspondent aux thèmes les plus récurrents dans la population de journaux disponibles. La structure thématique de chacun des deux journaux a été dégagée par l'analyse d'un échantillon aléatoire tiré des numéros de journaux couvrant la période 1990-1994, numéros qui représentent 70% de la population (cf. tableau 3 où la période 1990-1994 est couverte par 280 numéros sur 400 retenus).

---

[3] Pour le bimensuel *Ndongozi*, période de 1974-1980 par exemple, la probabilité de tirer deux fois la même colonne est de 1 / 480 (huit colonnes pour soixante numéros).

Dans l'ensemble, quatre thèmes reviennent dans presque tous les numéros des deux journaux : la politique intérieure, les problèmes de société, les informations religieuses pour *Ndongozi* qui, rappelons-le, appartient au clergé catholique, et l'humour qui est une particularité de *Ubumwe*.

Le format des deux journaux varie avec les années. Nous fournissons dans le tableau 5 les changements relatifs au nombre de pages et de colonnes qu'ont subis, au fil du temps, les deux journaux.

Années	*Ndongozi*		*Ubumwe*	
	Nombre de pages	Nombre de colonnes	Nombre de pages	Nombre de colonnes
1994	16	4	12 pages	4
1993	16	4	8 pages	4
1992	16	4	8 pages	4
1991	8	4	8 pages	4
1990	8	4	8 pages	4
1979-1980	8	4	8 pages	5
1978	8	4	8 pages	5
1977	8	4	8 pages	5
1976	8	4	8 pages	5
1974-1975	8	4	8 pages	5

Tableau 5 - *Nombre de pages et de colonnes des journaux dépouillés*

Comme on peut le constater, le nombre de pages varie de huit à seize pour *Ndongozi* et de huit à douze pour *Ubumwe*. Quant au nombre de colonnes, il est de quatre pour *Ndongozi* et varie de quatre à cinq pour *Ubumwe*.

Il devient donc difficile de déterminer une moyenne de nombre de pages qu'un journal consacre à un thème donné, ce qui permettrait de se faire une idée précise de la structure thématique des deux journaux.

Nous avons donc dû procéder à une estimation manuelle du nombre de pages couvertes par un thème dans les deux journaux à partir d'un échantillon aléatoire prélevé sur les journaux de la période 1990-1994.

Pour *Ndongozi*, nous avons pris comme référence les numéros à seize pages[4]; nous avons tenu compte du nombre de numéros retenus dans le corpus (92 sur 140 des numéros retenus ont 16 pages chacun, (cf. tableau 3 & 6). Pour *Ubumwe* nous avons opté pour les numéros à huit pages. Les thèmes se répartissent ainsi dans les pages des journaux :

Thèmes	*Ndongozi*		*Ubumwe*	
	16/8 pages	%	8 pages	%
Politique intérieure	6 / 3	38%	4	50%
Problèmes de société	3 / 1,5	19%	2	25%
Information religieuse	5 / 2,5	31%	0	0%
Humour[5]	0 / 0	0%	2	25%
Thèmes divers	2 / 1	12%	0	0%

Tableau 6 - *Structure thématique de l'échantillon*

On peut constater à la lecture du tableau 6 que le journal *Ubumwe* est dominé essentiellement par les thèmes de politique intérieure (50%) et l'humour (25%). Les problèmes de société (hygiène, fraude, sécurité, justice, sida, exiguïté des terres, etc.) couvrent les 25% des pages restantes avec, de temps à autre, quelques articles sur des thèmes divers comme l'environnement, la recherche scientifique, le sport mais aussi quelques commmuniqués de presse et des publicités.

---

4 Les données sont réduites de moitié pour les numéros à huit pages (cf. données du tableau 6 pour *Ndongozi*).

5 Les numéros parus de 1974 à 1980 ne contiennent pas d'écrits à vocation humoristique; à la place, on y trouve le thème du développement, qu'il soit national, provincial ou communal. Nous l'avons retenu comme thème important de *Ubumwe* pour la période de 1974-1980.

Dans *Ndongozi,* l'information religieuse (31%) rivalise avec l'information politique (38%). Les problèmes de société couvrent près de 20% du journal. Des thèmes divers (sport, économie, humour, etc.) couvrent les 12% restants.

Nous avons tenu compte de cette structure thématique dans la formation de l'échantillon stratifié. En restant le plus proche possible des proportions thématiques présentées dans le tableau 6, nous avons établi la structure de l'échantillon comme suit :

Sous-corpus	***Ndongozi***		***Ubumwe***		***Total***
	1974 - 1980	1990 - 1994	1974 - 1980	1990 - 1994	
Aléatoire	2 864 mots (w1)	5 268 mots (w2)	2 555 mots (w9)	4 933 mots (w10)	15 620 mots
Politique	5 019 mots (w3)	12 020 mots (w4)	4 988 mots (w11)	15 662 mots (w12)	37 689 mots
Social	2 601 mots (w5)	7 624 mots (w6)	3 801 mots (w13)	7 808 mots (w14)	21 834 mots
Religieux	5 015 mots (w7)	11 660 mots (w8)	----------	----------	16 675 mots
Humoristique / développement	----------	----------	3 893 mots (w15)	7 850 mots (w16)	11 743 mots
Total	15 499 mots	36 572 mots	15 237 mots	36 253 mots	103 561 mots

Tableau 7 - *Structure de l'échantillon*

Au total, notre corpus se compose de 16 sous-corpus qui, pour des raisons de conformité avec le logiciel d'indexation, sont notés de w1 à w16 (« work 1 » à « work 16 »).

Voici présentés sous forme de liste les 16 sous-corpus selon leur longueur, leurs thèmes, le journal dont ils sont tirés et les années qu'ils couvrent :

Sous-corpus	*N*	*Thèmes*	*Journal*	*Années*
w1	2 864	aléatoire	Ndongozi	1974 - 1980
w2	5 268	aléatoire	Ndongozi	1990 - 1994
w3	5 019	politique	Ndongozi	1974 - 1980
w4	12 020	politique	Ndongozi	1990 - 1994
w5	2 601	social	Ndongozi	1974 - 1980
w6	7 624	social	Ndongozi	1990 - 1994
w7	5 015	politique	Ndongozi	1974 - 1980
w8	11 660	religieux	Ndongozi	1990 - 1994
w9	2 555	aléatoire	Ubumwe	1974 - 1980
w10	4 933	aléatoire	Ubumwe	1990 - 1994
w11	4 988	politique	Ubumwe	1974 - 1980
w12	15 662	politique	Ubumwe	1990 - 1994
w13	3 801	social	Ubumwe	1974 - 1980
w14	7 808	social	Ubumwe	1990 - 1994
w15	3 893	humour	Ubumwe	1974 - 1980
w16	7 850	développement	Ubumwe	1990 - 1994

Tableau 8 - *Liste des sous-corpus*

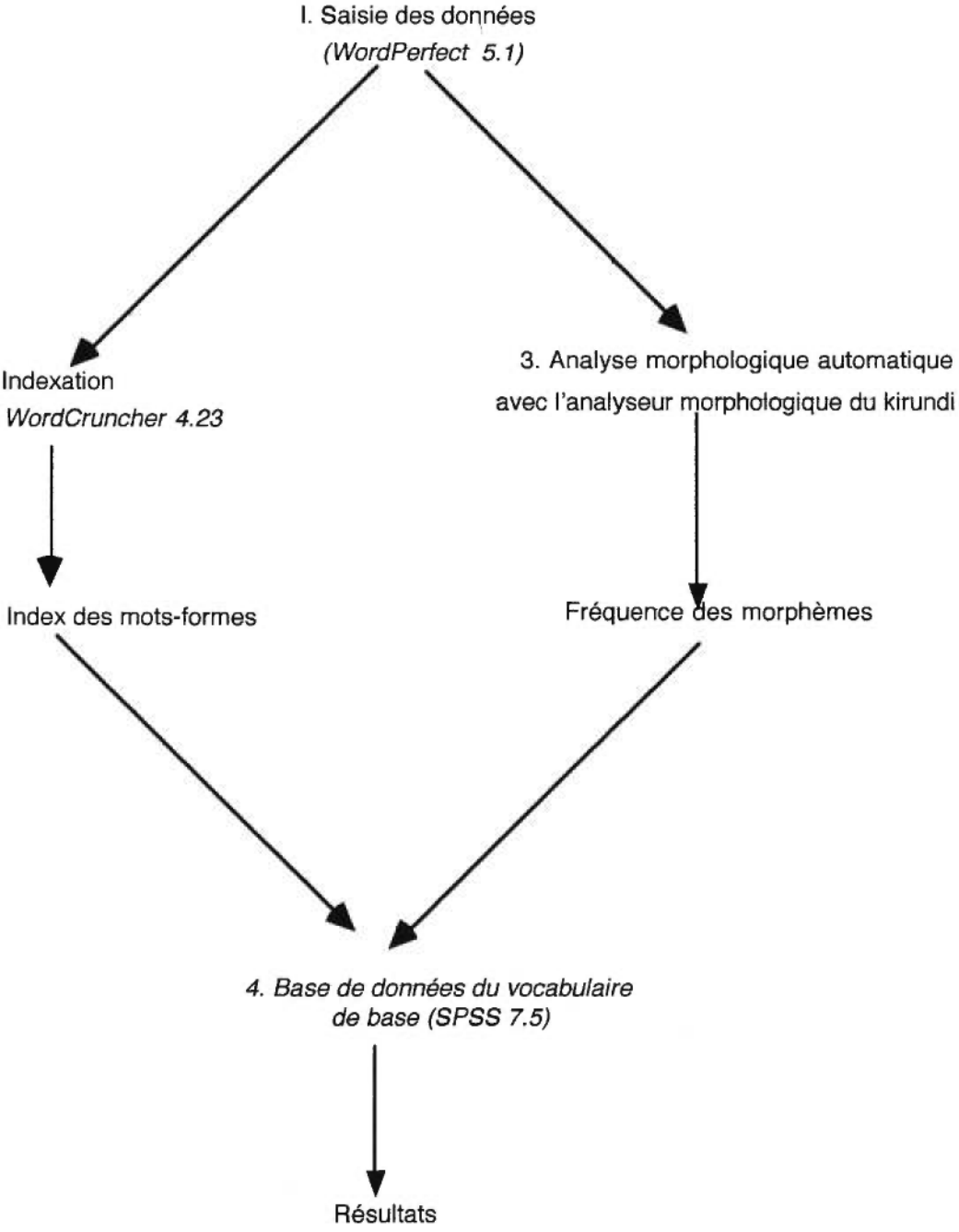
Signalons que pour les numéros du journal qui reprennent un thème sur plusieurs pages, c'est la première page qui traite du thème que nous avons retenue. L'annexe I présente thématiquement les numéros et les pages retenus dans notre échantillon. Le sous-corpus w1, par exemple, contient des extraits de dix numéros de Ndongozi (1974-1980)[6].

[6] Nous retenons une page de texte de chaque numéro, soit environ 250 mots.

2. LE TRAITEMENT DES DONNÉES

Les données sont traitées avec quatre logiciels : *WordPerfect 5.1* pour la saisie, *WordCruncher 4.23* pour l'indexation des mots-formes, l'analyseur morphologique pour l'analyse morphologique automatique et *SPSS 7.5* pour les calculs statistiques. Nous décrivons dans la suite les opérations effectuées avec chaque programme. Résumons d'abord dans un schéma notre démarche méthodologique.

Schéma 1- Le traitement des données



2.1. LA SAISIE DES DONNÉES AVEC WordPerfect 5.1

L'orthographe du kirundi est partiellement phonétique, ce qui gêne considérablement une étude fondée sur une délimitation rigoureuse du mot et des morphèmes qui le composent. Nous sommes donc intervenu sur diverses graphies. Nous présentons ici les principales modifications apportées lors de la saisie du corpus. Le but visé est de distinguer les homographes afin d'obtenir des index de mots-formes interprétables et de fournir à l'analyseur morphologique des mots-formes désambiguïsés.

2.1.1. LA DÉSAMBIGUÏSATION DES MOTS-FORMES HOMOGRAPHES

La notation de la tonalité et de la quantité vocalique permet déjà de désambiguïser certains homographes. Mais demeurent de nombreux cas d'homographies dont nous fournissons une synthèse dans le tableau ci-dessous, où SBVN = substantifs à base verbo-nominale, SBN = substantifs à base nominale et MG = mots grammaticaux.

	Verbe	SBVN	SBN	MG
Verbe	*kwandika* 'écrire' *kwandika* 'faire un tas'	*ababóna* 'il les voit' *ababóna* 'ceux qui voient'	-	-
SBVN	*abiíze* 'les intellectuels' *abiíze* 'qu'il leur apprenne'	*abatáanga* 'ceux qui arrivent les premiers' *abatáanga* 'ceux qui donnent'	-	-
SBN	*abáana* 'enfants' *abáana* 'en vivant avec *x*'	*ukwáaha* 'aisselle' *ukwáaha* 'fait de cueillir'	*umuruúndi* 'burundais' *umuruúndi* 'tibia'	-
MG	*aba* 'ceux-ci' *aba* 'il habite'	*iceése* 'publiquement' *iceése* 'récitation'	*icó* 'ce dont' *icó* 'liquide provenant du lavage de tissu'	*kó* 'que' (conj.) *kó* 'sur' (loc.)
Nom propre	*Bata* (nom d'une entreprise) *bata* 'ils jettent'	*Ubugesera* (région du Burundi) *ubugesera* 'chose terrible'	-	-

Tableau 9 - *Typologie de l'homographie en kirundi*

Tous ces cas d'homographie peuvent se ramener à trois grands types : l'homographie lexicale, syntaxique et morphologique. Nous les exemplifions dans cet ordre.

2.1.1.1. L'HOMOGRAPHIE LEXICALE

Les homonymes lexicaux appartiennent à une même partie du discours. Nous en fournissons ci-dessous quelques exemples.

2.1.1.1.1. L'homographie verbe / verbe

(5)	a. *kwandika*	'écrire'	*kwandika*	'faire un tas'
	b. *kurinda*	'supporter'	*kurinda*	'surveiller'
	c. *gutéera*	'causer'	*gutéera*	'planter'
	d. *gutwara*	'emporter'	*gutwáara*	'diriger'

2.1.1.1.2. L'homographie substantif à base verbo-nominale / substantif à base verbo-nominale

(6) a.	*ugucúra*	'faire de la métallurgie'
	ugucúra	'fait de manger plus que *x*'
b.	*ukuvyíigaana*	'fait de les raconter'
	ukuvyíigaana	'fait de les imiter'

2.1.1.1.3. L'homographie substantif à base verbo-nominale / substantif à base nominale

(7)	a. *itsíinda*	'victoire'	*itsíinda*	'poignée de céréales'
	b. *ubushaza*	'élégance'	*ubushaza*	'petit pois'
	c. *icaári*	'ce qui était'	*icaári*	'nid'

2.1.1.1.4. L'homographie substantif à base nominale / substantif à base nominale

(8)	a. *umutuúmba*	'colline'	*umutuúmba*	'tronc'
	b. *umuruúndi*	'un Burundais'	*umuruúndi*	'tibia'

2.1.1.1.5. L'homographie nom propre / substantif à base verbo-nominale

(9)	*Ubugesera*	(région du Burundi)	*ubugesera*	'chose terrible'

2.1.1.1.6. L'homographie mot grammatical / mot grammatical

(10)	*ubu*	'maintenant'	*ubu*	'ceux-ci' (dém. $_{cl.14}$)
	iyó	'si' (conj.)	*iyó*	'celui dont' (dém. $_{cl.\ 6\ \&\ 9}$)

La désambiguïsation des homonymes lexicaux se fait à l'aide de discriminants sémantiques soudés aux unités les moins fréquentes. Ces discriminants correspondent grossièrement aux gloses. Les homonymes en (5 a, 7 c, 8 b) sont, par exemple, désambiguïsés comme suit :

(11)	a. *kwandika*	versus	*kwandika*-tas
	b. *icaári*	versus	*icaári*-nid
	c. *umuruúndi*	versus	*umuruúndi* -tibia

2.1.1.2. L'HOMOGRAPHIE SYNTAXIQUE

Les homographes syntaxiques appartiennent à des catégories grammaticales différentes. Nous en fournissons ci-dessous la typologie et les exemples.

2.1.1.1.2.1. L'homographie mot grammatical / verbe

(12)	*aba*	'ceux-ci'	*aba*	'il habite'
	gusa	'seulement'	*gusa*	'ressembler à *x*'
	ica	'de' (connectif)	*ica*	'tue'
	aya	'ceux-ci'$_{cl.\ 6}$	*aya*	'gesticule'

2.1.1.2.2. L'homographie substantif à base verbo-nominale / verbe

(13)	a. *abakúze*	'ceux qui sont grands'	*abakúze*	'il les fait grandir'
	b. *agakúra*	'ce qui grandit'	*agakúra*	'il grandit'
	c. *abamúfasha*	'ceux qui l'aident'	*abamúfasha*	'il l'aide en s'occupant d'eux'

2.1.1.2.3. L'homographie substantif à base nominale / verbe

(14)	a. *amasé*	'bouse'	*amasé*	'il est collé à *x*'
	b. *amara*	'intestins'	*amara*	'il finit'
	c. *agahinda*	'chagrin'	*agahinda*	'il poussa'
	d. *ubwiira*	'enthousiasme'	*ubwiira*	'tu parles à *x*'

2.1.1.2.4. L'homographie nom propre / verbe[7]

De nombreux noms propres du kirundi correspondent à des formes verbales. Nous en donnons deux :

---

[7] Cette homographie, de même que celle en (9), est due aux limitations du logiciel d'indexation, qui ramène les majuscules aux minuscules.

(15) a. *Ngezé* (nom de personne) *ngezé* 'que j'essaie'
b. *Bubanza* (province du Burundi) *bubanza* 'ils commencent' (péjoratif)

2.1.1.2.5. L'homographie substantif à base nominale / mot grammatical

(16) a. *ivya* 'testicule' *ivya* 'de' (connectif $_{cl.\ 8}$)
b. *sé* 'son père' *sé* (interpellatif)

De manière générale la désambiguïsation des homographes syntaxiques se fait en soudant l'initiale de la catégorie morphosyntaxique (v pour verbe, n pour nom, etc.) à l'unité la moins fréquente. Ainsi, les unités en (12 a, 13 a, 14 b, 15 a, 16 a) sont désambiguïsées comme suit :

(17) *aba* versus *aba*-v
abakúze versus *abakúze*-n
amara versus *amara*-n
Ngezé versus *ngezé*-v
ivya versus *ivya*-n

2.1.1.3. L'HOMOGRAPHIE MORPHOLOGIQUE

Il y a homographie morphologique lorsque l'on a deux interprétations grammaticales différentes pour un même mot-forme. Ainsi la forme *peux* du français peut référer tant à la première personne du singulier qu'à la deuxième.

Pour le kirundi, une telle homographie concerne les substantifs dont les classes sont homographes : les classes 1 et 3 sont marquées par le même préfixe de classe [-mu-] tandis que les classes 9 et 10 ont le même préfixe de classe [-n-].

(18) a. *umuntu* 'personne humaine' $_{cl.\ 1}$
umurima 'champ' $_{cl.\ 3}$
b. *inka* 'vache' $_{cl.\ 9}$
inka 'vaches' $_{cl.\ 10}$

Au-delà de cette homographie des préfixes de classe, l'homographie morphologique en kirundi se trouve liée aux différents morphèmes qui composent les mots-formes. Elle est très fréquente comme en témoigne la longue typologie que nous fournissons ci-dessous.

2.1.1.3.1. L'homographie liée au préfixe personnel

Lorsque le préfixe personnel est identique à la première syllabe du radical, il se crée une homographie :

(19) *baaza* 'ils viennent' => [ba]préf.pers. [Ø]préd.v. [z]rad. [a]asp.
baaza 'fais de la menuiserie' => [Ø]préd.v. [baaz]rad. [a]asp.

2.1.1.3.2. L'homographie liée aux suffixes de dérivation

La présence d'un suffixe de dérivation dans certaines formes verbales crée une homographie avec des formes verbales dont le radical finit sur une séquence graphique identique à celle d'un suffixe.

2.1.1.3.2.1. L'applicatif [-ir-]

La présence d'un morphème applicatif [-ir-] dans certaines formes verbales crée une homographie avec des formes verbales dont le radical finit en [-ir-].

(20) *kurahira* 'ravir pour *x*' => [ku]préf.pers. [rah]rad. [ir]appl. [a]asp.
kurahira 'jurer' => [ku]préf.pers. [rahir]rad. [a]asp.
basangira 'ils surveillent pour'
=> [ba]préf.pers. [Ø]préd.v. [sang]rad. [ir]appl. [a]asp.

basangira 'ils partagent'
=> [ba]préf.pers. [Ø]préd.v. [sangir]rad. [a]asp.

2.1.1.3.2.2. Le causatif indirect [-ish-]

La présence d'un suffixe causatif indirect [-ish-] dans certaines formes verbales dont le radical finit en [-ish-] crée aussi une homonymie.

(21) *kuryá* 'manger' => [ku]préf.inf. [rí]rad. [a]asp.
kurîisha 'causer que *X* mange' => [ku]préf.inf. [rí]rad. [ish]caus. [a]asp.
kurîisha 'paître' => [ku]préf.inf. [rîish]rad. [a]asp.

2.1.1.3.2.3. L'associatif [-an-]

(22) *gukóra* 'travailler' => [ku]préf.pers. [kór]rad. [a]asp.
gukórana 'travailler ensemble' => [ku]préf.pers. [kór]rad. [an]ass. [a]asp.[8]
gukórana 'se réunir' => [ku]préf.pers. [kóran]rad. [a]asp.

2.1.1.3.2.4. L'oppositif [-uur-]

(23) *gutéga* 'installer un piège' => [ku]préf.pers. [tég]rad. [a]asp.
gutéguura 'enlever du piège' => [ku]préf.pers. [tég]rad. [uur]opp. [a]asp.
gutéguura 'préparer' => [ku]préf.pers. [téguur]rad. [a]asp.

2.1.1.3.3. L'homographie liée aux morphèmes compléments

Lorsqu'une forme verbale renferme un morphème complément qui ressemble formellement à un préfixe de classe, il se crée une homonymie. Nous illustrons le cas du complément [-mu-] qui ressemble formellement au préfixe de la classe 1 [-mu-].

(24) a. *umuhana* [u]préf.pers. [Ø]préd.v. [mu]compl. [han]rad. [a]asp.
'tu' 'présent' 'le' 'punir' 'inaccompli'

b. *umuhana* [u]augm. [mu]préf.cl. [hana]rad.
'enclos'

---

[8] La loi de Dhal rend compte du *g* initial en (22) et (23) (cf. chap.3 § 3.1.4.1.4).

2.1.1.3.4. L'homographie liée aux particules adverbales

Les particules adverbales du kirundi sont des morphèmes préfixés au radical verbal. Nous illustrons ceux qui sont à la base d'homographes dans le corpus.

2.1.1.3.4.1. Le morphème actualisateur [-ra-]

Lorsqu'une forme renferme le morphème actualisateur [-ra-] qui ressemble formellement à une syllabe initiale d'un radical verbal, il se crée une homonymie.

(25) *aragura* [a]$_{\text{préf.pers.}}$ [ra]$_{\text{act.}}$ [Ø]$_{\text{préd.v.}}$ [gur]$_{\text{rad.}}$ [a]$_{\text{asp.}}$
'il' 'présent' 'acheter' 'inaccompli'
'il achète'

aragura [a]$_{\text{préf.pers.}}$ [Ø]$_{\text{préd.v.}}$ [ragur]$_{\text{rad.}}$ [a]$_{\text{asp.}}$
'il' 'présent' 'pratiquer de la sorcellerie' 'inaccompli'
'il pratique de la sorcellerie'

2.1.1.3.4.2. Le morphème négateur [-tá-]

(26) *bitangáara* [bi]$_{\text{préf.pers.}}$ [tá]$_{\text{préf.nég.}}$ [Ø]$_{\text{préd.v.}}$ [angaar]$_{\text{rad.}}$ [a]$_{\text{asp.}}$
'ils' 'présent' 'errer' 'inaccompli'
'qui n'errent pas'$_{\text{cl.8}}$

bitangáara [bi]$_{\text{préf.pers.}}$ [Ø]$_{\text{préd.v.}}$ [tangaar]$_{\text{rad.}}$ [a]$_{\text{asp.}}$
'ils' 'présent' 's'étonner' 'inaccompli'
'ils s' étonnent' $_{\text{cl.8}}$

2.1.1.3.4.2. Le morphème de syndèse [-ka-]

Lorsqu'une forme renferme le morphème de syndèse [-ka-], il y a risque d'homographie avec des formes dont le radical commence par la séquence ka :

(27) *akabura* [a]préf.pers. [ka]synd. [Ø]préd.v. [búr]rad. [a]asp.
'il' 'présent' 'manquer' 'inaccompli'
'il manque alors de'

akabura [a]préf.pers. [Ø]préd.v. [kábur]rad. [a]asp.
'il' 'présent' 'exciter' 'inaccompli'
'il excite'

2.1.1.3.4.3. Le morphème de syndèse [-kí-]

Le fait que le morphème de syndèse [-kí-] soit homographe du préfixe de la classe 7 et corrolairement du morphème complément de cette même classe occasionne de nombreux homographes :

(28) *akíriha* [a]préf.pers. [kí]synd. [Ø]préd.v. [rih]rad. [a]asp.
'il' 'présent' 'payer' 'inaccompli'
'alors qu'il payait'

akíriha [a]préf.pers. [kí]compl. [Ø]préd.v. [rih]rad. [a]asp.
'il' 'le' 'présent' 'payer' 'inaccompli'
'il le paie'

2.1.1.3.5. L'homographie liée au prédicatif verbal [-oo-]

(29) *nooza* [n]préf.pers. [oo]préd.v. [z]rad. [a]asp.
'je' 'conditionnel' 'venir' 'inaccompli'
'je viendrais'

nooza [n]préf.pers. [Ø]préd.v. [ooz]rad. [a]asp.
'je' 'présent' 'laver' 'inaccompli'
'je lave'

2.1.1.3.6. L'homographie liée au morphème aspectuel accompli [-ye]

Les mots-formes verbaux à l'accompli, formés à partir d'un radical dont la consonne finale est /d/, /r/, et /g/ sont à la base de nombreux homonymes.

(30) *arunze* [a]$_{\text{préf.pers.}}$ [Ø]$_{\text{préd.v.}}$ [rund]$_{\text{rad.}}$ [ye]$_{\text{asp.}}$
'il' 'présent' 'entasser' 'accompli'
'il entasse'

arunze [a]$_{\text{préf.pers.}}$ [Ø]$_{\text{préd.v.}}$ [rung]$_{\text{rad.}}$ [ye]$_{\text{asp.}}$
'il' 'présent' 'huiler de la nourriture' 'accompli'
'il huile la nourriture'

areze [a]$_{\text{préf.pers.}}$ [Ø]$_{\text{préd.v.}}$ [rer]$_{\text{rad.}}$ [ye]$_{\text{asp.}}$
'il' 'présent' 'éduquer' 'accompli'
'il éduque'

areze [a]$_{\text{préf.pers.}}$ [Ø]$_{\text{préd.v.}}$ [rég]$_{\text{rad.}}$ [ye]$_{\text{asp.}}$
'il' 'présent' 'dénoncer' 'accompli'
'il dénonce'

La chute du yod du morphème aspectuel [-ye] est aussi à la base de l'homographie lorsqu'il s'agit de mots-formes verbaux dont le radical finit en /j/ comme en (31b) :

(31) a. *baamenye* [ba]$_{\text{préf.pers.}}$ [a]$_{\text{préd.v.}}$ [mén]$_{\text{rad.}}$ [ye]$_{\text{asp.}}$
'ils' 'passé récent' 'casser' 'accompli'
'ils ont cassé'

b. *baamenye* [ba]$_{\text{préf.pers.}}$ [a]$_{\text{préd.v.}}$ [meny]$_{\text{rad.}}$ [ye]$_{\text{asp.}}$
'ils' 'passé récent' 'savoir' 'accompli'
'ils ont su'

2.1.1.3.7. L'homographie liée au morphème réfléchi

La présence d'un morphème réfléchi [-íi-] dans certaines formes verbales crée une homographie avec des formes verbales dont le radical présente le phonème /i/ à l'initiale du radical[9].

(32) *biita* [ba]$_{\text{préf.pers.}}$ [Ø]$_{\text{préd.v.}}$ [íi]$_{\text{préf.réfl.}}$ [tá]$_{\text{rad.}}$ [a]$_{\text{asp.}}$
'ils' 'présent' 'se' 'jeter' 'inaccompli'
'ils se jettent dans *x*'

biita [ba]$_{\text{préf.pers.}}$ [Ø]$_{\text{préd.v.}}$ [íit]$_{\text{rad.}}$ [a]$_{\text{asp.}}$
'ils' présent' 'donner un nom' 'inaccompli'

Il faut signaler que ces types d'homographies peuvent se combiner, ce qui augmente le nombre d'homographes :

(33) a. *aka* 'demande'
b. *aka* 'brûle'
c. *aka* 'celui-ci' (démonstratif $_{\text{cl.12}}$)
d. *aka* 'de' (connectif $_{\text{cl.12}}$)

Entre (33 a) et (33 b), il y a homographie lexicale. Les deux sont des homographes syntaxiques de (33 c) et (33 d) alors que ces derniers sont entre eux des homographes morphologiques.

La désambiguïsation des homographes morphologiques se fait en utilisant soit des discriminants sémantiques, soit des symboles de discriminants morpho-syntaxiques soudés aux unités les moins fréquentes. Les unités en (24) et (31) sont, par exemple, désambiguïsées comme suit :

---

[9] Des règles morphophonologiques complexes expliquent la différence entre la structure phonique des mots-formes par rapport à leur structure morphologique (cf. aussi chap.3 § 3.2.1.2).

(34) *bamenye* versus *bamenye*-casser
umuhana versus *umuhana*-v

Quant aux unités en (33), elles sont désambiguïsées comme suit :

(35) a. *aka* [10]
b. *aka*-demander
c. *aka*-brûler

La désambiguïsation a été menée de front avec d'autres modifications au corpus. Nous les passons en revue au § 2.1.2.

2.1.2. LES AUTRES MODIFICATIONS APPORTÉES AU CORPUS

2.1.2.1. LA SÉPARATION DE LA PARTICULE DICTO-MODALE NTI *ET DU VERBE*

Certains auteurs comme Meeussen (1959) et Rodegem (1967) soudent la particule dicto-modale *nti* 'ne ...pas' à la forme verbale délimitée par le préfixe personnel sujet et le morphème aspectuel. Cette convention est adoptée dans les journaux de notre corpus. Nous l'illustrons par la séquence graphique en (36) :

(36) *ntibasoma* [nti]$_{dmod.}$ [ba]$_{préf.pers.}$ [Ø]$_{préd.v.}$ [som]$_{rad.}$ [a]$_{asp.}$
'négation' 'ils' 'présent' 'lire' 'non accompli'
'ils ne lisent pas'

---

[10] Nous ne désambiguïsons pas les mots-formes grammaticaux qui sont des homonymes syntaxiques. Ainsi nous ne distinguons pas *aka* (démonstratif) de *aka* (connectif). Leur comptage se fait à partir de concordances réalisées avec *WordCruncher*.

Dans Ntirampeba (1993 : 44), nous avons préconisé un modèle orthographique qui dissocie les mots et opté pour la séparation du verbe et de la particule dicto-modale. Goldsmith & Sabimana (1989 : 46) qui ont en effet démontré que la borne entre *nti* et le verbe équivaut à celle existant entre les mots (##).

Cette décision a occasionné 637 modifications au corpus, soit le nombre de formes verbales originellement soudées à la particule *nti*.

2.1.2.2. LA SÉPARATION DES MORPHÈMES LOCATIFS ET DU VERBE

Le kirundi dispose de quatre morphèmes locatifs : *yo* 'y', *ko* 'sur', *ho* 'y' et *mwo* 'dans'. Les journaux considérés, de même que Rodegem (1967) et Meeussen (1959), soudent les pronoms locatifs au verbe. L'exemple en (37) illustre le choix de cette orthographe pour le locatif *mwo* 'dans' :

(37) *basomamwo*
[ba]préf.pers. [Ø]préd.v. [som]rad [a]asp. [mwo]loc.
'ils 'présent' 'lire' 'inaccompli' 'dans'
'ils y lisent'

En accord avec Nkanira (1984 : 55), nous considérons que les morphèmes locatifs s'inscrivent en dehors des limites formelles du verbe. Nous optons donc pour les séparer de la forme verbale.

La séparation des morphèmes locatifs d'avec le verbe oblige à de nombreuses modifications du corpus. Pour le seul locatif *mwo* 'y', elle a occasionné 380 modifications. Pour les trois morphèmes locatifs restants, nous avons les chiffres suivants :

hó (loc.)	:	566 modifications
kó (loc.)	:	464 modifications
yó (loc.)	:	44 modifications.

2.1.2.3. LA SOUDURE D'UNE VARIANTE LEXICALE BI-MORPHÉMATIQUE

Lorsque la langue offre deux mots-formes grammaticaux synonymes dont l'un est monomorphématique, l'autre bi-morphématique, comme *kubéera* 'parce que' et *kubéera ko* 'parce que', nous soudons la deuxième unité en un seul mot-forme pour contourner l'homonymie. Ainsi, l'orthographe de *kubéera kó* 'parce que' devient *kubéerako*. Cette décision a occasionné 29 modifications au corpus.

2.1.2.4. LA RESTITUTION DES VOYELLES ÉLIDÉES AUX MOTS-FORMES

L'élision vocalique en finale de mot est à la base de nombreux homographes. Le mot-forme *n'* par exemple, correspond à deux unités distinctes : le prédicatif nominal *ni* 'c'est' et la conjonction de coordination *na* 'et'. En restituant aux mots-formes élidés leur forme lexicale, nous désambiguïsons le mot-forme *n'*.

2.1.2.5. LES VARIANTES ORTHOGRAPHIQUES

Dans les cas où un même mot connaît plusieurs variantes orthographiques, nous ne les avons pas modifiées lors de la saisie; l'index des mots-formes fournit la fréquence de chacune des variantes. Lors de la lemmatisation, nous avons cependant ramené leurs fréquences à la variante la plus fréquente. Ainsi le concept 'école' est rendu par trois mots-formes différents *ishuúre, ishuúri et ishuúle*. Nous avons ramené les fréquences de ces variantes à *ishuúre* qui est la variante la plus fréquente et qui respecte la structure de la langue[11]. Les fréquences des trois variantes orthographiques et de leurs formes au pluriel se présentaient comme suit :

---

[11] Le phonème /l/ qui n'est attesté que dans les emprunts, disqualifie par exemple *ishuúle* 'école' comme lemme au profit de *ishuúre* dont tous les phonèmes sont attestés en kirundi.

(38)	*ishuúre*	33
	ishuúle	22
	ishuúri	1
	amashuúre	54
	amashuúle	68
	amashuúli	2
	amashuúri	1

Lorsque les variantes orthographiques ont des fréquences proches, nous optons pour les ramener à celle qui ne pose pas de problème dans le traitement morphologique ultérieur. Par exemple, en retenant *umuúnsi* dans la paire de variantes orthographiques *umuúnsi* / *umuúsi* 'jour', nous résolvons pour l'analyseur morphologique une homonymie entre le radical [-si] 'mycoses' et [-nsi] 'jour'.

2.1.2.6. LE TYPE mweénewaányu ' ton frère' / 'ta sœur'

Les journaux dépouillés fournissent deux graphies différentes pour certains noms de parenté. Ainsi 'ton frère / 'ta soeur' est rendu tantôt par *mweénewáanyu* tantôt par *mweéne wáanyu.*

(39)	*mweénewáanyu*	~	*mweéne*	*wáanyu*
	'ton frère'/ 'ta soeur'		'celui des / celle des'	'vôtres'

La séparabilité est opératoire : on peut insérer un mot-forme comme *só* 'ton père' et former *mweéne só wáanyu* 'ton cousin germain' que nous glosons en (40) :

(40)	*mweéne*	*só*	*wáanyu*
	'celui des / celle des'	'ton père'	'vôtres'
	'ton/ta cousin(e) germain(e)'		

Nous concluons à la séparabilité et retenons deux mots pour les exemples du type de celui en (39). Nous avons donc procédé à des modifications orthographiques dans le corpus et avons scindé les termes de parenté du type *mweénewáanyu* 'ton frère' en deux unités *mweéne* 'celui de' et *wáanyu* 'vôtres'.

2.1.3. LA DÉLIMITATION DES MOTS-FORMES

La détermination du nombre de mots-formes d'un corpus se fait après la mise au point d'une norme lexicologique qui précise ce qui constitue un mot-forme et ce qui n'en est pas un. Cette norme doit être analytique (Muller 1979 c).

Afin de sauvegarder la constance des traitements et la comparaison des résultats, le praticien de la statistique lexicale adopte un dictionnaire de référence. Muller (1979 c) a choisi pour dictionnaire de référence la septième édition du *Dictionnaire général de la langue française* (Hatzfeld *et al.* 1964), dont il corrige parfois les décisions, par exemple pour *sur-le-champ,* que Muller (1979 c) compte pour un seul mot. Le nôtre est le dictionnaire de Rodegem (1970).

En français, la statistique lexicale bute notamment sur le traitement de quatre types de mots-formes qui posent des problèmes de délimitation. Il s'agit des composés, des locutions, des noms propres et des titres. Nous rapportons ce qui a été proposé pour le français et mettrons en parallèle des propositions pour le kirundi.

2.1.3.1 LES MOTS COMPOSÉS

La question qui se pose pour les composés est de déterminer s'ils sont des créations du discours (auquel cas on les scinde dans les mots-formes qui les constituent) ou des unités de la langue (auquel cas ils sont traités comme un seul mot-forme).

Muller (1977 : 16) identifie trois types de composés :

- composés avec soudure graphique; par ex. *madame, portefeuille, antigel*;
- composés avec trait d'union; par ex. *sous-marin, rendez-vous*;
- composés sans marque graphique; par ex. *chemin de fer, tiers monde.*

Cette typologie couvre bien les composés du kirundi; on en a :

- avec soudure graphique : *séerugo* 'chef de ménage' résultant de la composition de *seé* 'père' et de *urugó* 'ménage';
- avec trait d'union : *umugabo-mbwá* 'vaurien'[12] mettant en relation *umugabo* 'homme mâle' et *imbwa* 'chien';
- sans marque graphique : *intwáaro rusaángi* 'démocratie' formé à partir du substantif *intwáaro* 'gouvernement' et du qualifiant *rusáangi* 'collectif'.

Les composés soudés ne posent pas de problème de reconnaissance en tant que mots-formes. Pour les composés sans marque graphique, Muller (1979 c : 30) adopte une solution analytique. Pour des cas comme *boîte aux lettres, avion à réaction*, il compte plusieurs mots. Nous ferons de même pour le kirundi; en l'absence de trait d'union, nous adoptons la solution analytique. Les composés à trait d'union nécessitent un traitement cas par cas.

2.1.3.2. LES LOCUTIONS

Dans le traitement des locutions comme *avoir peur, à peine,* etc., Muller (1977, 1979 c) adopte la solution analytique et compte deux mots-formes. Nous ferons de même pour les locutions du kirundi comme celle en (41), où la glose de l'expression n'entretient aucun rapport sémantique avec le sens de chaque mot pris isolément :

(41)	*gufáta*	*imbwá*	*amabóko*	'être dans le pétrin'
	'prendre'	'chien'	'pattes'	

[12] Avec élision de l'augment [i-] de *imbwá* 'chien'.

Quoi qu'elle facilite le traitement des données en se fondant sur le blanc typographique comme délimiteur de mot, il faut reconnaître que la solution analytique est dommageable à la sélection des vocabulaires à des fins didactiques. Une locution comme celle en (41) correspond en effet à une unité sémantique qu'un apprenant doit comprendre et produire en bloc (Nattinger & DeCarrico 1992 : 133-134, Lewis 1977 : 11).

2.1.3.3. LES SIGLES

Conformément à la position de Muller (1977 : 17), nous comptons les sigles pour un mot : c'est que les sigles « représentent des mots dont la suite est fortement liée ».

En kirundi, on rencontre deux types de sigle différents : ceux formés en français et empruntés en kirundi comme en (42 a) et ceux formés en kirundi comme en (42 b) :

(42) a. U.PRO.NA : Union pour le Progrès National
FRO.DE.BU. : Front pour la Démocratie au Burundi
b. I.G.A.A. : *Ishírahámwe ryó Gutéeza imbere Abakényezi n'Ábáana* 'Association pour la promotion de la femme et de l'enfant'

Ces deux types de sigles ont reçu, pour des raisons de cohérence, le même traitement. Nous les comptons pour un mot-forme.

2.1.3.4. LES TITRES

Soit le titre en (43) et ses formes courtes vocatives (44) :

(43) *Nyakuubahwa Prezida wa Repuburiká y'Úburuúndi*
'Excellence Monsieur le Président de la République du Burundi'

(44) *Nyakuubahwa Prezida wa Repuburiká*
'Excellence Monsieur le Président de la République'
Prezida wa Repuburiká ' Président de la République'
Nyakuubahwa Prezida 'Excellence Monsieur le président'
Nyakuubahwa 'Excellence'

Faut-il compter (43) comme un mot-forme ou comme plusieurs ? Nous optons pour un traitement analytique et nous considérons les titres comme formés de plusieurs mots-formes. Nous nous appuyons sur le critère de séparabilité. On peut en effet insérer plusieurs mots-formes dans le titre :

(45) *Nyakuubahwa Melchior Ndadáye Prezida wa káne wa Repuburika y'Úburuúndi*
'Excellence Melchior Ndadaye quatrième président de la République du Burundi'.

2.1.3.5. LES ONOMATOPÉES

Soit le son de cloche exprimé par l'onomatopée : *nde nde nde*. S'agit-il là d'un ou de trois mots ? Nous optons pour une solution analytique et comptons trois mots; d'autant plus que l'onomatopée peut être réduite à *nde* ou amplifiée à *nde nde nde nde nde*.

Au terme de ces modifications, le corpus a été sauvegardé en code ASCII pour les fins de traitement avec le logiciel d'indexation *WordCruncher.*

2.1.4. LA DÉTERMINATION DU NOMBRE DE VOCABLES

La détermination du nombre de vocables d'un corpus constitue une étape importante dans l'élaboration d'un vocabulaire de base. Il importe en effet que l'ensemble des vocables soit constitué de façon rigoureuse et reproductible à défaut de quoi le rapport entre le nombre de vocables (V) et le nombre de mots-formes (N) ne sera pas fiable et la liste des vocables ne se prêtera pas à la comparaison avec d'autres listes.

Deux choix permettent de parer à ces écueils :

- adopter un dictionnaire de référence
- traiter un à un chacun des mots-formes du corpus et le rattacher à son lemme.

L'adoption d'un dictionnaire de référence permet de garantir la constance des traitements tandis que le traitement un à un de plus de 100 000 mots est irréalisable dans des délais raisonnables. L'on comprendra donc que nous adoptons en gros, la norme de Rodegem (1970). Nous en fournissons ci-dessous une description.

L'adoption d'un dictionnaire de référence est incontournable en lexicométrie. Dans ses recherches, Muller a pour dictionnaire de référence le *Dictionnaire général de la langue française.* Les décisions qu'il adopte s'écartent peu du classement retenu pour ce dictionnaire. Il retient par exemple l'option polysémique du *Dictionnaire général,* pour des unités comme *cour, siège, jalousie,* etc. (Muller 1979 c : 32-33).

Il s'en écarte cependant à quelques occasions par exemple pour *air* et *chapitre,* en distinguant deux vocables :

(46)	AIR I	'mine, apparence'
	AIR II	'atmosphère'
	CHAPITRE I	'partie d'un ouvrage'
	CHAPITRE II	'assemblée'

Nous fournissons dans la section suivante une description de Rodegem (1970) principalement axée sur les aspects suivants : les vocables recencés / exclus, le traitement de la polysémie / homonymie, le traitement des dérivés et des vocables grammaticaux. Nous indiquons dans la foulée certaines de nos décisions qui s'écartent de Rodegem (1970).

2.1.5. LE DICTIONNAIRE DE RÉFÉRENCE : Rodegem (1970)

2.1.5.1. LES VOCABLES RECENSÉS ET EXCLUS

Rodegem (1970) recense des vocables usuels (dont certains néologismes utilisés principalement dans l'enseignement primaire) et des vocables archaïques. Ces derniers visent à « mettre en évidence des aspects socio-culturels ». Rodegem (1970) retient également des noms propres (noms de clans par exemple). Il recense aussi bien des mots-pleins (verbes, substantifs et adjectifs) que des mots grammaticaux fléchis et non fléchis. Les termes argotiques ont été omis sauf s'ils sont jugés fréquents.

Le vocabulaire de base du kirundi écrit que nous cherchons à constituer résulte d'un dépouillement exhaustif du corpus. Nous n'omettons donc aucune catégorie de mot-forme ou de vocable.

2.1.5.2. LE TRAITEMENT DE LA POLYSÉMIE / HOMONYMIE

Rodegem (1970) privilégie une option polysémique. On peut l'illustrer avec le radical [-umv-] qui exprime trois sens 'toucher', 'entendre', 'goûter', et qui correspond à une seule entrée dans Rodegem (1970). Nous retenons cette option.

Signalons cependant que, vu l'option fortement polysémique de ce dictionnaire, nous aurons parfois à nous en écarter sur la base de la différence sémantique entre les vocables. Rodegem (1970) tient par exemple pour un simple cas de polysémie KUGABIRA 'faire un don à' et 'couper le cordon ombilical'.

Intuitivement, ces deux sens ne nous paraissent pas entretenir de relation sémantique. Nous nous fondons bien évidemment sur notre jugement de locuteur

natif. Nous distinguons donc deux vocables KUGABIRA I 'faire un don à' et KUGABIRA II 'couper le cordon ombilical'.

Nous nous écartons des choix de Rodegem (1970) dans deux autres cas :

- lorsqu'un sens d'un vocable a été omis dans Rodegem (1970); ainsi ne figure dans le dictionnaire que le sens en (47 b) alors que celui en (47 a) a été omis :

(47)	a.	*gutangaaza*	'diffuser'
	b.	*gutangaaza*	'étonner'

- lorsqu'un sens est inséré dans l'article du lexème homonyme; ainsi le sens (48 b) se retrouve dans l'article du lexème en (48 a) après deux barres verticales (II) qui l'identifient comme étant un autre sens :

(48)	a.	*gukáma*	'traire'
	b.	*gukáma*	's'assécher'

Nous tenons (48 a) et (48 b) pour des vocables distincts.

Le traitement de l'homographie est crucial dans l'analyse morphologique automatique. Soit le mot-forme substantif *umuntu* 'personne humaine' dont le radical est [-ntu] 'personne humaine'.

Ce radical est un homonyme de trois autres, ce qui donne quatre entrées différentes dans Rodegem (1970 : 286-287) et qui doivent correspondre à quatre entrées différentes dans le dictionnaire de l'analyseur morphologique. Nous adoptons pour ce faire les notations suivantes qui tiennent compte de la fréquence du radical et où F_O = fréquence.

(49)	[-ntu]	dans	*umuntu*	'personne humaine' (F_O = 838)
	[-ntu 1]	dans	*ikintu*	'chose' (F_O = 193)
	[-ntu 2]	dans	*ubuntu*	'humanité' (attributs de l'humain) (F_O = 29)
	[-ntu 3]	dans	*ahantu*	'endroit' (F_O = 23)

L'on remarquera que, pour des raisons d'économie, nous n'avons pas donné de discriminant au radical le plus fréquent.

Dans la présentation des résultats, nous contournons l'homographie des radicaux nominaux en listant les mots-formes et non les radicaux.

La désambiguïsation des radicaux verbo-nominaux est plus problématique du fait des phénomènes morphophonologiques opérant aux deux frontières morphologiques.

Soient les radicaux en (50) et les mots-formes verbaux en (51).

(50) a. [-rind-] 'attendre'
b. [-rind-] 'surveiller'

(51) *ndindé* 'que j'attende'
ndinze 'je surveille'

Comment désambiguiser ces radicaux de manière à ne pas bloquer la reconnaissance morphologique automatique des mots-formes ? On a deux choix :

- fournir un discriminant sémantique à ces deux mots-formes et les traiter manuellement lors de l'analyse morphologique;
- insérer un discriminant à l'intérieur du radical de manière à ce que la reconnaissance morphologique se fasse normalement. On aurait ainsi les radicaux : [-ri1nd-] 'attendre' et [-ri2nd-] 'surveiller'.

Ce dernier choix suppose que, aux fins de la reconnaissance des deux radicaux, les mots-formes doivent porter les mêmes discriminants. La tâche est lourde.

Nous avons opté pour la première alternative et nous nous sommes servi des concordances réalisées par *WordCruncher* pour compter les occurrences des radicaux verbaux homographes. Ainsi, l'analyseur fournit des radicaux verbo-nominaux sémantiquement ambigus et grâce à une concordance, nous départageons les fréquences des différents sens du radical.

Voici un inventaire des radicaux homographes rencontrés dans le corpus :

[-gir-] 'faire'	~	[-gir-] 'posséder'
[-shik-] 'arriver'	~	[-shik-] 'tirer'
[-am-] 'faire toujours *x* '	~	[-am-] 'porter des fruits'
[-táang-] 'donner'	~	[-táang-] 'devancer'
[-shír-] 'mettre'	~	[-shír-] 'finir'
[-mer-] 'être comme *x* '	~	[-mer-] 'germer'
[-túruk-] 'venir de'	~	[-túruk-] 'aller paître'
[-éer-] 'devenir blanc'	~	[-éer-] 'mûrir'
[-húur-] 'rencontrer'	~	[-húur-] 'frapper'
[-téer-] 'causer *x*'	~	a. [-téer-] 'planter'
		b. [-téer-] 'lancer'
[-hór-] 'faire continuellement *x*'	~	a. [-hór-] 'taire'
		b. [-hór-] 'devenir froid'

2.1.5.3. LE TRAITEMENT DES DÉRIVÉS

2.1.5.3.1. Les dérivés verbaux

Rodegem (1970) fournit une entrée aux dérivés verbaux prévisibles. Soit les dérivés verbaux en (52) et leur découpage morphologique. Les trois unités figurent comme entrées lexicales dans Rodegem (1970).

(52) *kurimira* 'cultiver pour *x*' [ku]$_{préf.inf.}$ [rim]$_{rad.}$ [ir]$_{suff.\ appl.}$ [a]$_{asp.}$
kuraabira 'regarder pour *x*' [ku]$_{préf.inf.}$ [raab]$_{rad.}$ [ir]$_{suff.\ appl.}$ [a]$_{asp.}$
kuririra 'pleurer pour *x*' [ku]$_{préf.inf.}$ [rir]$_{rad.}$ [ir]$_{suff.\ appl.}$ [a]$_{asp.}$

Comme on peut le constater, ces unités mettent toutes en jeu le suffixe applicatif [-ir-] et leur sens est prévisible.

Si on omet le suffixe applicatif [-ir-], on obtient les unités en (53), qui figurent également comme entrées lexicales dans Rodegem (1970) :

(53)	*kurima*	'cultiver'	[ku]$_{préf.inf.}$ [rim]$_{rad.}$ [a]$_{asp.}$
	kuraaba	'regarder'	[ku]$_{préf.inf.}$ [raab]$_{rad.}$ [a]$_{asp.}$
	kurira	'pleurer'	[ku]$_{préf.inf.}$ [rir]$_{rad.}$ [a]$_{asp.}$

Aux fins des traitements automatiques de la morphologie du kirundi que nous visons, nous avons retenu, pour représenter les unités en (52) et en (53), les radicaux en (54) :

(54)	a. [-rim-]	'cultiv-'
	b. [-raab-]	'regard-'
	c. [-rir-]	'pleur-'

Toutes les occurrences des mots-formes verbaux fléchis formés sur un radical ont donc été ramenées à ce radical.

Ainsi donc, dans la détermination du vocable verbal, nous ignorons certains morphèmes qui appartiennent au mot-forme verbal. L'on se fera une idée plus exhaustive des morphèmes omis au chapitre 3 § 3.2.1.2 où nous fournissons une description du verbe en kirundi. Nous indiquons au chapitre 3 § 3.2.1.1 comment nous délimitons le radical verbal.

2.1.5.3.2. Les substantifs à base verbo-nominale prévisibles

Rodegem (1970) ne retient pas les substantifs à base verbo-nominale qui sont prévisibles[13], soit les substantifs affirmatifs ou négatifs comme (55 a), les infinitifs substantivés réfléchis ou non (55 b) et certains substantifs à base verbo-nominale de la classe 5 (55 c). Nous les illustrons à partir du verbe *kurima* 'cultiver'.

---

[13] C'est-à-dire qu'à partir de n'importe quel radical de la langue, le locuteur peut former des substantifs déverbaux en appliquant tout simplement les mêmes règles.

(55)	a.	*abaríma*	'les cultivants'
		abatárimá	'les non cultivants'
	b.	*ukurima*	'le fait de cultiver'
		ukwiírima	'fait de se couper avec la houe'
	c.	*irima*	'temps de cultiver'

Nous divergeons avec Rodegem (1970) pour le traitement des substantifs dérivés en (55). Ces substantifs figurent dans notre corpus et nous devons leur accorder une entrée dans le vocabulaire. Nous avons adopté comme lemme le substantif dans sa forme du singulier (cf. Ntirampeba 1993 : 91).

Rodegem (1970) ne retient pas également tous les dérivés préfixaux résultant de la permutation des préfixes de classe. Dans Ntirampeba (1993 : 94), nous proposions aussi que le lemme des mots-formes substantifs soit le substantif au singulier, rejoignant ainsi Mel'cuk et Bakiza (1987 : 337) qui illustrent, par un processus de conversion de classe, la série des substantifs auxquels *inká* 'vache' donne cours; cf. (56) :

(56)	a. *agaka*	'petite vache'	diminutif
	b. *igika*	'grosse vache'	augmentatif simple
	c. *uruka*	'grosse vache laide'	augmentatif péjoratif
	d. *amaka*	'nombreuses vaches laides'	collectif péjoratif simple
	e. *ubuka*	'petites vaches laides'	collectif péjoratif diminutif

Vu la récurrence du sens 'vache' dans tous ces mots-formes, nous posons comme vocable le substantif à la classe neutre et au singulier, c'est-à-dire :

> *« la classe initiale, inhérente à ce radical : c'est la classe dans laquelle [le radical] est muni de la signification flexionnelle de singulier et n'exprime que sa signification lexicale pure. »* (Mel'cuk & Bakiza 1987 : 293).

Le vocable est pour les exemples en (56) *inka* 'vache'. Cette position rejoint celle de Rodegem (1970) qui n'accorde pas d'entrées aux substantifs dérivés formés par permutation des préfixes de classe.

2.1.5.3.3. Le traitement des adjectifs

En kirundi, l'adjectif s'accorde avec le substantif et prend son préfixe de classe. De ce fait, l'adjectif n'a pas de forme neutre. À quel vocable le rattacher alors? Rodegem (1970) enregistre les adjectifs selon leur radical. Nous retenons cette option.
Soient les exemples en (57) - rappelons que «augm.» doit être lu «(voyelle) augment» :

(57) a.

inká	*ntó*
[i]augm. [n]préf.cl.[ká]rad.	[n]préf.adj. [tó]rad.
'vache'	'petit'
'une petite vache'	

b.

agaká	*gató*[14]
[a]augm.[ka]préf.cl. [ká]rad.	[ka]préf.adj. [tó]rad.
'vache'	'très petit'
'une très petite vache'	

Nous retenons ainsi comme vocable le sens stable de l'adjectif *-to* 'petit'. Nous rattachons donc toutes les occurrences d'un adjectif à son radical. Les mots-formes *ntó, gató* 'petit' seront rattachés au vocable radical *-to* 'petit'.

Signalons l'existence de mots-formes adjectivaux issus de la reduplication des mots-formes adjectivaux simples.

(58) *inka ntóontó* 'les jeunes vaches'
'vache' 'jeune'$_{cl.9}$

Nous ramenons les occurrences de ce mots-forme au radical adjectival redoublé [-tó...tó] 'jeune'.

---

[14] Pour [k] ---> [g] cf. la loi de Dhal, chapitre 2 § 2.2.3.2.4.

2.1.5.3.4. Le traitement des mots grammaticaux

Rodegem (1970) écarte certains mots grammaticaux. Parmi les mots grammaticaux fléchis, il ne recense que les numéraux, les locatifs et les interrogatifs. Les seuls mots grammaticaux non fléchis recensés sont les conjonctions.

La nature de notre étude nous contraint à un inventaire exhaustif de tous les mots-formes grammaticaux du corpus. Nous n'en omettons donc aucun. Nous nous inspirons des travaux sur le français pour leur regroupement en vocables.

De façon générale, pour les mots grammaticaux du français, Muller (1979 c) et Lyne (1985 : 58) optent pour le regroupement et comptabilisent comme une même unité des unités différentes comme [*le*]art. et [*le*]pron., [*que*]conj. et [*que*]pron., [*si*]conj. et [*si*]adv., [*en*]prép. et [*en*]adv., etc.

Une telle décision s'appuie sur un argument psycholinguistique : les mots grammaticaux sont utilisés par le locuteur de façon moins consciente que les mots à signification lexicale; ils ne résultent donc pas du choix volontaire du locuteur.

Dans une étude prospective comme la nôtre, il eut été facile d'opter pour une lemmatisation des mots grammaticaux sur une base formelle, d'autant plus qu'elle permettait d'éviter la longue opération de désambiguïsation. Nous n'avons pas retenu cette option parce que l'analyse de l'homographie des mots grammaticaux permettra ultimement de faire des choix plus éclairés dans l'enseignement des structures grammaticales du kirundi. Grâce aux concordances fournies par *WordCruncher*, nous avons pu établir la fréquence de chacun des mots-formes grammaticaux homographes.

Pour le kirundi, les mots-formes grammaticaux se répartissent selon leur forme en trois groupes distincts : ceux dont les formes sont homographes, ceux dont les formes sont hétérographes et ceux à forme courte et à forme longue.

Les mots-formes grammaticaux hétérographes ne posent aucun problème de traitement. Sont concernés d'une part, les possessifs et les connectifs d'autre part. Rappelons que dans les deux cas, la différence ne tient qu'à la présence d'une voyelle identique à celle du préfixe de classe du substantif déterminé ou pronominalisé.

Restent les mots-formes à forme courte / longue et les mots-formes grammaticaux homographes.

2.1.5.3.4.1. Les vocables des mots-formes grammaticaux courts versus *longs*

Les pronoms allocutifs et substitutifs ont une forme brève et une forme longue. Il en est de même pour les prépositions. Faut-il rattacher ces deux types de mots-formes à un même vocable ?

Nous faisons observer plus loin (chapitre 3 § 3.3.13.) que les formes longues sont emphatiques par rapport aux formes courtes. Cette différence, qui est de nature pragmatico-sémantique, nous a conduit à rejeter l'hypothèse d'allomorphie. Elle nous permet aussi de fonder notre décision de regrouper les formes courtes sous un vocable et les formes longues sous un autre.

Par ailleurs, notre étude étant prospective en statistique lexicale du kirundi, il nous semble plus adéquat de poser deux types de vocables afin de fournir des données aux études ultérieures; d'autant plus qu'il n'est pas difficile de procéder à une fusion des fréquences des deux types de vocables même après la lemmatisation.

2.1.5.3.4.2. Les mots-formes grammaticaux homographes

On peut distinguer deux cas : soit les homographes appartiennent à la même catégorie morphosyntaxique, soit ils appartiennent à deux catégories morphosyntaxiques différentes.

Le premier cas est illustré par les démonstratifs, les numéraux et les allocutifs. Leurs formes sont identiques en emploi pronominal et déterminatif. Il s'agit là d'une simple homonymie fonctionnelle. Comme nous ne quantifions pas les fonctions des mots-formes, nous avons opté, dans ce cas, pour le regroupement sur une base formelle comme l'ont fait Muller (1979 c) et Lyne (1985) pour le français.

Le second cas est celui de mots-formes grammaticaux homographes appartenant à des catégories morphosyntaxiques différentes. Nous distinguons les cas où l'homographie opère entre deux catégories morphosyntaxiques et celui où elle opère entre trois.

L'homographie entre deux catégories morphosyntaxiques présente plusieurs cas. Nous les passons en revue.

- Homographie entre les connectifs et les substitutifs

C'est l'homographie grammaticale la plus fréquente. Nous avons par exemple les résultats suivants où les chiffres représentent la fréquence :

có (substitutif)	: 229	et	*có* (connectif)	: 32
vyó (substitutif)	: 155	et	*vyó* (connectif)	: 50
wó (substitutif)	: 21	et	*wó* (connectif)	: 78
ryó (substitutif)	: 37	et	*ryó* (connectif)	: 61
zó (substitutif)	: 16	et	*zó* (connectif)	: 53
rwó (substitutif)	: 11	et	*rwó* (connectif)	: 18
twó (substitutif)	: 1	et	*twó* (connectif)	: 3

- Homographie entre les démonstratifs et les substitutifs :

urwo (démonstratif)	: 53	et	*urwo* (substitutif)	: 4
izo (démonstratif)	: 55	et	*izo* (substitutif)	: 6
utwo (démonstratif)	: 2	et	*utwo* (substitutif)	: 11

- Homographie entre les connectifs et les démonstratifs :

aka (connectif)	: 13	et	*aka* (démonstratif)	: 6

- Homographie entre un adverbe et un prédicatif nominal :

si (adverbe)	: 30	et	*si* (prédicatif nominal)	: 106

- Homographie entre un numéral et un adverbe :

kamwé (numéral) : 5 et *kamwé* (adverbe) : 1

Quant à l'homographie entre trois catégories morphosyntaxique, l'on a aussi plusieurs cas :

- Homographie entre un démonstratif, un interrogatif et un substitutif :

ubwo (démonstratif) : 58
ubwo (interrogatif) : 14
ubwo (substitutif) : 6

- Homographie entre un substitutif, un connectif et un locatif :

yó (substitutif) : 166
yó (connectif) : 156
yó (locatif) : 44

- Homographie entre une conjonction, un locatif et un substitutif :

kó (conjonction) : 1 212
kó (locatif) : 464
kó (substitutif) : 92

Cette typologie de l'homographie des mots-formes grammaticaux et leurs fréquences permettent déjà de se faire une idée de l'ampleur du travail qu'aurait exigé la désambiguïsation des mots-formes grammaticaux lors de la saisie. Nous l'avons évité et avons préféré compter les occurrences de ces homographes à partir de concordances fournies par *WordCruncher*.

2.1.5.3.5. Les vocables pour les interjections, les conjonctions et les adverbes

Les interjections, les conjonctions et les adverbes ne posent aucun problème de regroupement de mots-formes sous un vocable dans la mesure où, en kirundi, l'homographie est inexistante dans ces catégories grammaticales. À chaque mot-forme de ces trois catégories correspond un vocable.

2.2. L'INDEXATION AVEC WORDCRUNCHER 4.23

WordCruncher est un logiciel qui opère en deux temps : il indexe d'abord les fichiers de textes DOS et crée un index de fréquences de mots-formes du texte que l'usager peut alors manipuler.

On appelle index le résultat d'un travail de dépouillement lexical. En lexicométrie, on distingue l'index de mots-formes, qui indique dans l'ordre alphabétique les mots-formes dépouillés, de l'index de vocables, qui distingue les unités du vocabulaire. *WordCruncher* fournit un index de mots-formes.

Pour les langues qui s'y prêtent, *WordCruncher* permet également de mener des recherches sur des séquences de caractères (un radical par exemple) ou des cooccurrents (Merrilees *et al.* 1992). Cette caractéristique du logiciel nous a permis de procéder à de nombreuses vérifications sur les radicaux.

Dans un deuxième temps, *WordCruncher* permet de visualiser l'index produit, d'effectuer des retours au texte et de mettre au point, pour n'importe quel mot de l'index, une concordance.

Une concordance est un index de mots-formes présentés dans leurs contextes d'occurrence. Elle permet d'étudier les différents emplois d'un vocable, ses différentes acceptions et sa combinatoire (cf. Mel'cuk 1993 : 123-129).

Notre recherche sur le vocabulaire de base du kirundi écrit exploite davantage les index que la concordance, par ailleurs essentielle pour les analyses syntaxiques.

Un index généré par *WordCruncher* comporte deux parties : des données générales sur le texte indexé et un index de mots-formes.

Les données générales sont le nombre de mots-formes différents du texte, le nombre total de mots-formes, le nombre total de caractères dans les mots-formes et le nombre de pages du corpus. *WordCruncher* étant un programme qui utilise l'anglais, les données générales se présentent comme en (59) où le sous-corpus w1 sert de référence :

(59) Unique Words Read = 1324
Total Words Read = 2 864
Total Chars Read = 16 829
Pages Found = 10

L'index de mots-formes généré par *WordCruncher* peut être ordonné selon la décision de l'usager. Celui de notre corpus est un index alphabétique croissant (de A à Z). Chaque mot-forme est suivi de sa fréquence. Ainsi l'index de w1 (sous-corpus 1) contient, par exemple les données suivantes que nous glosons dans la colonne de droite :

(60)	1 *aba*	'ceux-ci'
	2 *abakényezi*	'les femmes'
	5 *abaruúndi*	'les Burundais'
	11 *abantu*	'les personnes humaines'
	21 *ku*	'sur'
	4 *taanzaaniyá*	'la Tanzanie'
	7 *umuntu*	'la personne humaine'
	1 *yabídushiriye*	'il les a placés pour nous'

L'index de notre corpus comprend aussi bien des mots pleins, des mots grammaticaux que des noms propres.

À partir de l'index alphabétique, on réalise un index hiérarchique (les mots-formes y sont rangés par ordre de fréquence décroissante) qui permet de mettre en évidence la fréquence des mots-formes, donnée importante dans l'élaboration des vocabulaires fondamentaux.

Comme on le voit en (60), les mots-formes *abantu* 'personnes humaines' et *umuntu* 'personne humaine' ne diffèrent que par la flexion. Leurs fréquences devraient être cumulées et attribuées à la forme la moins marquée des deux, la forme au singulier. Ce travail est généralement effectué par un programme appelé lemmatiseur, outil qui n'existe pas pour le kirundi.

Il nous a fallu dès lors mettre au point un outil suppléant, un analyseur morphologique qui permette le regroupement des formes fléchies et de leurs fréquences, formes rattachées au radical *-ntu* 'personne humaine'.

2.3. L'ANALYSEUR MORPHOLOGIQUE

L'analyseur morphologique dont nous nous servons est un programme qui, pour des mots-formes en entrée, fournit à la sortie un découpage morphologique et un décompte des différents morphèmes du corpus[15]. Cet analyseur morphologique a trois composantes : un dictionnaire, une interface et un ensemble de règles.

Le dictionnaire comprend la liste de tous les radicaux du corpus, de tous les affixes de la langue et leurs positions à l'intérieur du mot-forme. Il comprend aussi la liste des noms propres du corpus ainsi que celle des mots grammaticaux. Quant à l'interface et aux règles utilisées par l'analyseur dans la reconnaissance des morphèmes, elles exigent, compte tenu de leur complexité, une présentation plus détaillée.

---

15 Nous en profitons pour remercier notre codirecteur de recherche et responsable du Groupe de Recherche en Linguistique du Texte (GRELT), Monsieur Nathan Ménard, qui a financé la mise au point de ce logiciel. Nous remercions, bien entendu, Monsieur Philippe Daviet, programmeur et condisciple au département de linguistique et de traduction, avec qui nous avons passé de longues heures à tester la validité du programme.

2.3.1. L'INTERFACE DE L'ANALYSEUR MORPHOLOGIQUE

L'interface de l'analyseur morphologique se présente sous la forme d'une fiche morphologique de chaque mot-forme à signification lexicale contenu dans le corpus, avec sa catégorie grammaticale et son numéro d'occurrence (qui va de 1 à N selon la longueur du texte) dans le texte analysé[16]. L'interface se présente comme suit pour l'ensemble du mot-forme *bazobirimiranira* 'il les cultiveront les uns pour les autres' :

N°1

morphème	mot-forme *bazoobirimiranira*	structure morphologique : *bazoobirimiranira*[17]
préfixe de classe	ba	ba
préfixe 1	zoo	zoo
préfixe 2	-	-
préfixe 3	-	-
complément 1	bi	bi
complément 2	-	
complément 3	-	-
radical	rim	rim
suffixe 1	ir	ir
suffixe 2	an	an
suffixe 3	ir	ir
aspect	a	a

CAT1

Précédente	Prochaine	Mauvais	Statistique
Quitter	Enregistrer	Ouvrir	Corriger

Tableau 10 - *Interface de l'analyseur morphologique*

---

16 Les noms propres et les mots grammaticaux ne sont pas candidats à l'analyse morphologique; ils sont consignés dans des fichiers supplémentaires distincts.

17 On notera l'absence de redoublement de la voyelle *o* dans le mot-forme. Nous revenons sur les choix orthographiques des journaux dépouillés au § 3.

La case **CAT** affiche la catégorie grammaticale du mot-forme (1 pour les verbes, 2 pour les adjectifs, 0 pour les substantifs). La case **N°** contient un numéro d'occurrence du mot-forme dans la base de données, ce qui permet d'avoir des repères dans le fichier.

La fonction **Ouvrir** permet d'ouvrir le fichier Résultat (nommé Base de données). Les mots-formes du corpus, du plus simple au plus complexe, s'affichent en même temps que le découpage morphologique proposé par l'analyseur[18]. Si le découpage est correct, on clique sur **Prochaine** et on passe au mot-forme suivant. S'il est mal découpé, on modifie la fiche du mot et on clique sur **Corriger** afin que toutes les occurrences du mot-forme soient corrigées en une seule opération. Si, erronément, il s'agit d'un mot grammatical ou d'un nom propre, on clique sur **Mauvais**; on le récupérera dans le fichier approprié nommé « Mots grammaticaux ».

La fonction **Précédente** permet de revenir aux mots-formes déjà corrigés. Lorsque le fichier est corrigé, la fonction **Statistique** permet de faire les comptages sur les morphèmes et de créer un fichier de statistiques portant sur le corpus, statistiques que l'on peut imprimer.

Comme on le voit, l'analyseur reçoit en entrée des mots du corpus et fournit en sortie leur découpage morphologique, ce qui permet de disposer de statistiques morphologiques en cliquant sur la fonction appropriée.

L'on se réfèrera à l'annexe 2 pour se faire une idée des résultats fournis par l'analyseur morphologique. Ces résultats permettent d'apprécier la distribution des différents radicaux et affixes dans les 16 sous-corpus. L'annexe 3 fournit pour quelques radicaux adjectivaux une compilation des résultats des 16 sous-corpus.

Quant aux règles utilisées par l'analyseur morphologique, nous les présentons au chapitre 3, chapitre consacré aux principaux aspects de la morphologie du kirundi dans une perspective lexicologique.

---

[18] Nous avons réduit la structure du mot-forme à 12 positions paradigmatiques. Les mots-formes plus longs sont rares; ils sont traités manuellement.

Disons pour l'instant que ces règles servent à la reconnaissance et à la reconstruction des mots-formes. Une fois les mots-formes décomposés, l'analyseur morphologique génère quatre fichiers : le fichier des résultats (qui constitue la base de données morphologiques), le fichier des mots grammaticaux du corpus (indexé avec *WordCruncher*), un fichier de noms propres contenus dans le corpus et un fichier de sortie constitué de mots-formes résiduels, non reconnus, dont le découpage morphologique est effectué manuellement.

Le fichier « résultats » est vérifié et corrigé manuellement pour s'assurer que le découpage morphologique effectué par l'analyseur est conforme. Environ 50% des découpages n'ont pas besoin d'être retouchés. Considérant que 40% des mots-formes du corpus sont des mots grammaticaux (ils sont groupés dans un autre fichier) et que 7% sont des noms propres, des abréviations et des mots-formes divers (dont on ne fait pas l'analyse), l'analyseur reconnaît ainsi correctement près de 75% des mots-formes du corpus.

Les 16 sous-corpus sont traités un à un avec l'analyseur morphologique et les résultats sont compilés dans des tableaux qui permettent d'apprécier la répartition des différents morphèmes pour l'ensemble du corpus (voir l'exemple de l'annexe 3). C'est à partir de ces données et des index de mots-formes générés par *WordCruncher* que nous mettons au point une base de données avec SPSS 7.5, base de données à partir de laquelle nous calculons les différents indices lexicométriques des vocables du corpus.

2.4. LE LOGICIEL STATISTIQUE SPSS 7.5

SPSS 7.5 a servi à deux niveaux de notre étude, au niveau lexical et au niveau morphologique.

Au niveau lexical, ce logiciel nous a permis de constituer une base de données des vocables du corpus, base de données fondée sur les résultats de *WordCruncher* et de l'analyseur morphologique.

Soit le mot-forme *ukuguru* 'jambe'. *WordCruncher* fournit sa fréquence ($F_O = 4$). Il fournit aussi la fréquence de *amaguru* 'jambes' ($F_O = 17$). L'analyseur quant à lui fournit la fréquence du radical [-guru] 'jambe' ($F_O = 21$) et des préfixes de classe $[ku]_{cl.15}$ et $[ma]_{cl.6}$. C'est ainsi que nous établissons que dans le corpus, le morphème $[ku]_{cl.15}$ a une fréquence de 925 et tandis que $[ma]_{cl.6}$ une fréquence

de 2 144, la voyelle initiale des deux mots-formes étant prévisible (c'est celle du préfixe).

Ainsi, nous avons retenu pour les mots-formes *ukuguru* 'jambe' et *amaguru* 'jambes' le vocable *ukuguru* 'jambe' de fréquence 21 (4 +17) dont nous avons calculé un indice de dispersion et d'usage (voir formules au §§ 2.5.2 et 2.5.3).

Au niveau morphologique, SPSS 7.5 nous a permis de calculer un coefficient de variation, un indice de fréquence, de dispersion et d'usage pour chaque morphème. Nous abordons en 2.5. les différents indices calculés et la manière dont nous les avons obtenus.

2.5. LES INDICES CALCULÉS AVEC SPSS 7.5

Le logiciel SPSS 7.5 nous a permis de calculer automatiquement trois indices théoriques pour chaque vocable et chaque morphème : un indice de fréquence, un indice de dispersion et un indice d'usage. Nous illustrons notre démarche à partir de quelques vocables.

2.5.1. L'INDICE DE FRÉQUENCE (F_o)

Avant d'indiquer comment nous avons calculé ces indices, il faut rappeler que notre corpus est composé de sous-corpus de longueurs inégales. Il nous a fallu recourir à un calcul de fréquences théoriques en supposant une répartition aléatoire des vocables sur le total du corpus.

Soit par exemple le vocable [-IIZÁ] 'beau'. Sa fréquence réelle ou observée (F_o) est de 98. Elle est la somme des sous-fréquences d'un vocable dans les 16 sous-corpus.

Pour calculer la fréquence théorique de ce vocable dans chacun des 16 sous-corpus, nous appliquons la formule suivante, où N = la longueur de chaque sous-corpus et 103 561 la taille de tout le corpus.

$$F_t = F_o \times \frac{N}{103561}$$

Nous avons ensuite calculé l'écart type et les indices de dispersion et d'usage, à partir des écarts observés entre les fréquences réelles et les fréquences théoriques. La démarche a été également utilisée par Beauchemin et *al.* (1983 : XI-XII). Le calcul est illustré ci-dessous pour le vocable [-IIZÁ] 'beau'. On a les données suivantes :

Tranches	*Longueur des tranches*	*Fréquences réelles*	*Fréquences théoriques*	*Écarts*	*Carré des écarts*
w1	2 864	2	2,71	-0,71	0,50
w2	5 268	5	4,99	0,01	0,00
w3	5 019	7	4,75	2,25	5,06
w4	12 020	11	11,37	-0,37	0,14
w5	2 601	3	2,46	0,54	0,29
w6	7 624	6	7,21	-1,21	1,46
w7	5 015	5	4,75	0,25	0,06
w8	11 660	15	11,03	3,97	15,76
w9	2 555	3	2,42	0,58	0,34
w10	4 933	3	4,67	-1,67	2,79
w11	4 988	4	4,72	-0,72	0,52
w12	15 662	6	14,84	-8,84	78,15
w13	3 801	8	3,60	4,40	19,36
w14	7 808	7	7,39	-0,39	0,15
w15	3 893	5	3,68	1,32	1,74
w16	7 850	8	7,43	0,57	0,32
Total	**103 561**	**98**	**98**	**0,00**	**126,65**

C'est à partir du carré des écarts que l'on calcule la variance en tenant compte du nombre de tranches soit :

$$\sigma^2 = \frac{126,65}{16} = 7,91$$

On peut ainsi déterminer l'écart type pour le vocable [-IIZÁ] 'beau' :

$$\sigma = \sqrt{7,91} = 2,81$$

Connaissant l'écart type et la moyenne (98 / 16 = 6,10) du vocable [-IIZÁ] 'beau', on calcule son coefficient de variation, soit :

$$v = \frac{\sigma}{\overline{X}} = \frac{2,81}{6,10} = 0,46$$

2.5.2. L'INDICE DE DISPERSION (*D*)

L'indice de dispersion est exprimé par la formule suivante (Juilland *et al.* 1970, Beauchemin *et al.* 1992) :

$$D = 1 - \frac{v}{\sqrt{n-1}} = 1 - \frac{v}{\sqrt{15}}$$

La dispersion rend compte de la plus ou moins grande stabilité d'un mot-forme ou d'un vocable dans les tranches du corpus. Quand les occurrences sont réparties de façon très régulière dans les différentes tranches, l'indice de dispersion tend vers 1; à l'opposé, il tend vers 0.

Dans le cas qui nous concerne, l'indice de dispersion atteint des valeurs négatives pour des vocables dont la fréquence est très basse ou qui ne se retrouvent que dans un seul des 16 sous-corpus.

Selon Juilland & Chang-Rodriguez (1964 : LIII-LV), un indice de dispersion négatif est normal avec les basses fréquences. Il est dû aux sous-fréquences «0» qui entrent dans le calcul de l'écart type, écart type qui sert au calcul du coefficient de variation. L'on consultera l'annexe 5 pour avoir quelques exemples dont D est inférieur à 0 et qui illustrent des fréquences théoriques très basses.

Mais quel statut à accorder aux vocables à indice de dispersion négatif ? Comme nous le verrons au chapitre 4, ces vocables sont au nombre de 836 dans notre corpus. Nous les retenons dans le vocabulaire que nous sélectionnons parce que notre inventaire est exhaustif.

D'un point de vue didactique, l'indice de dispersion est une donnée très importante. Les mots les plus utiles aux apprenants sont des mots à grande dispersion que l'apprenant rencontre dans des textes variés.

Mais, comme le soulignent Beauchemin *et al.* (1992 : XXX), il est impossible de déterminer des seuils au-delà ou en deçà desquels la dispersion serait significativement élevée ou basse.

Ils estiment, de manière empirique, que des valeurs au-dessus de 0,50 montrent une bonne dispersion et que des valeurs plus petites que 0,20 indiquent une mauvaise dispersion des occurrences d'un vocable donné (Beauchemin *et al.* 1992 : XXX). Nous nous servirons de la dispersion pour évaluer notamment le caractère thématique ou athématique d'un vocable. Les vocables thématiques ont généralement un indice de dispersion bas et des fréquences élévées alors que les vocables athématiques ont des indices de fréquence et de dispersion équilibrés.

2.5.3. L'INDICE D'USAGE (*U*)

L'indice d'usage est le résultat du croisement entre la fréquence et la dispersion. Dans notre cas, il est obtenu par la formule : $U = F_O \times D$

Certaines unités ont une fréquence élevée et une basse dispersion; à l'inverse, d'autres ont une fréquence basse et une grande dispersion. Le croisement des deux données permet de déterminer un indice d'usage, unique, qui permet de tenir compte à la fois de la fréquence et de la dispersion.

Soit par exemple les substantifs suivants où la F_O = la fréquence observée, D = l'indice de dispersion et U l'indice d'usage :

(61)	*Vocable*	*F*o	*D*	*U*	*Glose*
	a. *igihe*	317	0,93	278,78	'moment'
	b. *umugambwe*	412	0,65	268,55	'parti politique'
	c. *umwáana*	245	0,79	205,24	'enfant'

Si l'on se fondait sur la fréquence, on privilégierait le vocable en (61 b). Si l'on se fondait sur la dispersion, l'on privilégierait les vocables en (61 a) et (61 c) avec un avantage pour (61 a) si l'on tenait subsidiairement compte de la fréquence. Mais si l'on tient compte de U, on constate que le vocable en (61 a) constitue le premier choix - mais pas très loin devant *umugambwe* - ce dont nous convenons.

L'usage est un critère central pour la sélection des vocables à des fins didactiques que ce soit en enseignement de la langue maternelle ou de la langue seconde (Beauchemin *et al.* 1992 : XXXIV). Nous privilégierons ce critère tout au long de ce travail.

Ce critère présente cependant une limite. Dans leur étude sur le québécois parlé, Beauchemin *et al.* (1992) sont arrivés à la conclusion qu'il est difficile d'établir un seuil significatif de U. Dès lors, dans une étude comme la nôtre où, pour des raisons pédagogiques, l'indice d'usage U est central, il faut résoudre le problème des limites des tranches au sein de la liste des vocables de tout le corpus. Où s'arrête la tranche des vocables à haut, moyen et bas indice d'usage ? Les limites sont empiriques.

Nous avons établi trois tranches de vocables : ceux dont U est supérieur ou égal à 3, ceux dont U est compris entre 3 et 0 et ceux dont U est inférieur à 0. Les vocables dont U est supérieur ou égal à 3 représentent 87 % des mots-formes du corpus, ce qui nous semble suffisant pour affirmer que ces vocables constituent le noyau du vocabulaire de base du kirundi écrit.

Aux fins de la présente étude, nous adoptons les symboles suivants :

F_O	=	indice de fréquence
D	=	indice de dispersion
U	=	indice d'usage
v	=	coefficient de variation théorique
X	=	moyenne
σ	=	écart type

2.6. LES RÉSULTATS ATTENDUS

Les résultats attendus sont de deux types : lexicaux et morphologiques.

Au niveau lexical, notre recherche aboutit à une liste de vocables de tout le corpus. L'analyseur fournit du même coup le nombre de verbes, de substantifs à base verbo-nominale et à base nominale ainsi que le nombre d'adjectifs utilisés dans le corpus, ce qui permet d'analyser la fréquence et la stabilité des différentes catégories grammaticales.

Au niveau morphologique, notre recherche fournit des données sur la fréquence des morphèmes dans le corpus. Les calculs ont exclusivement porté sur les morphèmes du mot-forme verbal et du mot-forme substantif.

Pour le mot-forme verbal, l'analyseur morphologique donne la fréquence des radicaux, des affixes de dérivation et des morphèmes aspectuels.

Pour le substantif à base verbo-nominale, l'analyseur morphologique donne la fréquence des radicaux, des préfixes de classe, des affixes de dérivation et des voyelles thématiques nominales.

Pour le substantif à base nominale, l'analyseur morphologique donne la fréquence des radicaux et celle des préfixes de classe.

Nous fournissons au chapitre 3 de notre étude une description et un inventaire de ces morphèmes. Nous y précisons également les raisons qui ont guidé le choix des morphèmes objets de comptages. Avant cela, présentons les principaux aspects retenus pour l'analyse de la liste des vocables.

3. L'ANALYSE DE LA LISTE DES VOCABLES

Une fois la liste des vocables de notre corpus établie, commence le travail d'analyse. Il présente deux volets : un volet quantitatif et un volet qualitatif.

3.1. LE VOLET QUANTITATIF

Le volet quantitatif aborde différents aspects numériques de la liste. Nous la subdivisons d'abord en trois grandes tranches, la première regroupant des vocables dont l'indice d'usage (désormais U) est supérieur ou égal à 3, la seconde ceux dont U est compris entre 3 et 0, la dernière regroupant les vocables dont U est inférieur à 0. À l'intérieur de chaque tranche, nous examinons la part des différentes catégories grammaticales.

Ainsi, pour chacune des tranches, nous indiquons :

- le nombre de vocables par catégorie grammaticale;
- le nombre d'occurrences couvertes par la tranche et le rapport avec N;
- la fréquence moyenne des vocables de la tranche selon la catégorie grammaticale.

Pour chaque catégorie de vocable nous indiquons :

- le nombre d'occurrences couvertes par la catégorie de vocable à l'intérieur de la tranche;
- le nombre de vocables dont la fréquence dépasse la fréquence moyenne des vocables de la tranche.

Nous distinguons pour chaque catégorie de vocables ceux qui sont thématiques (caractérisés par un indice de fréquence élevé et un indice de dispersion bas) de ceux qui sont généraux (dont les indices de dispersion et d'usage sont équilibrés).

La distinction vocable thématique / vocable général est retenue pour les vocables dont U ≥ 3 et dont la moyenne de fréquence est ≥ 66. Lorsque la fréquence baisse, ce qui est le cas des vocables dont U est inférieur à 3, les vocables tombent dans la catégorie des mots rares. Nous les analysons à la lumière de la caractérisation formulée par Ménard (1978 : 37-39). Ce dernier a constaté pour le français que, d'un point de vue morphologique, les vocables rares sont généralement des vocables longs. Un bon nombre de vocables rares sont également des emprunts.

D'un point de vue sémantique, les vocables rares constituent des créations nouvelles ou des emprunts qui répondent à des besoins nouveaux de désignation. Il s'agit souvent de termes scientifiques et techniques. Certains vocables rares doivent leur présence à des choix stylistiques des auteurs (Ménard 1978 : 37-38).

Nous dressons l'inventaire des emprunts rencontrés dans notre corpus. Précisons cependant que nous ne discutons pas des problèmes théoriques liés à l'intégration des emprunts en kirundi.

Les résultats nous permettent de comparer les données sur la fréquence des catégories grammaticales en kirundi avec les résultats disponibles pour le français et l'anglais. Nous prolongeons cette comparaison sur les 50 vocables les plus fréquents. Nous concluons l'analyse par un regard sur les noms propres et les abréviations contenus dans le corpus.

3.2. LE VOLET QUALITATIF

Le volet qualitatif de l'analyse de la liste vise à cerner certains aspects linguistiques d'importance majeure en didactique des langues. Ces aspects sont essentiellement morphosyntaxiques et sémantiques.

Les aspects morphosyntaxiques concernent beaucoup plus les verbes que les noms ou les vocables grammaticaux. Nous abordons la structure morphologique des vocables verbaux et leurs aspects syntaxiques en rapport avec leurs fréquences.

Nous abordons les aspects sémantiques de la liste en analysant les relations lexico-sémantiques entre les vocables. Nous privilégions les relations paradigmatiques (Lipka 1990 : 140-150). Il s'agit principalement des relations hiérarchiques (hypéronymie, hyponymie), de la synonymie et de l'antonymie, relations de première importance en didactique des langues.

Signalons que les relations lexico-sémantiques privilégiées par l'analyse dépendent de la catégorie grammaticale. On sait par exemple qu'en français, les relations d'antonymie sont généralement exprimées par des adjectifs (Picoche 1977 : 101). L'on ne s'étonnera donc pas que l'accent soit mis sur certaines relations plutôt que d'autres selon les catégories de vocables qui feront l'objet d'analyse.

4. CONCLUSION

Le deuxième chapitre constituait une présentation de la méthodologie de l'étude. Celle-ci étant fondée sur un corpus écrit, nous y avons présenté comment s'est effectué l'échantillonnage aléatoire simple et l'échantillonnage aléatoire stratifié proportionnel.

Nous avons indiqué ensuite comment s'effectue le traitement des données en décrivant le rôle des différents outils informatiques utilisés : *WordPerfect 5.1* pour la saisie, *WordCruncher 4.23* pour l'indexation, l'analyseur morphologique pour l'analyse morphologique automatique, et *SPSS 7.5* pour la détermination du vocabulaire de base et des indices statistiques sur les vocables et les morphèmes. Nous avons enfin indiqué les principaux aspects retenus pour l'analyse des résultats.

Il nous faut maintenant fournir une description et une typologie des mots-formes et des morphes de la langue. Tel est l'objet du chapitre 3.

CHAPITRE 3

LA MORPHOLOGIE DU KIRUNDI

1. INTRODUCTION

On peut distinguer de manière schématique trois types de morphologie en regard des unités considérées comme élémentaires dans la structure du mot (Corbin 1987 : 182-183).

Il y a d'abord une morphologie du morphème. Elle est développée par les théories structuralistes (Bloomfield 1933) et par certains générativistes comme Halle (1973). Elle postule que la formation des mots opère à partir de morphèmes.

Il existe aussi une morphologie du mot. De nombreuses recherches, comme celles d'Aronoff (1976), Booij (1977) et Scalise (1984) vont dans ce sens. Dans ce modèle morphologique, l'on considère que les mots se forment à partir d'autres mots et non à partir d'unités plus petites comme peuvent l'être nombre de morphèmes.

Finalement, on peut reconnaître une morphologie du mot-morphème, une option qui combine les deux précédentes. Le lexique est alors conçu comme contenant des mots et des morphèmes. Corbin (1987) illustre ce courant de recherche.

Les écrits que nous avons pu consulter sur le kirundi se situent tous dans une morphologie du morphème. Ceci explique que nous admettons, quand besoin est, la présence d'un morphème zéro dans la décomposition des mots-formes.

A priori, le choix d'un modèle de morphologie ou d'un autre est indépendant de la statistique lexicale, dont les unités de travail sont les mots-formes et les vocables.

Il reste que, dans une langue agglutinante, traditionnellement conçue comme fondée sur la concaténation de morphèmes en mots-formes, la seule liste de ces derniers (ou celle des vocables) les plus fréquents ne nous paraît pas constituer une

donnée suffisante en regard de notre objectif : l'élaboration d'un vocabulaire de base du kirundi écrit à l'usage des didacticiens de la langue.

Bref, il nous semble essentiel de mettre la morphologie à contribution de façon à compléter les index des mots-formes et des vocables par un index de fréquence de morphèmes, pour en privilégier l'exploitation dans les vocabulaires de base.

Dans les lignes qui suivent, nous situons d'abord brièvement les recherches sur la morphologie du kirundi dans leurs cadres théoriques respectifs[1]. Nous abordons ensuite la structure et la typologie des mots-formes du kirundi.

2. LES MODÈLES MORPHOLOGIQUES APPLIQUÉS AU KIRUNDI

Les études déjà parues sur la morphologie du kirundi relèvent de la morphologie compositionnelle (concaténatoire) et privilégient la théorie *Items et Arrangements*. C'est dans ce cadre que se situent les descriptions du kirundi de Meeussen (1959) et Rodegem (1967) dont nous nous servons dans la présente étude.

Dans une morphologie compositionnelle, deux représentations de la structure du mot sont possibles : soit une représentation linéaire, soit une représentation hiérarchique.

Des études récentes sur le kirundi ont opté pour une représentation hiérarchique. On peut citer Nkanira (1984) pour une analyse des temps verbaux du kirundi dans une perspective psychomécanique, Jouannet (1989) et Goldsmith & Sabimana (1989) pour leurs études des déplacements tonaux dans une perspective autosegmentale et métrique.

---

[1] L'aperçu que nous donnons s'inspire largement des commentaires de M. John Reighard, membre de notre jury de maîtrise. Qu'il en soit remercié.

Goldsmith et Sabimana (1989) proposent par exemple pour le verbe du kirundi une structure hiérarchique à trois niveaux inspirée de la phonologie lexicale (cf. Kiparsky 1982). Au premier niveau se trouvent, dans l'ordre : le radical, les suffixes de dérivation verbale et le morphème aspectuel. Au second niveau se trouvent, dans l'ordre : le morphème réfléchi, le(s) morphème(s) complément(s), le morphème temporel, le négateur *-ta-* et le préfixe verbal personnel. Au troisième niveau s'inscrit le morphème dicto-modal *nti-* (dont nous adoptons la séparabilité d'avec le verbe).

Concrètement, pour un mot-forme verbal comme *nti bamurimira* 'ils ne cultivent pas pour lui', on obtient la représentation suivante :

Niveau 1	radical	[-rim-]
	suffixe	[-ir-]
	aspect	[-a]
Niveau 2	complément	[-mu-]
	temps	[-Ø-]
	préfixe verbal personnel	[ba-]
Niveau 3	nti	[nti]

Tableau 11 - *Représentation hiérarchique d'un verbe selon Goldsmith & Sabimana (1989)*

Dans cette représentation, le mot-forme contient un morphème central (le radical) que d'autres morphèmes modifient de manières diverses. Dans notre analyse linéaire, le même mot-forme est représenté comme suit :

(62) [nti]$_{dmod.}$ [ba]$_{préf.pers.}$ [Ø]$_{préd.v.}$ [mu]$_{compl.}$ [rim]$_{rad.}$ [ir]$_{suff.}$ [a]$_{asp.}$
'ne...pas' 'ils' 'présent' 'lui' 'cultiver' 'pour' 'inaccompli'
'ils ne cultivent pas pour lui'

Dans l'optique de notre recherche, le choix d'une analyse linéaire ou hiérarchique n'est pas crucial. Pour nous, le problème de base consiste plutôt à décider du sort à réserver aux morphèmes et à dégager ceux qui contribueraient le plus à caractériser un vocabulaire de base du kirundi.

Pour ce faire, nous décrivons ci-dessous, dans un modèle linéaire la structure des mots-formes du kirundi et en proposons une typologie.

La description permettra d'analyser les structures des mots-formes et de mieux mettre en évidence les morphèmes que nous privilégions dans l'élaboration du vocabulaire de base; nous estimons en effet nécessaire d'inclure dans notre recherche l'index de fréquence de certains morphèmes. Nous optons ainsi pour le découpage en morphèmes des mots-formes lexicaux car leurs radicaux sont essentiels dans un vocabulaire de base; ce découpage ne s'impose pas pour les mots-formes grammaticaux qui sont régis par la syntaxe de la langue.

La typologie des mots-formes du kirundi permet quant à elle d'établir des catégories grammaticales et leurs sous-catégories morphosyntaxiques; ces données sont essentielles notamment dans le traitement informatique des données.

3. LA STRUCTURE MORPHOLOGIQUE DES MOTS-FORMES

ASPECTS GÉNÉRAUX

3.1.1. ASPECTS TONALS

Le kirundi est une langue à tons avec une opposition pertinente ton haut / ton bas. Le site du ton est la voyelle de la syllabe. Du fait que le ton bas est le plus fréquent, il est usuel de ne noter que les tons hauts. Nous examinerons au chapitre 5 § 4.2. les conséquences sur la lecture.

Comme les journaux que nous dépouillons ne notent pas les tons (ni la quantité vocalique), nous avons décidé (cf. chapitre 2 § 2.1.1), aux fins de la désambiguïsation, de restituer au mot-forme ses tons de manière à produire des index de mots-formes interprétables. Sans cela, on ne saurait pas le sens du mot-forme du corpus *abana* qui peut avoir les prononciations et les sens suivants :

(63) *abáana* 'les enfants'
abaaná 'qui vit avec *x*'

Cette décision de noter les tons connaît deux contraintes importantes.

D'une part, nous évacuons de notre étude le problème des changements de tons dûs au contexte linguistique. Soient les mots-formes *indá* 'pou', *inda* 'ventre' (rendus uniformément dans le corpus par le mot-forme *inda*) et *na* 'et'; les tons des deux mots-formes lexicaux changent comme suit :

(64) *indá n'índa* 'le pou et le ventre'

Aux fins de la reconnaissance de toutes les occurrences d'un mot-forme, nous les décontextualisons dans la saisie en adoptant la réalisation de base, c'est-à-dire le mot-forme tel qu'il se réalise en isolation. Nous adoptons ici la notion de ton lexical. Dans le contexte en (64), nous avons saisi *inda* 'ventre' et non **índa* , *inda* représentant 'pou'.

D'autre part, nous ignorons le déplacement tonal à l'intérieur des mots-formes. Soit l'infinitif *kurima* 'cultiver' et sa forme au passé éloigné[2] *naríma* 'je cultivais' en (65 c) :

(65) a. *kurima* [ku]préf.inf. [rim]rad. [a]asp. 'cultiver'
b. *naríma* [n]préf.suj. [á]passé éloigné [rim]rad. [a]asp. 'je cultivais'
c. **nárima*

Le mot-forme régulier serait (65 c); il est agrammatical. Le mot-forme grammatical est celui en (65 b), où le ton haut se trouve sur la voyelle /i/ alors qu'il était attendu sur la voyelle /a/ de la syllabe précédente, compte tenu de la structure morphologique. Il y a eu déplacement du ton haut.

L'insertion de morphèmes entre le morphème du passé éloigné [á] et le radical [-rim-] provoque également un déplacement tonal de la voyelle /a/ sur /i/ en (66 a) et de /i/ sur /a/ en (66 b) :

---

[2] Le morphème du passé éloigné est [-á-].

(66) a. *nabírima* 'je les cultivais'

[n]préf.pers. [á]préd.v. [bi]compl. [rim]rad. [a]asp.

'je 'passé éloigné 'les' cultiver' 'inaccompli'

b. *nahábirima* 'je les y cultivais'

[n]préf.pers. [á]préd.v. [ha]compl. [bi]compl. [rim]rad. [a]asp.

'je' 'passé éloigné' 'y' ' les' cultiver' 'inaccompli'

Goldsmith & Sabimana (1989), de même que Jouannet (1989), proposent dans le cadre de la phonologie autosegmentale et métrique une analyse du déplacement tonal et des règles qui le régissent en kirundi. Nous en faisons ici l'économie, d'autant plus que l'analyseur morphologique que nous utilisons n'est pas à même d'en fournir un traitement ni qualitatif ni quantitatif.

Les mots-formes que nous retenons en (66) sont *nabírima* 'je les cultivais' et *nahábirima* 'je les y cultivais' et en (65) *kurima* 'cultiver' et *naríma* 'je cultivais'.

3.1.2. DURÉE VOCALIQUE

En kirundi, la quantité vocalique est distinctive mais l'orthographe actuelle de la langue ne la marque pas. Aux fins de la désambiguïsation des homographes, nous la marquons en doublant la voyelle, ce qui permet, par exemple, de distinguer *gusesa* 'fouiller' de *guseesa* 'renverser' qui dans les textes sont rendus uniformément par *gusesa*.

Nous n'avons pas noté la longueur vocalique prévisible. Il s'agit notamment de la longueur de la voyelle précédant une séquence NC (consonne nasale + consonne) et de celle qui porte sur l'initiale vocalique du suffixe de dérivation [-ish-] (cf. chapitre 3 § 3.2.1.3.1).

3.1.3. CLASSES D'ACCORD

Un autre aspect majeur du mot-forme en kirundi est l'accord de classe : le verbe, l'adjectif, le déterminant et le pronom s'accordent avec le substantif comme illustré dans l'exemple ci-dessous :

(67) ***ab**antu **baa**we **ban**íni **ba**zoorima* 'tes grandes personnes cultiveront'

abantu*n.*	**baa**we*det.*	**ban**íni*adj.*
'personnes'	'tes'	'grandes'

bazoorima*v.*

[ba]préf.pers. [zoo]préd.v. [rim]rad. [a]asp.

'cultiveront'

On observe en (67) la récurrence du morphème [ba] que ce soit pour le préfixe de classe (préf.cl.), le préfixe déterminatif (préf.dét.), le préfixe adjectival (préf.adj.) ou le préfixe personnel (préf.pers.). Il y a homographie des préfixes.

Le tableau suivant synthétise l'accord en kirundi. Rappelons que c'est le substantif qui, par son préfixe de classe, commande l'accord. Cela explique pourquoi le tableau est construit sur les préfixes de classe.

préfixe de classe / adjectival	*préfixe déterminatif / pronominal*	*préfixe personnel*[3]	*préfixe objet*
1. mu-	u-	a-	-mu-
2. ba-	ba-	ba-	-ba-
3. mu-	u-	u-	-u-
4. mi-	i-	i-	-i-
5. ri- / Ø	ri-	ri-	-ri-
6. ma-	a-	a-	-a-
7. ki-	ki-	ki-	-ki-
8. bi-	bi-	bi-	-bi-
9. n- /Ø	i-	i-	-i-
10. n- /Ø	zi-	zi-	-zi-
11. ru-	ru-	ru-	-ru-
12. ka-	ka-	ka-	-ka-
13. tu-	tu-	tu-	-tu-
14. bu-	bu-	bu-	-bu-
15. ku-	ku-	ku-	-ku-
16. ha-	ha-	ha-	-ha-

Tableau 12 - *Les préfixes d'accord du kirundi*

Nous distinguons dans ce tableau deux types d'homographies morphologiques : une homographie par rapport à une classe d'accord (la classe 1 par exemple) et une homographie entre classes d'accord (entre les classes 1 et 3 par exemple).

---

[3] Il faut y ajouter les préfixes verbaux personnels suivants dont certains sont homographes à ceux du tableau : [n-] '1re pers.sg.', [u-] '2e pers.sg.', [a-] '3e pers.sg.', [tu-] '1re pers.pl.', [mu-] '2e pers.pl.', [ba-] '3eme pers.pl.'.

3.1.3.1. HOMOGRAPHIE À L'INTÉRIEUR D'UNE CLASSE D'ACCORD

L'on peut, à partir du tableau ci-dessus, distinguer deux types d'homographies à l'intérieur d'une classe d'accord : une homographie totale et une homographie partielle.

3.1.3.1.1. Homographie totale

L'homographie totale concerne les classes 2 (*ba*), 7 (*ki*), 8 (*bi*), 10 (*n*), 11(*ru*), 12 (*ka*), 13 (*tu*), 14 (*bu*), 15 (*ku*) et 16 (*ha*). Elle est illustrée avec la classe 2 (*ba*) par l'exemple déjà donné en (67). Il y a identité de tous les préfixes (**ba**), qu'ils soient du nom (*abantu* 'personnes'), du possessif en emploi déterminantif (*baawe* 'tes'), de l'adjectif (*baníni* 'grandes', ou du verbe (*bazoorima* 'ils cultiveront *x*').

3.1.3.1.2. Homographie partielle

L'homographie partielle à l'intérieur d'une classe d'accord touche les classes 1, 3, 4, 5, 6, 9. Nous indiquons dans le tableau suivant les formes que prennent les préfixes d'accord selon qu'ils appartiennent à un déterminant, à un pronom et à un verbe soit, dans ce dernier cas, comme préfixe personnel ou objet.

Classes des substantifs	*Préf.cl.*	*Préf.pron.*	*Verbe*	
			préf.pers.	*préf.obj.*
cl.1	mu	u	a	mu
cl.3	mu	u	u	wu
cl.4	mi	i	i	yi
cl.5	Ø	i	i	yi
cl.6	ma	a	a	ya
cl. 9	n	i	i	yi
cl.10	n	zi	zi	zi

Tableau 13 - *Homographie partielle des préfixes d'accord*

Comme on peut le constater, ces morphèmes d'accord sont tantôt homographes tantôt hétérographes. Ainsi par exemple, pour les substantifs de la classe 1, la forme [mu] correspond à la fois au préfixe de classe et au préfixe objet. Lorsque l'accord se fait avec un pronom ou un verbe, le préfixe prend les formes respectives [u] et [a] comme l'illustre l'exemple suivant à partir du substantif *umuntu* 'une personne' de la classe 1:

(68) *umuntu* *umwé* *azoomubona*
'personne' 'un'$_{cl.1}$ 'il' futur''le' 'voir'
'une personne le verra'

Un traitement automatique quantitatif de tous les préfixes d'accord supposerait leur désambiguïsation; ce travail est extrêmement fastidieux et peu susceptible de livrer des informations substantielles en matière d'élaboration d'un vocabulaire de base du kirundi. Nous avons donc dû choisir la sous-catégorie des préfixes d'accord qui ferait l'objet de comptages.

Nous considérons que les préfixes personnels, objets et pronominaux / déterminantifs sont régis par la syntaxe. Ils sont de ce fait exclus de notre recherche.

Par contre, les préfixes de classe (homographes aux préfixes adjectivaux), qui commandent l'accord en kirundi, font l'objet de quantification. Nous ne désambiguïsons pas cependant les préfixes de classe homographes. Ainsi, nous ne distinguons pas, par exemple, [mu]$_{cl.1}$ de [mu]$_{cl.3}$.

L'homographie des préfixes d'accord est à l'origine de nombreux homographes. Soit les deux mots-formes suivants où le morphe [-n] est ambigu :

(69) a. *ngira* [n]$_{préf.\ pers.}$ [Ø]$_{préd.v.}$ gir]$_{rad.}$ [a]$_{asp.}$
'je fais'

b. *ngira* [Ø]$_{préd.v.}$ [n]$_{préf.\ compl.}$ [gir]$_{rad.}$ [a]$_{asp.}$
'fais-moi *x*'

Face à ces homographes, on a deux choix : soit on désambiguïse les mots à l'entrée de l'analyseur et on s'astreint à d'innombrables opérations manuelles lors du découpage morphologique, soit on omet la désambiguïsation et on ne distingue pas les deux mots-formes. Nous avons opté pour la deuxième solution d'autant plus que, comme indiqué plus haut, les préfixes personnels et les préfixes compléments ne font pas l'objet de comptages.

Signalons également que l'homographie des morphèmes d'accord est à la base d'une ressemblance formelle entre le verbe à préfixe personnel vocalique (70 b) et le substantif (70 a), ce qui oblige à la désambiguïsation. Soit par exemple les mots-formes en (70) :

(70) a. *umuhana* [u]augm. [mu]préf.cl. [hana]rad.
'enclos'

b. *umuhana* [u]préf.pers.[Ø]préd.v.[mu]préf.objet [han]rad.[a]asp.
'tu' 'le' 'punir' 'inaccompli'
'tu le punis'

Il est clair qu'il s'agit en (70) de mots-formes distincts. Il faut dès lors désambiguïser les deux mots-formes en collant au mot-forme le moins fréquent un discriminant morphosyntaxique (v = verbe), ce qui donne : *umuhana*-v 'tu le punis' et *umuhana* 'enclos'. L'on comprendra que tous les mots-formes portant des discriminants sont à traiter manuellement lors de l'analyse morphologique.

Signalons que les préfixes d'accord permettent notamment de distinguer, en kirundi, l'adjectif du déterminant et du pronom : s'il y a identité du préfixe d'accord avec le préfixe de classe du substantif, l'unité considérée est un adjectif (71 a); dans le cas contraire, c'est un déterminant ou un pronom comme en (71 b).

(71) a. ***umuntu*** ***munini***
[u]augm. [mu]préf.cl. [ntu]rad. [mu]préf.adj. [níni]rad.
'une personne de grande taille'

b. *umuntu* *wanje*
[u]augm. [mu]préf.cl. [ntu]rad. [u]préf.dét. [anje]rad.
'mon homme'

3.1.4. ASPECTS MORPHOPHONOLOGIQUES

Le mot-forme lexical du kirundi est marqué également par le caractère agglutinant de la langue, particulièrement prononcé pour le verbe et les substantifs à base verbo-nominale. Dans Ntirampeba (1993 : 23-27), nous avons montré que la formation de ces deux types de mots-formes opère à partir d'un noyau (le radical) auquel s'agglutinent des morphèmes en nombre variable.

La formation des mots-formes recourt à un ensemble de règles phonologiques et morphophonologiques que nous avons évoqué au chapitre 2 lors de la présentation de l'analyseur morphologique du kirundi dont nous nous servons dans cette étude. Ces règles servent à la reconnaissance des morphèmes à partir des mots-formes du corpus.

La présentation de ces règles permettra de comprendre les découpages morphologiques que nous proposons tout au long du chapitre 3 et de se faire déjà une idée des principales difficultés que rencontre le traitement automatique de la morphologie du kirundi.

Ces règles sont au nombre de treize, soit six règles phonologiques et sept règles morphophonologiques. Nous les présentons en nous inspirant directement de Mel'cuk & Bakiza (1987).

Selon Mel'cuk (1996 : 18-19), les règles correspondent à des modifications du signifiant entraînées soit par un contexte phonologique, soit par des facteurs phonologiques agissant dans un contexte morphologique, soit tout simplement par un contexte morphologique. Dans le premier cas, il parle de règles phonologiques,

dans les second et troisième cas de règles morphophonologiques. Nous retenons cette distinction.

Nous partons des mots-formes tels qu'ils apparaissent dans le corpus. Ils sont notés en italique. Les formes à droite de la flèche correspondent à la représentation phonologique que nous postulons pour les divers morphèmes[4]. Par convention, et pour fins de lisibilité, nous les notons entre crochets [ ]. Le signe => signifie que l'analyseur morphologique opère la reconstruction morphologique résultante.

3.1.4.1. RÈGLES PHONOLOGIQUES

3.1.4.1.1. Assimilation régressive de la consonne nasale /n/ devant une consonne labiale

On peut illustrer l'assimilation régressive de la nasale /n/ devant une consonne labiale avec le préfixe de classe /n-/ et le préfixe verbal personnel /n-/ qui se réalisent [m] à une frontière morphologique quand ils sont suivis des labiales /p/ et /b/ :

(72) ***m**piima* 'je mesure' => [**n**]$_{\text{préf.pers.}}$ [**p**iim]$_{\text{rad.}}$ [a]$_{\text{asp.}}$

***m**boma* 'je cogne' => [**n**]$_{\text{préf.pers.}}$ [**b**om]$_{\text{rad.}}$ [a]$_{\text{asp.}}$

---

[4] L'on comprend dès lors les nombreuses distorsions formelles entre le mot-forme tel qu'orthographié et sa structure morphologique.

3.1.4.1.2. Consonantisation des voyelles

La consonantisation des voyelles a lieu à la frontière des morphèmes. Les voyelles /i/ et /a/ se résolvent en la semi-voyelle [j] devant une autre voyelle tandis que /u/ se résoud en [w] tel qu'illustré en (73) :

(73) a. *yandika* =>[**i**]préf.pers.[Ø]préd.v. [**a**ndik]rad. [a]asp.'il écrit' (+ inanimé)[5]
b. *yandika* =>[**a**]préf.pers. [Ø]préd.v. [**a**ndik]rad. [a]asp.'il écrit' (+ humain)[6]
c. *wandika* =>[**u**]préf.pers. [Ø]préd.v. [**a**ndik]rad. [a]asp. 'tu écris'

3.1.4.1.3. Assimilation progressive de /r/ après la nasale /n/

Nous illustrons en (74) l'assimilation progressive de /r/ après la nasale /n/ située à la frontière morphologique :

(74) ***n**dima* 'je cultive' => [**n**]préf.pers. [Ø]préd.v. [**r**im]rad. [a]asp.

3.1.4.1.4. Assimilation progressive de /h/ devant la nasale /n/

Nous illustrons en (75) l'assimilation progressive de /h/ après la nasale /n/ située à la frontière morphologique, /n/ s'assimilant ensuite comme au § 3.1.4.1.1 :

(75) *im**p**éne* 'chèvre' => [i]augm. [**n**]préf.cl. [**h**éne]rad.

---

[5] Comme dans : *Ikáraámu yandika néeza*
'le stylo écrit bien'

[6] Comme dans : *Yohaáni yandika néeza*
'Jean écrit bien'

3.1.4.1.5. Troncation de /a, i/ après /C/ devant une voyelle

La troncation de /a, i/ après /C/ devant une des cinq voyelles de la langue est illustrée en (76) :

(76) *abeémera* => [a]préf.pers.[Ø]préd.v. [**ba**]préf.compl. [**éemer**]rad. [a]asp.
'il les croit'
azuugara => [a]préf.pers. [Ø]préd.v. [**zi**]préf.compl. [**uugar**]rad. [a]asp.
'il les ferme'

3.1.4.2. RÈGLES MORPHOPHONOLOGIQUES

3.1.4.2.1. Chute des consonne

La chute de consonne s'observe pour les phonèmes /n/ et /j/. Lorsque la consonne initiale du radical est /n/ et que ce morphème se répète en position pré-radicale comme morphème complément ou personnel, les deux consonnes se résolvent en une seule.

(77) *niga* => [**n**]préf.pers. [Ø]préd.v. [**n**ig]rad. [a]asp.
'je' 'présent' 'étouffer' 'non accompli'
'j'étouffe'

azooniga => [a]préf.pers. [zoo]préd.v. [**n**]préf.compl. [**n**ig]rad. [a]asp.
'il' 'futur' 'moi' 'étouffer' 'non accompli'
'il m'étouffera'

Nous illustrons la chute de [j] en (78). Lorsque /j/ se retrouve en finale de radical et que le morphème aspectuel est /-je/, les deux semi-voyelles se résolvent en une seule.

(78) *nagaye* => [n]préf.pers. [a]préd.v. [gay]rad. [ye]asp.
'moi' 'passé récent' 'mépriser' 'accompli'
'j'ai méprisé'

3.1.4.2.2. Spirantisation

Les phonèmes impliqués dans la spirantisation sont /p/, /b/, /t/, /d/, /g/, /k/, /r/ en finale de radical au contact du morphème de dérivation causatif [i] ou du morphème aspectuel /ye/.

(79)	*acafye*	=> [a]préf.pers. [Ø]préd.v. [cap]rad. [ye]asp.	'il dessine'
	abivye	=> [a]préf.pers. [Ø]préd.v. [bib]rad. [ye]asp.	'il ensemence'
	afise	=> [a]préf.pers. [Ø]préd.v. [fit]rad. [ye]asp.	'il possède'
	arunze	=> [a]préf.pers. [Ø]préd.v. [rund]rad. [ye]asp.	'il entasse'
	kuduza	=> [ku]préf.inf. [duug]rad. [i]caus.[a]asp.	'faire monter
	kuriza	=> [ku]préf.inf. [rir]rad. [i]caus. [a]asp.	'faire pleurer'
	asutse	=> [a]préf.pers. [Ø]préd.v. [suk]rad. [ye]asp.	'il verse'

Comme on le voit, /p/ se résoud en /f/, /b/ en /v/, /t/ en /s/, /d, g, r/ en /z/ et /k/en /ts/.

3.1.4.2.3. Contraction des voyelles

La contraction se rencontre essentiellement dans la formation des adjectifs qualificatifs. Lorsque le préfixe de classe porte en finale la voyelle /a/ et que le radical adjectival commence par la voyelle longue /ii/[7] les deux voyelles se résolvent en /eé/ tel qu'illustré en (80) :

(80)	*beénshi*	=>	[ba]préf.cl. [iínshi]rad.	'nombreux'
	beezá	=>	[ba]préf.cl. [iizá]rad.	'beaux'

[7] Avec ou sans ton haut sur la deuxième voyelle /i/.

3.1.4.2.4. Loi de Dhal

La loi de Dhal exprime le fait que la cooccurrence immédiate de deux syllabes à consonne sourde initiale se résoud par le voisement de la première. Les consonnes sujettes à cette loi sont /k/ et /t/ qui deviennent respectivement /g/ et /d/. Par exemple, *gusaba* 'demander' et *kudátebá* 'ne pas traîner' sont analysés comme suit :

(81) a. ***g**usaba* [**k**u]préf.inf. [**s**ab]rad. [a]asp.
'demander'

b. *kudátebá* [ku]préf.inf. [**t**á]préf.nég. [**t**eb]rad. [a]asp.
'ne pas traîner'

3.1.4.2.5. Palatalisation

Soit les noms en (82) et leur découpage morphologique. Lorsque, en frontière de morphème, les morphèmes /ki/ et /bi/ sont adjacents à un autre morphème portant à l'initiale une des cinq voyelles de la langue, il s'opère une troncation de la voyelle /i/ et /k/ se résoud en /ʧ/ (l'orthographe du kirundi le note *c*) tandis que /b/ se résoud en /vy/.

(82)	a. *icaátsi*	=> [i]augm. [**ki**]préf.cl. [**aá**tsi]rad.		'herbe'
	b. *ivyaátsi*	=> [i]augm. [**bi**]préf.cl. [**aá**tsi]rad.		'herbe'[pl.]
	c. *icúuma*	=> [i]augm. [**ki**]préf.cl. [**úu**ma]rad.		'métal'
	d. *ivyúuma*	=> [i]augm. [**bi**]préf.cl. [**úu**ma]rad.		'métal' [pl.]

3.1.4.2.6. Épenthèse

L'épenthèse touche exclusivement les lexèmes monosyllabiques de la forme CV qui entrent dans la formation de mots-formes verbaux ou nominaux. Elle intervient pour rompre des séquences interdites dans la langues notamment dans la

construction des mots-formes avec le suffixe du passif [-u-]. Soit les exemples en (83) :

(83) a. *kuryá* 'manger'
b. *kurîibwa* 'être mangé'

Ces formes ont la structure morphologique en (84) où on observe pour (84 b) une suite de quatre voyelles (VVV)

(84) a. *kuryá* [ku]préf.inf. [**rí**]rad. [a]asp.
'manger'
b. *kurîibwa* [ku]préf.inf. [**rí**]rad.[u]suff.pass. [a]asp.
**kurîiua*
'être mangé'.

La séquence VVV n'étant pas admise par la langue, il y a épenthèse de /b/ précédée d'une insertion de /i/ qui allonge le radical et suivie de la formation d'une glissante /w/.

3.1.4.2.7. Harmonie vocalique

L'harmonie vocalique est un phénomène d'assimilation vocalique qui peut affecter plusieurs voyelles d'un même mot et qui a souvent pour but de rapprocher le timbre d'une voyelle de celui d'un phonème voisin. En kirundi l'harmonie vocalique s'observe dans les dérivés mettant en jeu le morphème applicatif [-ir-] qui devient [-er-] lorsque le radical porte la voyelle /e/ ou /o/ tel qu'illustré en (85) :

(85) *kugendera* [ku]préf.inf. [**gend**]rad.[**ir**]suff.appl. [a]asp. 'partir pour'
**kugendira*

kuvoomera [ku]préf.inf. [**vo om**]rad.[**ir**]suff.appl. [a]asp. 'puiser de l'eau pour'
**kuvomira*

Les exemples en (85) montrent que le suffixe applicatif s'harmonise à l'aperture des voyelles (/e/, /o/). Signalons que l'harmonie vocalique est inopérante entre la voyelle fermée du morphème [-ir-] et les autres voyelles de la langue /u/, /a/ et /i/.

Telles sont les règles avec lesquelles fonctionne notre analyseur morphologique. Elles servent à la reconnaissance des morphèmes du corpus. Nous analysons au chapitre 4 les problèmes spécifiques au traitement automatique de la morphologie du kirundi. Pour l'instant, revenons à la structure des mots-formes, mots-formes à partir desquels nous déterminerons le vocabulaire de base du kirundi écrit.

Nous distinguons au sein des mots-formes du kirundi deux grandes classes : les mots-formes lexicaux et les mots-formes grammaticaux.

Les mots-formes lexicaux comprennent les verbes, les substantifs et les adjectifs; ils sont tous fléchis. Parmi les mots-formes grammaticaux, nous distinguons ceux qui sont fléchis de ceux qui ne le sont pas (prédicatifs nominaux, prépositions, conjonctions, adverbes, interjections et onomatopées). Les mots grammaticaux fléchis assurent la fonction de pronom ou de déterminant. Nous décrivons toutes ces catégories aux §§ 3.2. et 3.3.

3.2. MOTS-FORMES LEXICAUX

3.2.1. MOT-FORME VERBAL

Sous sa forme la plus courte, le mot-forme verbal du kirundi est formé de trois morphèmes : un radical, un prédicatif verbal et un morphème aspectuel, appelé aussi dérivatif thématique verbal (cf. Houis 1977 : 26-27); ce mot-forme correspond à l'impératif; cf. ex. (86) :

(86) *rima* [Ø]$_{\text{préd.v.}}$ [rim]$_{\text{rad.}}$ [a]$_{\text{asp.}}$ 'cultive' (impératif)

3.2.1.1. RADICAUX ET MORPHÈMES ASPECTUELS

Les radicaux verbaux constituent un ensemble ouvert alors que les morphèmes aspectuels forment un ensemble fermé de trois éléments. Ntirampeba (1993 : 39) fournit la description des trois morphèmes aspectuels du kirundi qui sont :

- [-a] 'action qui se déroule encore' (inaccompli)
- [-ie] 'action qui s'est totalement déroulée' (accompli)
- [-e] 'action qui va se dérouler ' (inaccompli inchoatif)

La catégorie de l'aspect est fondamentale pour le verbe. D'un point de vue psycholinguistique (Mel'cuk 1994 : 97), l'aspect est la catégorie à partir de laquelle l'enfant apprend les distinctions temporelles.

De plus, pour les langues qui marquent l'aspect morphologiquement[8], les distinctions aspectuelles s'apprennent bien avant toutes les autres (personnelles, temporelles, etc.)[9].

Quant au radical, sa délimitation est essentielle pour notre recherche. Elle permet d'isoler les différents affixes qui feront l'objet des comptages et de décider de ce qui est dérivé et de ce qui ne l'est pas.

Mel'cuk (1993 : 310-315) distingue deux types de dérivés : des dérivés au sens fort (dont le sens est compositionnel) et des dérivés au sens faible (qui doivent être stockés dans le dictionnaire à cause de leur lexicalisation). Pour nous, les premiers sont des dérivés et les seconds sont des lexèmes.

Notre position tranche ici avec la plupart des recherches sur le kirundi effectuées dans le cadre distributionnaliste (Meeussen 1959, Rodegem 1967) où la compositionnalité ou la non-compositionnalité sémantique du dérivé ne constituaient pas le critère de reconnaissance du lexème.

---

8 Notons que tel n'est pas le cas pour le français par exemple.

9 Nous parlons bien entendu de l'apprentissage de la langue maternelle.

Soit par exemple les infinitifs suivants figurant dans le corpus :

(87) a. *kugenda* 'marcher'
b. *kugenduura* 'inspecter'
c. *kugenza* 'épier'

Une perspective distributionnelle laisserait conclure que (87b) et (87c) sont dérivés de (87a) avec respectivement les suffixes oppositif [-uur-] et causatif [-i-] tel qu'illustré en (88)[10] :

(88) a. [ku]préf.inf.[gend]rad.[uur]suff.[a]asp.
b. [ku]préf.inf.[gend]rad.[i]suff.[a]asp.

Nous rejetons le découpage en (88 a, b) sur la base de l'argument que les suffixes posés n'ont ni le sens causatif (pour [-i-]), ni le sens oppositif (pour [-uur]). Nous optons donc pour les radicaux [-genduur-] 'inspecter' et [-genz-] 'épier' :

(89) a. [ku]préf.inf.[genduur]rad.[a]asp. 'inspecter'
b. [ku]préf.inf.[genz]rad.[a]asp. 'épier'

Ainsi, seule la compositionnalité ou la non-compositionnalité du sens des unités nous permet de déterminer le radical.

L'avantage de l'option en (89) est de fournir des index de radicaux sémantiquement interprétables; ce qui ne serait pas le cas des index de radicaux fondés sur le découpage en (88), où les mots-formes formés sur les radicaux de sens 'marcher', 'inspecter' et 'épier' se retrouveraient tous sous le radical [-gend-] malgré des différences sémantiques évidentes.

---

[10] Il suffirait en effet de reconnaître sur une base formelle pour (88 a), le suffixe [-i-] et une règle de spirantisation et pour (88 b) le suffixe [-uur-], suffixes dont le sens resterait à préciser.

3.2.1.2. AUTRES MORPHÈMES DU MOT-FORME VERBAL

Hormis les formes impératives, les mots-formes verbaux du kirundi se composent minimalement de quatre morphèmes :

(90) *bazoorima*
[ba]préf.pers. [zoo]préd.v. [rim]rad. [a]asp.
'ils' 'futur' 'cultiver'
'ils cultiveront'

Le kirundi possède seize préfixes personnels (cf. tableau 12). Ce nombre se ramène à 14 si l'on tient compte de l'homographie et à 15 si l'on y ajoute le préfixe verbal de la 1re personne du singulier [-n-][11].

Le kirundi connaît également cinq prédicatifs verbaux[12] dont la combinatoire est régie par des règles complexes dont nous faisons ici l'économie.

L'on peut dénombrer 225 mots-formes verbaux à partir d'un lexème comme KURIMA 'cultiver' (soit 15 préfixes x 3 morphèms aspectuels x 5 prédicatifs verbaux).

---

[11] Les autres préfixes verbaux sont homographes à des préfixes personnels du tableau 12 p.110.

[12] L'on distinguera les prédicatifs verbaux des particules adverbales *-na-* (à valeur associative) et *-ta-* (à valeur négative); ex. :

(i) *bazoonarima* 'ils cultiveront également'
[ba]préf.pers. [zoo]préd.v. [na]part.adv. [rim]rad. [a]asp.
'ils' 'futur' 'également' 'cultiver'

(ii) *batazóorima* 'ils ne cultiveront pas'
[ba]préf.pers. [ta]préf.nég. [zóo]préd.v. [rim]rad. [a]asp.
'ils' 'négation' 'futur' 'cultiver'

En outre, des morphèmes supplémentaires, optionnels, peuvent s'intercaler et accroître d'autant le nombre de mots-formes sous lesquels tout radical verbal peut se présenter dans la langue, soit :

- un morphème réfléchi *-íí-*, adjacent au radical :

(91) a. *arabóna*
[a]préf.pers. [ra]act. [bón]rad. [a]asp.
'il' 'voir'
'il voit'

b. *arííbona*
[a]préf.pers. [ra]act. [íí]réfl. [bón]rad. [a]asp.
'il' 'se' 'voir'
'il se voit'

- de un à trois pronoms objets du verbe :

(92) a. *bararimira Yohani ibiharage mu murima*
'ils cultivent pour Jean compl. les haricots compl. dans le champ compl.'

b. *barahabimurimira*
[ba]préf.pers. [ra]act. [ha]compl. [bi]compl. [mu]compl. [rim]rad. [ir]suff. [a]asp.
'ils' 'y' 'les' 'lui' 'cultiver' 'pour' 'inaccompli'
'ils les lui cultivent là'

- de un à 6 suffixes de dérivation :

(93) *bazoosomanishirizwa*
[ba]préf.pers.[zoo]préd.v.[som]rad. [an]suff. [ish]suff. [ir]suff. [ir]suff. [i]suff. [u]suff. [a]asp.
'ils' 'fut' 'embrasser' 'ass.' 'caus.' 'appl.' 'appl.' 'caus.' 'pass.' 'inacc.'
'On leur fera embrasser l'un l'autre'

Théoriquement, un mot-forme verbal peut donc comporter jusqu'à 16 morphèmes soit un préfixe personnel, un morphème prédicatif, trois morphèmes compléments, un préfixe réfléchi, un radical, six suffixes de dérivation et un morphème aspectuel.

En combinant tous ces morphèmes, sans tenir compte de la combinatoire à l'intérieur de chaque paradigme, on peut donc estimer à 2 200 le nombre de mots-formes verbaux possibles à partir d'un radical verbal[13]

Mais dans les faits, toutes ces possibilités ne se réalisent pas. Pour le radical [-bón-] 'voir' par exemple, nous avons identifié dans notre corpus 191 mots-formes verbaux ; ils totalisent 462 occurrences.

Les combinaisons qui ne se réalisent pas tiennent de l'incompatibilité sémantique entre les morphèmes, incompatibilité que Mel'cuk (1993 : 358-383) appelle défectivité systématique sémantique. Nous les illustrons dans le tableau 13 pour le radical [-bón-] 'voir'.

Préfixe personnel	*Prédicatifs verbaux*	*Préfixes compléments*	*Radical*	*Suffixes de dérivation*	*Aspect*
94 -	-	-	bon	-	*ye
94a. -	-	-	bon	*am	ye
94b. ba	*zoo	-	bon	-	*ye
94c. ba	a	-	bon	-	*e
94d. ba	á	-	bon	-	*e

Tableau 14 - *Incompatibilité des morphèmes dans le mot-forme verbal*

[13] Soit 225 x 2 (absence / présence du morphème réfléchi) x 2 (absence / présence d'un suffixe de dérivation) x 2 (absence / présence d'un morphème complément) = 2 200.

Le tableau 13 permet de distinguer pour le verbe du kirundi :

- Une incompatibilité entre le prédicatif verbal et l'aspect. On peut distinguer les cas suivants :

  . l'impératif est incompatible avec l'aspect accompli [-ye] (cf. 94); d'où l'agrammaticalité de **bonye*;
  . le morphème du futur [-zoo] est incompatible avec l'aspect accompli (cf. 94 b); d'où l'agrammaticalité de **bazoobonye*;
  . le morphème du passé récent [-a-] est incompatible avec celui de l'aspect inaccompli inchoatif [-e] (cf. 94 c) **baabone*;
  . le morphème du passé éloigné [-á] est incompatible avec l'aspect inaccompli inchoatif [-e] (cf. 94 d) **baábone*.

- Une incompatibilité entre un radical et certains affixes de dérivation; on peut distinguer deux cas :

  . des radicaux ne sont pas suffixables : c'est le cas de [-ri] 'être';
  . un radical exclut certains suffixes de dérivation parce que son sens est incompatible avec celui du suffixe. Le radical [-bón-] 'voir' est par exemple incompatible avec le suffixe statif [-am-] (cf. ex. 94 a).

Hormis l'exemple en (94 a), qui constitue un cas de dérivation, tous les cas d'incompatibilité signalés sont de nature flexionnelle.

Dans le but de cerner de plus près la productivité des processus de dérivation, donnée essentielle dans l'élaboration d'un vocabulaire de base, nous isolerons les suffixes de dérivation les plus fréquents de sorte à dégager les suffixes les plus récurrents dans la formation des verbes et des substantifs à base verbo-nominale. Nous en présentons l'inventaire ci-dessous.

3.2.1.3. SUFFIXES DE DÉRIVATION

À l'exception du morphème réfléchi [-íi-] et du morphème de négation [-ta-] préposés au radical, les suffixes de dérivation du kirundi, ou dérivatèmes (Mel'cuk 1994 : 314), constituent des morphèmes situés à droite du radical verbal. Leur nombre dans le mot-forme varie de 0 à 6. Nous nous servons essentiellement de la classification proposée par Mel'cuk (1994 : 317-387) à laquelle nous associons les travaux de Meussen (1959) et Rodegem (1967) pour l'inventaire.

Mel'cuk (1994) ramène les suffixes de dérivation, à trois grandes classes : les dérivatèmes de contact, les dérivatèmes de manière et une classe de dérivatèmes hétérogène.

3.2.1.3.1. Dérivatèmes de contact

Les dérivatèmes de contact changent le nombre d'actants sémantiques du lexème de départ. *« Ils signalent le lien, ou le type de contact, entre l'actant ajouté / soustrait et la situation de départ».* (Mel'cuk 1994 : 318). Ils regroupent les factitifs, les coopératifs et les applicatifs. Les coopératifs ne sont pas représentés en kirundi.

Selon Mel'cuk (1994 : 318), les factitifs sont des morphèmes de dérivation qui expriment la composante sémantique 'causer'. On distingue en kirundi deux morphèmes causatifs [-i-] et [-ish-] respectivement appelés « causatif direct » et « causatif indirect » du fait que le premier modifie la structure actantielle avec un seul complément (ex. 95) alors que le second (cf. ex. 96) peut en induire deux. Nous les illustrons ci-dessous :

(95) *kurira* [ku]$_{\text{préf.inf.}}$ [rir]$_{\text{rad.}}$ [a]$_{\text{asp.}}$ 'pleurer'
kuriza [ku]$_{\text{préf.inf.}}$ [rir]$_{\text{rad.}}$ [i]$_{\text{suff.caus.}}$[a]$_{\text{asp.}}$ 'faire pleurer
kuriza umwáana 'faire pleurer un enfant'

(96) *kurima* [ku]$_{\text{préf.inf.}}$ [rim]$_{\text{rad.}}$[a]$_{\text{asp.}}$ 'cultiver'
kurimisha [ku]$_{\text{préf.inf.}}$ [rim]$_{\text{rad.}}$ [ish]$_{\text{suff.caus.}}$[a]$_{\text{asp.}}$'faire cultiver'
kurimisha abantu ikawa 'faire cultiver le café aux personnes'

L'applicatif est un morphème typique des langues bantoues et son sémantisme est variable d'une langue à l'autre (Mel'cuk 1994 : 334). Les applicatifs sont représentés en kirundi par l'applicatif [-ir-], il signifie 'pour' comme illustré en (97) :

(97)	*kurima*	[ku]préf.inf. [rim]rad. [a]asp.	'cultiver'
	kurimira	[ku]préf.inf. [rim]rad. [ir]suff.appl. [a]asp.	'cultiver pour'

3.2.1.3.2. Dérivatèmes de manière

Mel'cuk (1994 : 337-346) distingue trois classes de dérivatèmes de manière : les dérivatèmes exprimant un mode d'action, les complémentatifs et le réciproque. Les complémentatifs ne sont pas illustrés en kirundi. Les dérivatèmes exprimant le mode d'action se retrouvent en kirundi sous la forme du perduratif [-iir-], du répétitif [-agur-], de l'intensif [-agan-], de l'intensif duratif [-ang-], de l'intensif ampliatif [-agir-], de l'intensif intransitif [agar-] tandis que le réciproque est représenté en kirundi par le morphème [-an-].

Le dérivatème perduratif [-iir-] est un morphème qui exprime le caractère duratif de l'action :

(98)	*kumiija*	[ku]préf.inf.[miij]rad.[a]asp. 'éparpiller'
	kumijiira	[ku]préf.inf.[miij]rad.[iir]suff.perd.[a]asp. 'éparpiller longuement'

Le dérivatème répétitif [-agur-] est un morphème qui signifie que l'action exprimée par le radical est répétitive.

(99)	*kurima*	[ku]préf.inf.[rim]rad.[a]asp. 'cultiver'
	kurimagura	[ku]préf.inf.[rim]rad.[agur]suff.répét.[a]asp. 'cultiver de façon suivie'

Le suffixe intensif [-agan-] traduit un degré élevé de l'action exprimée par le radical :

(100)	*kuvyíimba*	[ku]préf.inf.[vyíimb]rad. [a]asp.
		'gonfler'
	kuvyiímbagana	'gonfler très fort'
		[ku]préf.inf.[vyíimb]rad.[agan]suff.intens.[a]asp.

Le suffixe intensif duratif [-ang-] signifie que l'action exprimée par le radical atteint un degré élevé et dure dans le temps :

(101)	*kuména*	[ku]préf.inf.[mén]rad. [a]asp.
		'casser'
	kuménanga	[ku]préf.inf.[mén]rad.[ang]suff.intens.dur.[a]asp.
		'concasser'

Le suffixe intensif ampliatif [-agir] traduit que le degré élevé exprimé par le radical correspond à un mouvement répétitif :

(102)	a. *kuniha*	[ku]préf.inf.[nih]rad. [a]asp.
		'gémir'
	b. *kunihagira*	[ku]préf.inf.[nih]rad.[agir]suff.intens.ampl. [a]asp.
		'gémir beaucoup et régulièrement'

Le suffixe intensif péjoratif [-agar-] signifie que l'action exprimée par le radical atteint un degré élevé et véhicule une connotation péjorative.

(103)	a. *gupinda*	[ku]préf.inf.[pind]rad. [a]asp.[14]
		'être bien habillé'
	b. *gupindagara*	[ku]préf.inf. [pind]rad.[agar]suff.intens.péj. [a]asp.
		'être très chic et l'afficher'

[14] Pour *gu* ==> *ku*, cf. loi de Dhal.

Le suffixe réciproque [-an-] signifie 'l'un l'autre'. En kirundi, il transforme le verbe transitif en verbe intransitif.

(104)	*gusoma*	[ku]préf.inf. [som]rad. [a]asp.
		'embrasser'
	gusoma umuntu	'embrasser une personne'
	gusomana	[ku]préf.inf. [som]rad. [an]suff.ass. [a]asp.
		's'embrasser'

3.2.1.3.3. Dérivatèmes divers

La classe des dérivatèmes divers comprend les oppositifs [-uur-] et [-uuk-], l'aptitif [-ik-], le passif [-u-], le statif [-am-] et le transmutatif [-ar-].

L'oppositif sert à dériver des antonymes. On en distingue deux en kirundi : un actif [-uur-] (cf. 105 b) et un passif [-uuk-] (cf. 105 c).

(105)	a. *gupfúka*	[ku]préf.inf.[pfúk]rad. [a]asp.
		'couvrir'
	b. *gupfúkuura*	[ku]préf.inf.[pfúk]rad.[uur]suff.actif [a]asp.
		'découvrir' (enlever la couverture)
	c. *gupfúkuuka*	[ku]préf.inf.[pfúk]rad.[uuk]suff.passif [a]asp.
		'être découvert'

L'aptitif [-ik-] exprime le sens de 'aptitude à être facilement'. En kirundi, on le retrouve par exemple dans :

(106) a. *kurima* [ku]préf.inf.[rim]rad. [a]asp.
'cultiver'
b. *kurimika* [ku]préf.inf.[rim]rad.[ik]suff.apt.[a]asp.
'être cultivable'

Le passif [-u-] exprime que c'est le sujet grammatical qui subit l'action exprimée par le radical.

(107) a. *gusoma* [ku]préf.inf.[som]rad.[a]asp.
'embrasser'
b. *gusomwa* [ku]préf.inf.[som]rad.[u]suff.pass.[a]asp.
'être embrassé'

Le statif [-am-] signifie que l'action exprimée par le radical est achevée et que les conséquences de l'action persistent :

(108) a. *kugonda* [ku]préf.inf.[gond]rad.[a]asp.
'courber'
b. *kugondama* [ku]préf.inf.[gond]rad.[am]suff.stat.[a]asp.
'être courbé'

Le transmutatif [-ar-] exprime le sens 'devenir R[15]'.

(109) a. *ikimúga* 'un infirme'
[i]augm.[ki]préf.cl.[múga]rad.
b. *kumúgara* 'devenir un infirme'
[ku]préf.inf.[múga]rad.[ar]suff.transmut. [a]asp.[16]

---

[15] R = radical.

[16] On notera l'élision de la voyelle /a/ du radical.

3.2.2. MOT-FORME SUBSTANTIF

Le kirundi connaît deux types de substantifs : ceux ayant une base verbo-nominale et ceux ayant une base nominale. Les deux types de substantifs présentent des structures différentes.

Tout substantif du kirundi a cependant un préfixe de classe. À la manière du genre en français, la répartition des seize préfixes de classe paraît en grande partie aléatoire. Nous en fournirons un tableau des fréquences qui devrait informer sur la plus ou moins grande représentation de chaque classe en kirundi.

3.2.2.1. SUBSTANTIF À BASE VERBO-NOMINALE

Le substantif à base verbo-nominale a trois morphèmes obligatoires : un préfixe de classe, un radical et un dérivatif thématique nominal; nous indiquons en (113 a) le verbe infinitif qui partage le même radical que le substantif en (113 b) :

(113) a. *kurima*
[ku]préf.inf. [rim]rad. [a]asp.
'travailler'

b. *ku murimo*[17]
[ku]prép.[u]augm.[mu]préf.cl. [rim]rad. [o]dér.th.n.
'sur' 'travail'
'au travail'

Le substantif à base verbo-nominale du kirundi porte en position finale un morphème appelé « dérivatif thématique nominal » (Houis 1977 : 26-27).

Cette position paradigmatique est susceptible d'être occupée par les dérivatifs thématiques nominaux suivants :

---

[17] On notera l'élision de l'augment ; elle n'a aucune incidence sur le sens.

[-yi] : indique l'agent de l'action exprimée par le radical

umurimyi 'cultivateur' < [-rim-] 'cultiver'

[-i] : indique le résultat de l'action

ibisígi 'héritage' < [-síg-] 'laisser'

[-e] : indique l'état

ubukéne 'pauvreté' < [-ken-] 'avoir besoin de'

[-o] : exprime le résultat, le lieu ou l'instrument de l'action

urugendo 'voyage' < [-gend-] 'marcher'

[-a] : exprime le résultat de l'action exprimée par le radical

umurírna 'champ' < [-rim-] 'cultiver'

[-u] : indique le résultat de l'action

umunigu 'rétrécissement' < [-nig-] 'serrer la gorge'

Le substantif à base verbo-nominale du kirundi peut, en plus de ces trois morphèmes de base, accueillir l'un ou l'autre des quatre morphèmes facultatifs suivants : un augment (114 a-c)[18], un suffixe de dérivation (114 a), un morphème complément (114 b) et un morphème réfléchi (114 c) -mis en gras :

---

18 Certains substantifs ne possèdent pas d'augment : ex. *soókuru* 'grand-père', *daawé* 'papa', etc.

(114) a. ***u**kurimira*
[**u**]augm. [ku]préf.cl. [rim]rad. [ir]suff. [a]dér.th.n.
'action de cultiver pour'

b. *uku**b**irima*
[u]augm. [ku]préf.cl. [**bi**]compl.[rim]rad. [a]dér.th.n.
'action de les cultiver'

c. *ukw**íi**rima*[19]
[u]augm. [ku]préf.cl. [**íi**]réfl. [rim]rad. [a]dér.th.n.
'action de se couper avec la houe'

Des règles complexes, que nous ignorons ici, régissent la combinatoire des sept morphèmes constitutifs du substantif à base verbo-nominale.

Signalons une forme particulière du substantif à base verbo-nominale. Sa particularité réside en la présence d'une base complexe à laquelle s'adjoignent le préfixe de classe et l'augment tel qu'illustré en (115 b), forme que l'on peut rapprocher de (115 a) :

(115) a. *abantu* *bazoorima*
[ba]préf.v.suj. [zoo]préd.v. [rim]rad. [a]asp.
'les personnes' 'cultiveront'

b. *abazóoríma*
[a]augm. [ba]préf.cl. [zoo]fut. [rim]rad. [a]dér.th.n.
'ceux qui cultiveront'

---

[19] Où [w] < /u/ à la frontière de morphème (cf. consonantisation des voyelles au § 3.1.4.1.2).

Le mot-forme verbal en (115 a) permet d'obtenir le substantif en (115 b) en lui préfixant un augment; il donne ainsi cours à un substantif à base verbo-nominale de structure particulière. Même si ce substantif constitue un dérivé prévisible, nous le retenons dans le vocabulaire de base du kirundi écrit; notre inventaire des substantifs étant exhaustif.

3.2.2.2. SUBSTANTIF À BASE NOMINALE

Le substantif du kirundi à base nominale met généralement en jeu trois morphèmes : l'augment, le préfixe de classe et le radical. L'augment est souvent élidé sans aucune incidence sur le sens, si bien que minimalement, le substantif à base nominale se compose de deux morphèmes : un préfixe de classe et un radical :

(116) *ku muntu*
[ku]prép. [u]augm. [mu]préf.cl. [ntu]rad.
'sur' 'personne humaine'
'sur une personne'

Le tableau 12 (p.110) fournit les seize préfixes de classe, les seize préfixes pronominaux et les seize préfixes compléments correspondants.

Quant aux radicaux des substantifs à base nominale, ils constituent pour leur part un ensemble ouvert.

Signalons que les substantifs composés sont comptés parmi les substantifs à base nominale. Il en est de même pour les substantifs qui partagent leur radical avec un adjectif.

3.2.3. MOT-FORME ADJECTIF

Le mot-forme adjectif du kirundi contient deux morphèmes : un préfixe adjectival, identique au préfixe de classe du substantif avec lequel l'adjectif s'accorde, et un radical - rappelons que «augm.» doit être lu «(voyelle) augment» :

(117) *ikiraato*$_{n.}$ *kiníni*$_{adj.}$
'un soulier' 'grand'
[i]$_{augm.}$ [ki]$_{préf.cl.}$ [raato]$_{rad.}$ [ki]$_{préf.adj.}$ [níni]$_{rad.}$
'un grand soulier'

ibiraato$_{n.}$ *biníni*$_{adj.}$
'des souliers' 'grands'
[i]$_{augm.}$ [bi]$_{préf.cl.}$ [raato]$_{rad.}$ [bi]$_{préf.adj.}$ [níni]$_{rad.}$

On déduit à partir de ces exemples, que le radical adjectival [-níni] 'grand' permet la formation de seize mots-formes adjectifs, c'est-à-dire d'autant de mots-formes par radical adjectival qu'il y a de préfixes de classe.

Le kirundi connaît un autre type d'adjectifs. Il s'agit de mots-formes invariables qui déterminent les noms. Ils ont surtout cours dans le discours politique et forment une classe ouverte. On peut les illustrer avec le mot-forme *kaminúuza* 'de très haut niveau' qui reste invariable lorsque varient les substantifs qu'il détermine :

(118) *ishuúre kaminúuza* 'université'
'école' 'de très haut niveau'

***ama**shuúre kaminúuza* 'universités'
'écoles' 'de très haut niveau'

Ces adjectifs invariables sont importants dans le vocabulaire de base. Ils qualifient des notions nouvelles comme « l'école ».

3.3. MOTS-FORMES GRAMMATICAUX

3.3.1. MOTS-FORMES GRAMMATICAUX FLÉCHIS

Les mots-formes grammaticaux fléchis connaissent des emplois déterminatifs et pronominaux. En emploi déterminatif, ils suivent un substantif sans s'y agglutiner; les auteurs s'entendent en effet pour considérer qu'ils sont à l'extérieur des limites formelles du substantif.

Nous utilisons la notion de «déterminatif» dans le sens de Dubois *et al.* (1994 : 140) selon lequel les déterminants constituent *«une classe de morphèmes grammaticaux dépendants en nombre du nom qu'ils spécifient. »*

Dans ce sens, les articles, les possessifs, les démonstratifs, etc. sont des déterminants en français.

Quoique intéressante, cette classification fonctionnelle présente une limite en lexicométrie : elle ne tient pas compte de l'homographie et des formes courtes ~ formes longues des mots-formes.

Nous distinguons donc à l'intérieur des mots-formes grammaticaux ceux qui sont homographes et ceux qui sont hétérographes. Ces derniers regroupent entre autres les formes courtes ~ formes longues.

3.3.1.1. MOTS-FORMES GRAMMATICAUX FLÉCHIS HOMOGRAPHES

Les mots-formes grammaticaux fléchis homographes regroupent des démonstratifs, des allocutifs et des numéraux. Nous en fournissons un inventaire.

3.3.1.1.1. Démonstratifs

Le kirundi utilise cinq démonstratifs en emploi déterminatif dont le choix dépend de la distance qui sépare le locuteur de l'objet désigné. Nous illustrons ces cinq démonstratifs (en position pré-nominale) en combinaison avec le substantif à

base nominale de la classe 2 *abantu* 'personne humaine' dont l'augment [a] s'élide[20] :

(119)	*aba*	*bantu*	*bararima*	'ces personnes-ci cultivent'
	bárya	*bantu*	*bararima*	'ces personnes-là cultivent' (proches)
	báriíya	*bantu*	*bararima*	'ces personnes-là cultivent' (éloignées)
	báa	*bantu*	*bararima*	'ces personnes-là cultivent' (anaphorique)
	báno	*bantu*	*bararima*	'ces personnes en question cultivent' (anaphorique)

Les substantifs se répartissant en seize classes et chacune pouvant être précédée d'un démonstratif, les mots-formes démonstratifs du kirundi se chiffrent à 80.

Tel qu'illustré en (120), les démonstratifs connaissent des emplois pronominaux :

(120)	*aba*	*bararima*	'ceux-ci / celles-ci cultivent'
	bárya	*bararima*	'ceux-là / celles-là cultivent' (proches)
	báriíya	*bararima*	'ceux-là / celles-là cultivent' (éloignées)
	báa	*bararima*	'ceux-là / celles-là cultivent' (anaphorique)
	báno	*bararima*	'ceux-là / celles- là en question cultivent' (anaphorique)

À toute classe de substantif, on peut donc substituer cinq mots-formes démonstratifs, ce qui fait 80 démonstratifs en emploi pronominal pour les seize classes.

Comme nous n'avons pas désambiguïsé les démonstratifs, nous ne distinguons pas au niveau de leurs fréquences les emplois déterminatifs des emplois pronominaux.

---

[20] Tous les augments du kirundi s'élident après un démonstratif.

3.3.1.1.2. Indéfinis

La catégorie des mots-formes grammaticaux indéfinis en emploi déterminatif met en jeu trois morphèmes : *-óóse* 'tous' ~ *-óómpi* 'tous ensemble', *-mwé* 'certains' et *-ndi* 'autre'. Nous l'illustrons avec le substantif à base nominale *abantu* 'personne humaine'de la classe 2 :

(121) *abantu bóóse bazooza* 'toutes les personnes viendront'
abantu bóómpi bazooza 'toutes les personnes ensemble viendront'
abantu báandi bazooza 'les autres personnes viendront'
abantu bamwé bazooza 'certaines personnes viendront'

Ainsi, pour chaque substantif considéré dans une classe, il y a quatre mots-formes indéfinis possibles. Pour les seize classes, les mots-formes indéfinis en emploi déterminatif se chiffrent à 64.

Comme illustré en (121), les indéfinis connaissent des emplois pronominaux :

(122) *bóóse bazooza* 'tous / toutes viendront'
bóómpi bazooza 'tous / toutes ensemble viendront'
abaándi[21] *bazooza* 'les autres viendront'
bamwé bazooza 'certains viendront'

Pour chaque substantif considéré dans une classe, il y a quatre mots-formes indéfinis possibles en emploi pronominal, ce qui fait 64 mots-formes.

Mais comme nous ne désambiguïsons pas les mots-formes indéfinis selon qu'ils connaissent un emploi déterminatif ou pronominal, nous considérons qu'au total, les mots-formes indéfinis se chiffrent, pour les 16 classes, à 64.

---

[21] La voyelle /a/ est idiosyncrasique; le mot-forme régulier serait **báandi* 'autres'.

3.3.1.1.3. Interrogatifs

Les mots-formes interrogatifs en emploi déterminatif mettent en jeu quatre morphèmes : *-ndé* 'qui', *-kí* 'quoi' et *-ngáaheé* 'combien', *-hé* 'quel' :

(123)	*abantu*	*bandé ?*	'qui (sont) ces personnes ?'
	abantu	*bakí ?*	'quels types de personnes ?'
	abantu	*bangáahé ?*	'combien de personnes ?'
	abantu	*báahé ?*	'quelles personnes?'

Les interrogatifs construits sur-*ngáahé* 'combien' n'existent qu'avec les substantifs exprimant le pluriel, soit pour les classes : 2, 4, 6, 10, 14, 15 et 16, ce qui donne sept mots-formes pour 'combien'.

Ceux construits sur *-ndé* 'qui' et *-kí* 'quoi' et *-hé* 'quel' sont réguliers pour les seize classes, ce qui fournit 48 mots-formes interrogatifs pour ces trois morphèmes interrogatifs. Au total, on compte donc 55 mots-formes interrogatifs.

Comme illustré en (124), les interrogatifs connaissent des emplois pronominaux :

(124)	*bandé ?*	'qui ?'
	abakií ?	'lesquels [types de personnes] ?'
	bangáahé ?	'combien [de personnes] ?'
	báahé ?	'lesquels [les personnes] ?'

Les interrogatifs en emploi pronominal construits sur *-ngáahé* 'combien' n'existent que pour les substantifs exprimant le pluriel; il s'agit donc de sept mots-formes.

Les interrogatifs construits sur *-ndé* 'qui' et *-kí* 'quoi' et *-hé* 'quel' sont réguliers pour toutes les classes, ce qui donne 48 interrogatifs. Le total des interrogatifs en emploi pronominal s'élève donc à 55. En nous fondant sur l'identité formelle, nous ne comptons au total que 55 mots-formes interrogatifs.

3.3.1.1.4. Numéraux cardinaux

Le kirundi connaît six numéraux (de 1 à 6) cardinaux en emploi déterminatif (cf. 125)[22]. Les six numéraux mettent en jeu les morphèmes suivants : *-mwé* 'un', *-biri* 'deux', *-tatu* 'trois', *-né* 'quatre', *-taanu* 'cinq', *-tandátu* 'six'. Nous les illustrons avec les substantifs à base nominale *umuntu* 'personne humaine' de la classe 1 et *abantu* 'personnes humaines' de la classe 2 :

(125) a. *umuntu umwé ararima* 'une personne cultive '
b. *abantu babiri bararima* 'deux personnes cultivent'
b. *abantu batatu bararima* 'trois personnes cultivent'
c. *abantu bané bararima* 'quatre personnes cultivent'
d. *abantu bataanu bararima* 'cinq personnes cultivent'
f. *abantu batandátu bararima* 'six personnes cultivent'

Ainsi, pour un substantif considéré dans une classe, on a six mots-formes numéraux possibles; cela revient pour les seize classes à 96 mots-formes.

Comme illustré en (126), les numéraux connaissent des emplois pronominaux :

(126) a. *umwé ararima* 'une [personne] cultive'
b. *babiri ararima* 'deux [personnes] cultivent'
b. *batatu ararima* 'trois [personnes] cultivent'
c. *bané ararima* 'quatre [personnes] cultivent'
d. *bataanu ararima* 'cinq [personnes] cultivent'
f. *batandátu ararima* 'six [personnes] cultivent'

---

[22] À partir de 7, les numéraux ont la structure du substantif à base nominale ; ils ne s'accordent pas alors avec le substantif mais commandent l'accord comme tous les substantifs.

Ainsi, pour un substantif considéré dans une classe, on a six mots-formes numéraux possibles en emploi pronominal; cela revient pour les seize classes à 96 mots-formes.

En nous fondant uniquement sur le critère formel, nous chiffrons donc les mots-formes grammaticaux numéraux cardinaux à 96.

3.3.1.1.5. Locatifs

Les mots-formes locatifs sont postposés au verbe. Ils correspondent à quatre mots-formes : *kó* 'sur quelque chose', *mwó* 'dans quelque chose', *hó* 'quelque part' et *yó* 'dans un endroit'. Selon Meeussen (1959 : 103), ils constituent un emploi particulier du sustitutif bref (cf. §3.3.1.3.2.). Nous les illustrons en (127) :

(127)	*agiiye kw'ipikipiki*	'il part à moto'
	agiiye ***k ó***	'il part dessus'
	agiiye mu ndeége	'il part dans un avion'
	agiiye ***m w ó***	'il part dedans [dans un avion]'
	agiiye i Bujumbura	'il part à Bujumbura'
	agiiye ***yó***	'il y part [à Bujumbura]
	agiiye ku butegetsi	'il va au pouvoir'
	agiiye ***h o***	'il y va [au pouvoir]'

3.3.1.1.6. Interpellatifs

Les interpellatifs sont des termes de la langue utilisés dans la communication directe pour interpeller l'interlocuteur auquel on s'adresse (Dubois *et al.* 1994 : 45).

Le kirundi dispose de quatre interpellatifs pour les deuxièmes personnes du singulier et du pluriel, distinguant ainsi les cas où l'on a un ou plusieurs interlocuteurs :

(128)	*eéwe*	'eh toi!'	*eémwe*	'eh vous!'
	waa	'holà toi!'	*mwaa*	'holà vous!

3.3.1.1.7. Adverbes

Certains adverbes du kirundi sont flexionnels. Ils marquent notamment l'interrogation, la comparaison, la manière, etc. Nous les illustrons ci-dessous :

(129)	*akora gúte ?*	'il travaille comment ?'
	bakora báte ?	'ils travaille comment ?
	akora wéenyéne	'il travaille seul'
	bakora bóónyéne	'ils travaillent seuls'

3.3.1.2. MOTS-FORMES GRAMMATICAUX FLÉCHIS HÉTÉROGRAPHES

Les mots-formes grammaticaux hétérographes sont des possessifs et des connectifs. Nous les décrivons dans cet ordre.

3.3.1.2.1. Possessifs

Les possessifs connaissent des emplois déterminatifs. Nous illustrons ce fait avec le substantif à base nominale *abantu* 'personne humaine' de la classe 2 :

(130)	*abantu*	*banje*	*barakóra*	'mes personnes travaillent'
	abantu	*baawe*	*barakóra*	'tes personnes travaillent'
	abantu	*biíwe*	*barakóra*	'ses personnes travaillent'
	abantu	*báacu*	*barakóra*	'nos personnes travaillent'
	abantu	*báanyu*	*barakóra*	'vos personnes travaillent'
	abantu	*báabo*	*barakóra*	'leurs personnes travaillent'

Vu les seize classes de substantifs qu'ils peuvent suivre, les mots-formes possessifs se chiffrent à 96.

Le possessif connaît des emplois pronominaux. Nous en fournissons quelques exemples en (131) :

(131)	*abáanjé barakóra*	'les miens' / 'les miennes travaillent '
	abáawé barakóra	'les tiens' / 'les tiennes travaillent'
	abiíwé barakóra	'les siens' / 'les siennes travaillent'
	abaácu barakóra	'les nôtres travaillent'
	abaányu barakóra	'les vôtres travaillent'
	abaábo barakóra	'les leurs travaillent'

Ainsi donc, pour les seize classes, les mots-formes possessifs en emploi pronominal se chiffrent à 96. La voyelle initiale permet les distinguer de ceux en emploi déterminatif.

3.3.1.2.3. Connectifs

Les connectifs connaissent des emplois déterminatifs. Ils se présentent obligatoirement entre deux substantifs qu'ils lient, tel qu'illustré en (132) ou entre un substantif et une préposition.

Ceux qui lient deux substantifs sont construits sur le radical [-a] (cf. 132 a, 132 b) tandis que ceux qui lient un substantif et une préposition sont construits sur le radical [-ó] (cf. 132 c).

Les connectifs s'accordent avec le substantif, qu'ils suivent conformément aux accords résumés dans le tableau 12 (p. 110).

(132) a. *abantu* baa *Yóhaáni*
[ba]préf.pron. [a]rad.
'les personnes de Jean'

b. *umugóre* wa *Yóhaáni*
[u]préf.pron. [a]rad.[23]
'la femme de Jean'

c. *inzu* *zó* *mu* *Burundi*
[zi]préf.pron. [ó]rad.
'les maisons' 'de' 'dans' 'Burundi'
'les maisons du Burundi'

Pour les seize classes des substantifs du kirundi, on compte seize mots-formes connectifs construits sur [-a] et 16 autres construits sur [-ó] soit 32 mots-formes connectifs.

Les connectifs connaissent des emplois pronominaux tel qu'illustré en (133)

(133) a. *abaa Yóhaáni* 'ceux de Jean'
b. *uwa Yóhaáni* 'celui de Jean'
c. *izó mu Buruúndi* 'celles du Burundi'

L'emploi pronominal se traduit donc par la présence d'une voyelle identique à l'augment du substantif (cf. exemple (132)).

Ainsi, aux 16 classes des substantifs du kirundi répondent seize connectifs en emploi pronominal. Nous avons donc 32 connectifs soit 16 construits sur [-a] et 16 construits sur [-ó]. Les connectifs construits sur [-a] sont utilisés en position prénominale, tandis que ceux construits sur [-ó] sont utilisés en position pré-prépositionnelle.

---

[23] Il y a consonantisation de la voyelle /u/.

3.3.1.3. MOTS-FORMES COURTS versus *LES MOTS-FORMES LONGS*

Les pronoms allocutifs et substitutifs ont une forme brève et une forme longue. Il en est de même pour les prépositions.

Les formes longues sont emphatiques par rapport aux formes courtes. Cette différence qui est de nature pragmatico-sémantique, nous amène à rejeter l'hypothèse d'allomorphie. Nous compterons donc chaque forme courte pour une unité et chaque forme longue pour une autre.

3.3.1.3.1. Pronoms allocutifs

Les pronoms allocutifs représentent les participants au discours. Ils se présentent sous une forme courte ou une forme longue, emphatique; tel qu'illustré en (134) :

(134) *je* *Yohaáni* *nzoorima*
'moi Jean je cultiverai'

jeewé *Yohaáni* *nzoorima*
'moi Jean je cultiverai' [+ emphase]

Le kirundi dispose de quatre mots-formes allocutifs courts et de quatre longs, emphatiques. Nous les présentons dans le tableau suivant :

Forme courte	*Glose*	*Forme longue*	*Glose*
je	'moi'	*jeewé*	'moi' [+ emphase]
we	'toi'	*wewé*	'toi' [+ emphase]
twe	'nous'	*tweebwé*	'nous' [+ emphase]
mwe	'vous'	*mweebwé*	'vous' [+ emphase]

Tableau 14 - *Pronoms allocutifs du kirundi*

Tenant compte du seul critère formel, nous comptons quatre mots-formes allocutifs courts et quatre longs.

3.3.1.3.2. Pronoms substitutifs

Les pronoms substitutifs du kirundi se rencontrent eux aussi sous une forme courte (136 a) et une forme longue (136 b), emphatique :

(136)	a.	*abantu*	*bó*	*bazoorima*
		'les gens	eux	cultiveront'
	b.	*abantu*	*bóobó*	*bazoorima*
		'les gens	eux	cultiveront' [+ emphase]

À chacune des 16 classes nominales correspond un substitutif, c'est dire qu'ils se présentent sous 32 formes, soit seize formes courtes et seize formes longues.

Signalons une homonymie entre les locatifs et les substitutifs courts. Les locatifs sont, rappelons-le, quatre unités : *hó* 'quelque part', *kó* 'sur quelque chose' et *yó* 'dans un endroit' et *mwó* 'dans'. À l'exception de *mwó*, les locatifs sont homographes des pronoms substitutifs courts comme l'illustrent les exemples en (138) :

(138) *ukurima* *kó* *kuraruhisha*
'le fait de cultiver' 'lui' 'il fatigue'
'faire de l'agriculture ça fatigue'

imódoka *yó* *irazimvye*
'voiture' 'elle' 'est chère'
'la voiture, elle, est chère'

ahantu *hó* *harakanye*
'endroit' 'dont il est question' 'est froid'
'l'endroit dont il est question est froid'

Nous avons établi les fréquences de ces mots-formes locatifs et substitutifs courts à partir de concordances fournies par *WordCruncher*. Il nous faut en effet rappeler ici que ces mots-formes grammaticaux n'ont pas été désambiguïsés lors de la saisie.

Quant aux formes longues, elles ne posent pas de problème d'homographie; leurs fréquences sont celles fournies par *WordCruncher* et sont consignées dans l'index.

3.3.1.3.4. Numéraux ordinaux

Le mot-forme ordinal est une combinaison du connectif et du numéral; ex. :

(139) a. *abantu* *baa kábiri*
'les personnes humaines' 'deuxièmes'
'les deuxièmes personnes humaines'

b. *ahantu* *haa kábiri*
'endroit' 'deuxième'
'le deuxième endroit'

Fondamentalement, on ne peut dénombrer les mots-formes ordinaux, qui sont en nombre illimité. Aux fins de notre étude, nous ignorerons la sous-catégorie morphosyntaxique des ordinaux. Nous scinderons les mots-formes ordinaux en leurs composants (connectifs et numéraux) et ce sont ces composants qui feront l'objet de comptages.

Le numéral ordinal connaît des emplois pronominaux tel qu'illustré en (140) :

(140) a. *abaa kábiri* 'les deuxièmes ' (personnes humaines)
b. *ahaa kábiri* 'le deuxième' (endroit)

On le voit, les mots-formes ordinaux en emploi pronominal sont une combinaison des connectifs et des numéraux.

3.3.1.3.5. Adverbes

Comme on le constatera plus loin, la plupart des adverbes du kirundi sont non flexionnels. Notre corpus a cependant permis d'identifier quelques adverbes flexionnels. Il s'agit des unités suivantes :

hamwé	'ensemble'
rwóóse	'très'
gúte ?	'comment ?'
gúrtyo ?	'comme cela / 'ainsi'
kumwé	'ensemble' cl.15
nkó	'comme'
iwaácu	'chez nous'
iwaábo	'chez eux'
bóonyéne	'eux seul' (cl.2)
rimwé na rimwé	'des fois' (cl.11)
wéenyéne	'lui seul' (cl.1)
iwé	'chez lui'
yóonyéne	'lui seul' (cl.4,5,9)

3.3.2. MOTS-FORMES GRAMMATICAUX NON FLÉCHIS

Les mots-formes grammaticaux non fléchis regroupent en kirundi les prépositions, les prédicatifs nominaux, les conjonctions, les adverbes, les interjections et les onomatopées. Nous les présentons dans cet ordre en nous servant essentiellement de l'inventaire réalisé par Rodegem (1967 : 79-94) et des éléments de notre corpus.

3.3.2.1. PRÉDICATIFS NOMINAUX

La classe des prédicatifs nominaux compte quatre mots-formes : *ni, si, ntaa, atáa,* illustrés en (141) :

(141) a. *ni* *umuntu* 'c'est une personne'
'c'est' 'personne humaine'

b. *si* *umuntu* 'ce n'est pas une personne'
'ce n'est pas' 'personne humaine'

c. *ntaa* *muntu* 'pas de personne'[24]
'pas de' 'personne humaine'

d. *atáa* *muntu* 'sans personne humaine'
'sans' 'personne humaine

e. *nyaa* *muntu* 'personne dont il est question'
'dont il est question' 'personne humaine'

3.3.2.2. PRÉPOSITIONS

La classe des prépositions de notre corpus est constituée de huit mots-formes[25] :

(142)	a. *mu*	'dans'		d. *murí*	'dans'
	b. *ku*	'sur'		e. *kurí*	'sur'
	c. *i*	'à'		f. *kuva*	'depuis'
	g. *kugeza*	'jusqu'à'		i. *nayó*	'quant à'

3.3.2.3. LES CONJONCTIONS

On appelle conjonction, un mots-outil qui assure un lien syntaxique entre des phrases ou des membres de phrases. Piot (1988 : 3-18) distingue deux types de

---

[24] On notera l'élision de l'augment [u-] en (141 c, d).

[25] L'étude de la distribution des formes en (142 a, d) et (142 b, e) reste à faire.

conjonctions : des conjonctions de coordination et des conjonctions de subordination. Nous adoptons cette typologie.

3.3.2.3.1. Conjonctions de coordination

Les conjonctions de coordination relient des phrases ou des membres de phrases. Soit l'énoncé en (143) :

(143) a. *Yohaáni* ***na*** *Péetéro bararima* 'Jean et Pierre cultivent'
b. *Yohaáni ararima* ***canké*** *arasoma ?* 'Jean cultive ou il lit'

Commme on peut le constater, la conjonction *na* 'et' en (143 a) relie deux membres de phrases - deux substantifs -. En (143 b) la conjonction *canké* 'ou bien' relie deux phrases.

Nous dressons ci-dessous la liste de conjonctions de coordination que nous avons identifiées dans notre corpus :

(144)	*na*	'et'	*kiiburé*	'ou bien'
	kaándi	'ensuite'	*canké*	'ou bien'
	mugábo	'mais'	*hanyuma*	'après'
	bé	'et'	*nkakó*	'de fait'

3.3.2.3.2. Conjonctions de subordination

Les conjonctions de subordination relient des phrases dont l'une est indépendante de l'autre (Piot 1988 : 3-18). En kirundi, le paradigme des conjonctions de subordination est assez riche comme en témoigne l'inventaire suivant, tiré de notre corpus :

(145)	*aríko*	'mais'	*máze*	'dès lors que'
	mugábo	'mais'	*náahó*	'même si'
	kanásinda	'du reste'	*nakó*	'ou plutôt'

ngo	'que'[26]	*nkakó*	'de ce fait même'
ko	'puisque'	*nkaaswe*	'à plus forte raison'
kó	'que'[27]	*nkeeka*	'peut-être'
kubéera	'à cause de'	*reeró*	'donc'
kugíra	'pour que'	*rího*	'plutôt'
kukó	'parce que '	*nkanátangaaye*	'*a fortiori*'

3.3.2.4. ADVERBES

La liste des adverbes est fournie. En voici quelques-uns tirés de notre corpus :

(146)	*buhóro*	'lentement'	*hiíno*	'de ce côté-ci'
	caane	'beaucoup'	*hiíyo*	'de ce côté-là'
	eegó	'oui'	*imbere*	'devant'
	ejó	'hier / demain'	*inyuma*	'derrière'
	epfó	'en contre-bas'	*inzé*	'dehors'
	erega	'sûrement'	*keéra*	'jadis'/ 'autrefois'
	kure	'loin'	*haáfi*	'pas loin'
	haákuno	'en-deçà'	*muusí*	'sous'
	haákurya	'au delà'	*náabí*	'mal'
	haasí	'par terre'	*namba*	'nullement, du tout'
	hambavu	'à côté'	*néezá*	'bien'
	hambere	'auparavant'	*níngoga*	'rapidement'
	handihato	'tout à l'heure'	*nkána*	'exprès'
	hanyuma	'ensuite'	*nooné*	'maintenant'
	hanzé	'dehors'	*oya*	'non'

---

26 *Ngo* 'que' introduit le discours indirect :

(i) *avuga ngo azoorima* *'il dit qu'il cultivera'*
'il dit' 'que' 'il cultivera'

27 On notera l'homographie de *kó* 'puisque' et *kó* 'que'.

haruguru	'au-dessus'	*ryáarí*	'quand'
heejuru	'en haut'	*ubu*	'maintenant'
hé	'où'	*vubá*	'vite'

3.3.2.5. INTERJECTIONS

Dubois *et al.* (1994) définissent l'interjection comme un mot invariable exprimant une relation affective vive, comme *zut!* en français. Le corpus contient les interjections suivantes :

(147)	*yooo*	oh!	*ohooo*	oh!
	eee	eh!	*yeee*	oui
	hiii	eh!	*ehe*	eh!
	ashiii	ouf!		

3.3.2.6. ONOMATOPÉES ET IDÉOPHONES

Tournier (1985 : 41) définit l'onomatopée comme une imitation phonique reposant sur une analogie entre la forme et la chose nommée; /kokoriko/ est une onomatopée du français. En kirundi, les onomatopées suivantes suggèrent le son ou le bruit :

(148)	*diri diri diri*	des pas
	ndee ndee ndee	d'une cloche
	dobwe	de la goutte d'eau
	pwa pwa pwa	d'une gifle
	kwe kwe	d'un rire
	ruru ruru ruru	du feu
	mbomboro	du métal
	zye	d'une scie

Quant aux idéophones, ce sont des mots invariables qui impliquent une association motivée entre une image acoustique à une notion (Tournier 1985 : 139). Notre corpus en contient deux. Nous indiquons entre parenthèses la notion exprimée.

(149) *pe* 'tout à fait' (perfection)
de 'tout à fait' (joie)

Telles sont les classes de mots-formes du kirundi. Nous avons fait une brève description de chacune d'elles en insistant particulièrement sur les mots-formes lexicaux. Nous avons distingué parmi eux les verbes, les substantifs (à base verbo-nominale et à base nominale) et les adjectifs. Rappelons que le découpage morphologique opéré par l'analyseur morphologique aboutira à un index des fréquences des différents morphèmes tel qu'illustré par l'annexe 2.

Parmi les mots-formes grammaticaux, nous avons distingué les fléchis des non fléchis. Parmi les premiers, nous avons distingué trois types :

- les mots-formes grammaticaux fléchis homographes : les démonstratifs, les indéfinis, les numéraux, les interrogatifs, les locatifs et les interpellatifs. Considérant qu'il s'agit d'un cas d'homographie fonctionnelle où un même item connaît un emploi déterminatif ou pronominal, nous ne les avons pas désambiguïsés.

- les mots-formes grammaticaux fléchis hétérographes; on a deux cas :

  a. les mots-formes diffèrent par une voyelle initiale : c'est le cas des possessifs et des connectifs;

  b. les mots-formes qui ont une forme courte et une forme longue : c'est le cas des allocutifs et des substitutifs.

Parmi les mots-formes grammaticaux non fléchis, nous avons distingué six classes : les prédicatifs nominaux, les prépositions, les conjonctions, les adverbes, les interjections et les onomatopées.

Quelle est la représentativité de chacune de ces sous-catégories morphosyntaxiques dans le corpus ? À combien de vocables correspondent-elles ? Avec quelles occurrences ? Et quelle est la part des catégories lexicales ? Quels sont leurs morphèmes les plus fréquents ? Nous abordons toutes ces questions et bien d'autres au chapitre 4 où nous présentons les résultats de notre recherche.

CHAPITRE 4

LE VOCABULAIRE DE BASE DU KIRUNDI ÉCRIT

Le nombre d'unités lexicales constitutives d'un vocabulaire de base est variable d'une langue à l'autre et même d'une recherche à l'autre selon les critères que l'on prend en compte.

Gougenheim *et al.* (1964) ont retenu les vocables apparaissant dans cinq sous-corpus. La fréquence des vocables retenus devait être supérieure ou égale à 29. Ils ont ainsi sélectionné 1 475 mots dont 1 222 mots lexicaux et 253 mots grammaticaux. Par contraste, Vander Beke (1935) avait retenu des mots de fréquence et de répartition supérieures ou égales à 5, ce qui a fourni 6 067 mots à son *French word book on a count of 400 000 running words.*

Juilland *et al.* (1970) a de son côté retenu 5 000 vocables satisfaisant aux critères suivants : une fréquence minimale de 5, une dispersion minimale de 5,83 et un coefficient d'usage minimal de 3.

Dans le *Dictionnaire de fréquence des mots du français parlé au Québec*, dictionnaire qui n'est pas de type fondamental mais de type exhaustif, Beauchemin *et al.* (1992) utilisent les critères de dispersion, d'usage et d'écart réduit entre les fréquences d'un même vocable dans plusieurs tranches.

Tout en voulant dégager un noyau fondamental du vocabulaire du kirundi écrit, notre étude veut aussi aboutir à un inventaire exhaustif des vocables du corpus.

Nous nous servons de trois critères dans notre recherche : la fréquence, la dispersion et l'usage. De ces trois critères, l'usage constitue le critère qui permet de répondre au mieux à la préoccupation didactique (Beauchemin *et al.* 1992 : XXXIV). Il permet de tenir compte, pour un vocable, à la fois de sa fréquence et de sa dispersion.

La présentation des résultats privilégiera donc le critère de l'usage. Nous fournirons des index exhaustifs des vocables du corpus selon l'usage, la fréquence, la dispersion et complémentairement selon un ordre alphabétique.

En utilisant le critère d'usage, nous distinguons trois tranches dans le vocabulaire de base. La première tranche regroupe des vocables dont l'indice d'usage (U) est supérieur ou égal à 3. La deuxième regroupe des vocables dont U est compris entre 3 et 0. La troisième tranche comprend des vocables dont U est inférieur à 0; dans ce dernier cas, ce sont presque tous des hapax.

Nous présentons donc les résultats en tenant compte de ces trois grandes tranches de vocables. Nous présentons d'abord des résultats généraux sur le corpus. Nous analysons ensuite la structure de chaque tranche de vocables selon les différentes catégories grammaticales qui la dominent. Nous vérifions nos hypothèses de recherche sur la hiérarchie et la stabilité des fréquences des catégories grammaticales. Nous verrons ensuite comment se composent, en termes de catégories grammaticales, les 50 vocables les plus fréquents en kirundi écrit. Nous établirons une comparaison avec les données sur le français et l'anglais. Avant de clore le chapitre, nous jetterons un regard sur les noms propres et les abréviations que l'on retrouve dans le corpus analysé.

1. RÉSULTATS QUANTITATIFS GÉNÉRAUX

Les 103 561 mots-fornes de notre corpus correspondent à 4 025 vocables appartenant à diverses catégories grammaticales. Nous les avons répartis en trois tranches selon l'usage (U) et fournissons pour chaque catégorie des données sur sa représentativité dans l'ensemble du corpus. Le tableau 16 résume l'essentiel des données générales sur le corpus.

	V				*N*			
Classes de V	U ≥ *3*	*3*<U≥ *0*	U< *0*	***Total***	U≥ *3*	*3*< U≥ *0*	U< *0*	***Total***
Verbaux	406	327	71	**804**	24 590	971	90	**25 651**
Nominaux à base nominale	349	578	238	**1 165**	18 047	1 542	304	**19 893**
Nominaux à base verbo-nominale	189	766	416	**1 371**	5 003	1 506	473	**6 982**
Adjectivaux	24	5	48	**77**	1 330	28	191	**1 549**
Grammaticaux	320	229	59	**608**	41 359	602	66	**42 027**
Total	**1 288**	**1 905**	**832**	**4 025**	**90 334**	**4 649**	**1 124**	**96 102**
% de *V*	**32,00**	**47,3**	**20,6**	**100%**				
% de *N*					**87,22**	**4,48**	**1,08**	**92,79**

Tableau 16 - *Les classes de V selon U et fréquence des mots-formes*

Plusieurs conclusions se dégagent de ce tableau :

- Les différentes classes de vocables couvrent 92,7% des mots-formes de tout le corpus. Les 7% restants sont couverts par des noms propres (noms de personnes, de lieu, de mois), des abréviations, des chiffres non réécrits et des mots en langues étrangères - surtout en français -).

- Des trois tranches de vocables, celle dont U est supérieure ou égale à 3 (1 288 vocables) représente 32% du vocabulaire et 87% des mots-formes de tout le corpus. Cela revient aussi à 94% des occurrences de tous les vocables. Nous estimons ces pourcentages assez élevés pour dire qu'ils constituent le noyau du vocabulaire de base du kirundi écrit. Nous privilégierons donc les vocables de cette classe tout au long de ce chapitre.

- Le nombre de vocables dont U est compris entre 0 et 3 est élevé (1 905 vocables). Ils forment la tranche de vocables la plus fournie. Ils représentent 47% du vocabulaire total mais seulement 4% des mots-formes du corpus.

- Les vocables dont U est inférieur à 0 sont au nombre de 832; ils représentent 20% du vocabulaire du corpus et 1% des mots-formes.

- Dans l'ensemble, notre liste est dominée par les vocables verbaux (804), les nominaux à base verbo-nominale (1 371), les nominaux à base nominale (1 165) et les vocables grammaticaux (608). Les vocables adjectivaux sont peu nombreux.

Nous concentrerons donc notre attention sur les quatre premières catégories de vocables les plus fournies.

Rappelons que les substantifs, adjectifs et verbes sont reconnus par l'analyseur morphologique sur la base des morphèmes flexionnels qui les caractérisent et du dictionnaire des radicaux.

Pour des raisons de commodité, nous présentons les vocables verbaux sous forme de radicaux. Ceci permet de faire l'économie de nombreux changements morphophonologiques typiquement localisés entre le préfixe personnel et l'initiale vocalique ainsi qu'entre la consonne finale du radical et le morphème [-ye] de l'aspect accompli tel qu'illustré ci-dessous :

(150) *twiipfuuza* [tu]$_{\text{préf.pers.}}$ [Ø]$_{\text{préd.v.}}$ [íipfuuz]$_{\text{rad.}}$ [a]$_{\text{asp.inacc.}}$
'nous souhaitons'

twiipfuuje [tu]$_{\text{préf.pers.}}$ [Ø]$_{\text{préd.v.}}$ [íipfuuz]$_{\text{rad.}}$ [ye]$_{\text{asp·accompli}}$
'nous avons souhaité'

Il faut donc comprendre que nos radicaux verbaux correspondent aux vocables verbaux. Il en est de même pour les adjectifs. Les substantifs sont présentés dans leur forme du singulier.

Quant aux vocables grammaticaux, nous avons distingué deux sous-catégories : les flexionnels et les non flexionnels. Les deux sont reconnus grâce au dictionnaire des mots grammaticaux incorporé dans l'analyseur. Ce dernier les enregistre dans un fichier à part et nous les y avons récupéré pour créer un index de leurs fréquences avec *WordCruncher*.

Pour la classe des vocables, nous avons opté pour le codage suivant :

1 = vocable verbal
2 = vocable substantif à base nominale
3 = vocable substantif à base verbo-nominale
4 = vocable adjectival
5 = vocable grammatical

Ce codage est utile pour la consultation des index des vocables réalisés avec SPSS 7.5 que nous fournissons sur la disquette en annexe.

2. ANALYSE DU VOCABULAIRE DE BASE DU KIRUNDI ÉCRIT

2.1. LES VOCABLES DONT U ≥ 3

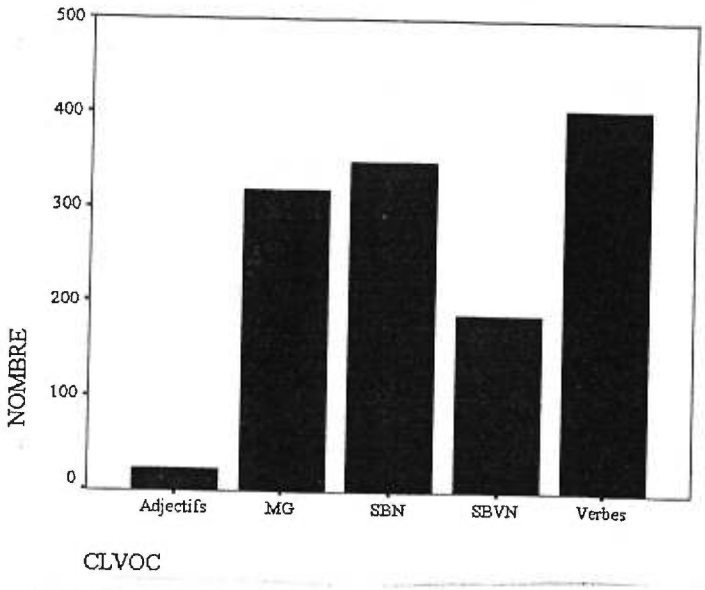
Les vocables dont U ≥ 3 constituent le noyau du vocabulaire du kirundi écrit. Ils comprennent, comme on l'a vu, 1 288 vocables qui représentent 32% du vocabulaire et 87% du nombre de mots-formes de tout le corpus.

La tranche des vocables dont U ≥ 3 est dominée, dans l'ordre, par les vocables verbaux, grammaticaux, nominaux à base nominale et nominaux à base verbo-nominale. Les vocables adjectivaux sont très peu nombreux. Nous fournissons les données dans le tableau 17.

Classes de V	*V*	$\bar{X}$	σ
Verbaux	406	60,56	156,53
Nominaux à base nominale	350	51,57	101,79
Nominaux à base verbo-nominale	189	26,47	37,91
Adjectivaux	24	55,41	80,85
Grammaticaux	320	129,24	378,76
Total	**1 288**	**70,00**	**218,35**

Tableau 17 - *Les vocables dont U ≥ 3*

Comme on le voit, les vocales verbaux sont les plus nombreux suivis par les vocables nominaux à base nominale et à base verbo-nominale. Viennent enfin les vocables grammaticaux. L'histogramme suivant, où CLVOC = classes de vocables, MG = mots grammaticaux, SBVN = substantifs à base verbo-nominale, SBN = substantif à base nominale illustre la représentativité des différentres catégories de vocables dans cette tranche importante du corpus.



Graphique 1 - *Rapports entre les catégories grammaticales dont U ≥ 3*

Nous analysons la composition de chacune de ces classes de vocables.

2.1.1. LES VOCABLES VERBAUX DONT U ≥ 3

Les vocables verbaux dont U ≥ 3 sont au nombre de 406. Ils représentent un peu plus de la moitié des vocables verbaux du vocabulaire du kirundi écrit (406 / 804), soit 50% des vocables verbaux du corpus. Si l'on tient compte du fait que les vocables verbaux dont U ≥ 3 constituent à eux seuls 23% de N (24 590 / 103 561), on peut conclure que les vocables verbaux dont U ≥ 3 occupent une place centrale dans le vocabulaire de base du kirundi écrit.

D'un point de vue qualitatif, on peut catégoriser les vocables verbaux sous plusieurs aspects. Touratier (1983 : 179-199) en relève cinq : morphosémantique, fonctionnel, morphologique, syntaxique et sémantique.

La catégorisation morphosémantique des verbes utilise des notions sémantiques indirectement morphologisantes. L'on dira par exemple que le verbe est « *ce qui exprime le temps (...), la personne* » (Touratier 1983 : 184-185). Nous ne la retenons pas; elle est redondante à la classification morphologique.

Quant à la catégorisation fonctionnelle, elle se fonde sur l'idée que le verbe est le noyau du prédicat (Martinet 1979 : 84). Nous ne retenons pas cette catégorisation; elle est plus proche de la syntaxe que de la lexicologie.

Nous adoptons une catégorisation des vocables fondée sur les aspects morphologiques, syntaxiques et sémantiques.

2.1.1.1. ASPECTS MORPHOLOGIQUES

Les aspects morphologiques du verbe concernent notamment la conjugaison. L'on distinguera par exemple pour le français les verbes en [-er], en [-ir], etc.

Pour une langue agglutinante comme le kirundi, la morphologie permet d'aborder deux aspects importants des verbes à savoir la défectivité morphologique et la structure des radicaux.

2.1.1.1.1. Les vocables verbaux morphologiquement défectifs

Au sein des vocables verbaux dont U ≥ 3 se retrouvent des verbes morphologiquement défectifs. Ces verbes sont caractérisés par une défectivité formelle systématique. Mel'cuk (1993 : 358-360) la définit comme une défectivité due à l'impossibilité de former des mots-formes par une combinaison de morphèmes tout à fait concevable.

Les verbes concernés ne peuvent pas recevoir comme les verbes réguliers le morphème marqueur du futur [-zoo-], le morphème de l'infinitif [-ku-] ou une des trois marques aspectuelles ([-a] 'inaccompli', [-e] 'inaccompli inchoatif', [-ye] 'accompli'). De plus, ces verbes n'acceptent aucun suffixe de dérivation. Nous en présentons les radicaux avec leurs indices de fréquence, de dispersion et d'usage.

Vocable	F_O	*D*	*U*	*Glose*
[-RI]	2 466	0,94	2 311,4	'être'
[-TI]	450	0,73	330,71	'dire'
[-FÍT-]	273	0,91	249,67	'avoir'/'posséder'
[-ZI]	248	0,88	217,60	'connaître'/'savoir'

Signalons que le vocable [-ri] 'être' possède les indices de fréquence et d'usage les plus élevés de tous les vocables verbaux.

2.1.1.1.2. Les vocables verbaux de structure CV

Un regard sur la morphologie du verbe kirundi nous permet aussi de faire un lien entre la structure phonologique des radicaux et leurs fréquences.

En effet, lorsque l'on sait que la structure canonique des radicaux verbaux du kirundi est CVC (consonne - voyelle - consonne), on constate que parmi les radicaux les plus fréquents et à U ≥ 3 viennent en tête ceux de structure CV (consonne-voyelle), une structure de radical plutôt rare dans le lexique (Faïk-Nzuji 1992) des langues bantoues. Ce sont les vocables verbaux suivants :

Vocable	F_O	*D*	*U*	*Glose*
[-BÁ-]	1 105	0,96	1 059,4	'être' / 'habiter'
[-CÍ-]	747	0,86	639,30	'faire consécutivement'
[-GÍ-]	556	0,89	494,34	'aller'
[-VU-]	245	0,93	226,87	'quitter'
[-HÁ-]	215	0,91	195,85	'donner'
[-PFÚ-]	103	0,81	83,85	'faire malgré'
[-TÁ-]	94	0,87	81,56	'jeter'
[-GU-]	45	0,90	40,34	'tomber'
[-SE-]	47	0,68	32,19	'moudre'
[-RÍ-]	47	0,66	31,12	'manger'
[-NYÓ-]	39	0,61	23,93	'boire'
[-HÍ-]	25	0,66	17,43	'brûler'

Comme on le voit, ces vocables verbaux ont, dans l'ensemble, des indices de fréquence, de dispersion et d'usage élévés. La structure CV des radicaux verbaux est donc importante dans le vocabulaire du kirundi écrit.

2.1.1.2. ASPECTS SYNTAXIQUES

D'un point de vue syntaxique, le verbe est considéré comme le point d'ancrage des autres unités de la phrase verbale. L'intérêt d'une catégorisation syntaxique des vocables verbaux réside dans le fait qu'elle permet de mettre en évidence des fonctions syntaxiques particulières à certains vocables. Elle nous permet notamment d'isoler au sein des vocables verbaux dont $U \geq 3$ un ensemble de verbes auxiliaires modaux.

Suivant Dubois *et al.* (1994 : 305), on appelle modaux ou auxiliaires modaux, *« la classe des auxiliaires du verbe qui expriment les modalités logiques »*. Ces modalités relèvent des oppositions comme contingent / nécessaire, probable / possible, etc. En français, *devoir* et *pouvoir* suivis de l'infinitif sont par exemple des auxiliaires modaux.

Greimas & Courtés (1993 : 231) distinguent trois catégories de modalités fondées sur le « *parcours tensif menant à la réalisation* » : modalités virtualisantes (*devoir, vouloir),* modalités actualisantes (*pouvoir, savoir)* et modalités réalisantes (*faire, être*).

Nous utilisons cette classification de Greimas & Courtés (1993 : 231) pour catégoriser les auxiliaires modaux de notre corpus[1]. Nous reconnaissons cependant que les limites entre les classes ne sont pas étanches et qu'il existe toujours des possibilités de classer différemment un vocable.

2.1.1.2.1. Les auxiliaires virtualisants

Les auxiliaires modaux virtualisants sont des vocables verbaux de sens essentiellement volitif. Ceux dont U ≥ 3 sont fournis ci-dessous.

Vocable	F_O	*D*	*U*	*Glose*
[-SHAAK-]	132	0,87	115,10	'vouloir *x*'
[-ÍIPFUUZ-]	85	0,86	72,91	'souhaiter *x*'
[-GOMB-]	61	0,80	48,92	'vouloir *x*'
[-ÍIZIGIR-]	19	0,81	15,42	'espérer *x*'
[-BWÍIRIZW-]	18	0,69	12,41	'ordonner'/'diriger'
[-ÍIZEER-]	9	0,79	7,08	'espérer *x*'

2.1.1.2.2. Les auxiliaires actualisants

Les auxiliaires actualisants sont liés à la réalisation du procès. Ceux dont U ≥ 3 sont, dans notre corpus, au nombre de quatre; nous les présentons ci-dessous.

---

[1] Notons qu'il n'y a aucune contradiction qu'un verbe soit défectif au plan morphologique et modal au plan syntaxique.

Vocable	F_O	*D*	*U*	*Glose*
[-SHÓBOR-]	280	0,89	250,41	'être capable de faire *x*'
[-TÚM-]	202	0,93	187,00	'causer *x*'
[-GERAGEZ-]	47	0,82	38,73	'essayer de faire *x*'
[-SHÓBOK-]	36	0,79	28,34	'être possible'

2.1.1.2.3. Les auxiliaires réalisants

Les auxilaires réalisants sont très nombreux dans notre corpus Ce sont des vocables verbaux qui expriment souvent des modalités de « faire » et de « être ». Nous en dressons la liste ci-dessous.

Vocable	F_O	*D*	*U*	*Glose*
[-BÁ-]	1 105	0,96	1 059,4	'être/habiter'
[-CÍ-]	747	0,86	639,30	'faire consécutivement *x*'
[-GI-]	556	0,89	494,34	'aller'
[-GIR-]	503	0,94	472,87	'faire'
[-MAR-]	318	0,95	300,94	'finir *x*'
[-SHIK-]	285	0,92	262,10	'arriver'
[-SANG-]	260	0,91	237,90	'trouver *x* '
[-Z-]	261	0,89	232,46	'venir'
[-GEND-]	236	0,91	219,99	'partir'
[-TÁANGUR-]	212	0,89	196,87	'commencer *x*'
[-SÚBIR-]	197	0,92	180,53	'recommencer *x*[2]'
[-ÁAM-]	182	0,93	169,83	'être toujours en train de'
[-ONGER-]	140	0,91	127,28	'faire ensuite *x*'
[-GUM-]	117	0,86	100,48	'faire constamment'
[-HÓR-]	112	0,87	97,07	'faire d'habitude *x*'
[-PFÚ-]	103	0,81	83,85	'faire *x* avec peu de chances'
[-SÍGAR-]	100	0,81	81,44	'faire *x* désormais'
[-BANZ-]	77	0,84	64,70	'commencer par *x*'

[2] Où *x* est un mot-forme verbal.

[-BANDANY-]	75	0,80	59,85	'continuer'
[-ÍIGER-]	58	0,87	50,29	'avoir fait *x*'
[-HÉRUUK-]	52	0,88	45,94	'avoir eu lieu dernièrement'
[-HÍT-]	30	0,65	19,59	'faire *x* en passant'
[-IMIRIZ-]	19	0,75	14,33	'faire bientôt *x*'
[-ÍISHIING-]	13	0,73	9,49	'faire surtout *x*'

Les indices élevés de ces vocables verbaux sont à relier à leur fonction syntaxique d'auxiliaire. L'auxiliaire constitue l'élément verbal qui accompagne le verbe lexical dans l'énoncé. Les vocables auxiliaires expliquent aussi la haute fréquence des mots-formes verbaux.

De nombreux autres vocables verbaux dont U ≥ 3 sont lexicaux. Ils recouvrent des champs lexicaux divers. Leur classement relève de critères sémantiques.

2.1.1.3. ASPECTS SÉMANTIQUES

La catégorisation sémantique des vocables verbaux est fondée sur la notion de « procès »; mais comme le soulignent Baylon & Fabre (1978 : 27), la difficulté réside dans le fait que *« le procès recouvre bien des choses différentes : actions, sensations, sentiments, activités intellectuelles, etc. ».*

Mel'cuk (1994 : 69-74) oppose par exemple des verbes dynamiques qui représentent un événément, (par ex. *construire*) à des verbes statiques, qui représentent un état (par ex. *aimer*) .

Les auteurs des vocabulaires de base tel Gougenheim *et al.* (1964) et Rivière (1979) ont tenté, pour le français, d'identifier un ensemble de champs lexicaux qui engloberaient tout le vocabulaire de l'activité humaine et dans lesquels seraient intégrés tous les vocables verbaux de base sélectionnés.

Dans la même veine, mais pour l'anglais, Mazareno & Mazareno (1988) distinguent 61 classes de vocables de l'anglais contenus dans les ouvrages didactiques de l'école primaire.

De son côté, Levin (1993) fournit une classsification systématique des verbes de l'anglais dans 49 classes sémantiques.

Dans les limites de notre recherche, nous ne pouvons procéder à la manière de Levin (1993). Une telle classification des vocables verbaux ferait l'objet d'une autre recherche. Il faudrait inventorier d'abord les classes de vocables puis classer ces derniers.

Sans minimiser l'intérêt de tels découpages de la réalité à des fins lexicologiques, il reste que l'acceptation d'un vocable dans un champ lexical est souvent subjective et discutable.

Nous distinguerons, aux fins des applications pédagogiques que nous envisageons plus loin, les vocables verbaux thématiques des vocables athématiques. Les premiers sont caractérisés par une haute fréquence et une faible dispersion tandis que les seconds ont une fréquence et une dispersion équilibrées. Nous indiquons également les vocables de notre corpus qui couvrent des relations lexico-sémantiques cruciales dans l'enseignement du vocabulaire.

2.1.1.3.1. Les vocables verbaux lexicaux thématiques

Les vocables verbaux lexicaux thématiques sont caractérisés par une fréquence élevée et une dispersion faible. Leur fréquence est due à la récurrence des thèmes auxquels ils sont liés.

Nous contrastons ci-dessous des vocables de même fréquence et de dispersions différentes pour illustrer les écarts. Nous indiquons entre parenthèses le thème de notre corpus auquel est liée la haute fréquence du vocable. Nous indiquons (général) quand le vocable n'est relié à aucun thème particulier.

Vocable	F_O	*D*	*U*	*Glose*
[-BÍIK-]	10	0,38	3,76	'conserver'(développement)
[-NÓONOOSOR-]	10	0,77	7,66	'approfondir *x*' (général)
[-BÁTIZ-]	11	0,46	5,04	'baptiser *x*' (religieux)
[-NYÚR-]	11	0,75	8,30	'intérioriser' (général)
[-VAAVAANUR-]	13	0,51	6,65	'rompre définitivement' (politique)
[-REKUR-]	13	0,70	9,10	'relâcher' (général)

[-ONK-]	16	0,45	7,17	'téter' (social)
[-HÉREKEZ-]	16	0,69	11,10	'accompagner' (général)
[-GENG-]	18	0,57	10,26	'régir' (politique)
[-RINDIIR-]	18	0,74	13,40	'attendre' (général)
[-HAKAN-]	18	0,78	14,06	'nier' (général)
[-JEEJW-]	45	0,73	33,00	'être responsable de' (politique)
[-GU-]	45	0,90	40,34	'tomber' (général)

Les vocables thématiques sont importants en didactique des langues. L'enseignement du vocabulaire est en effet fondé sur des textes articulés autour de thèmes centralisateurs. Nous illustrons au chapitre 5 § 3 l'utilisation de vocables thématiques en classe de vocabulaire du kirundi.

2.1.1.3.2. Les vocables verbaux lexicaux athématiques

Les vocables verbaux non thématiques sont des vocables lexicaux qui ont une fréquence et une dispersion équilibrées. Ils sont essentiels dans un vocabulaire de base dans la mesure où ce sont les verbes pleins communs à tous les types de textes. Nous fournissons ci-dessous la liste de ceux dont la fréquence dépasse la moyenne de 60, moyenne établie pour les vocables dont U ≥ 3. L'on consultera la liste sur disquette pour s'en faire une idée exhaustive. Il s'agit des vocables suivants :

Vocable	F_O	*D*	*U*	*Glose*
[-BÓN-]	462	0,89	410,73	'voir'
[-VÚG-]	431	0,90	389,56	'parler'/'dire'
[-KÓR-]	329	0,91	298,79	'travailler'
[-ÚUMV-]	296	0,91	270,32	'entendre'/'écouter'
[-FÍT-]	273	0,91	249,67	'avoir *x*'/'posséder *x*'
[-MENY-]	269	0,90	243,34	'savoir'/'connaître'/'réussir'
[-TÉER-]	265	0,92	242,93	'causer *x*'

[-Z-]	261	0,89	232,46	'venir'
[-VU-]	245	0,93	226,87	'quitter un endroit'
[-SHING-]	255	0,87	220,95	'décider'
[-FÁT-]	238	0,92	219,26	'prendre'/'toucher un salaire'
[-ZI]	248	0,88	217,60	'connaître
[-GEND-]	236	0,91	213,99	'aller'/'partir'
[-HÁ-]	215	0,91	195,85	'donner *x* à *y*'
[-ÉEREK]	200	0,91	181,39	'montrer'
[-RONK-]	189	0,92	174,04	'recevoir *x*'
[-ÉEMER-]	205	0,81	165,48	'croire'/'accepter'
[-BWÍIR-]	199	0,82	162,83	'dire *x* à *y*'
[-BÁZ-]	207	0,78	162,21	'demander *x*'
[-SAB-]	168	0,94	157,78	'demander'
[-ÍIT-]	174	0,89	155,22	'donner un nom à *x*'
[-RAAB-]	167	0,91	152,12	'regarder'
[-ÍIG-]	146	0,91	133,41	'étudier'
[-TÓOR-]	152	0,87	132,65	'élire'/'choisir'
[-KWÍIR-]	148	0,82	120,99	'être convenable'
[-RONGOOR-]	146	0,82	120,26	'diriger'
[-TÁANG-]	132	0,91	119,63	'donner *x*'
[-ÚUBAK-]	124	0,84	103,48	'construire'
[-IIBUK-]	115	0,85	97,68	'se souvenir de *x*'
[-SHÍR-]	110	0,88	96,32	'mettre *x* [PRÉP.] *y*'
[-KOMER-]	106	0,89	94,46	'être en bonne santé'
[-KEN-]	103	0,85	87,05	'manquer de *x*'
[-RONDER-]	102	0,85	86,81	'chercher'
[-GOOR-]	101	0,84	84,87	'être difficile'
[-TÉGEREZ-]	96	0,86	82,23	'obliger *x* à faire *y*'
[-RÁAR-]	94	0,87	81,66	'passer la nuit'
[-TÁ-]	94	0,87	81,56	'jeter'
[-TÉGUUR-]	95	0,82	78,04	'préparer'
[-RÉENG-]	88	0,87	76,86	'outrepasser *x*'
[-KÚUND-]	95	0,80	76,46	'aimer'

[-SHIIM-]	86	0,85	72,69	'apprécier'
[-TWÁAR-]	82	0,87	71,43	'gouverner'
[-RANGUUR-]	82	0,83	67,90	'réaliser *x*'
[-SHÍMIK-]	85	0,78	66,07	'insister'
[-SANGIR-]	81	0,81	65,23	'partager'
[-GARUK-]	73	0,88	64,54	'revenir'/'retourner'
[-RWAAN-]	78	0,79	61,70	'se battre'
[-KÓRAN-]	68	0,86	58,73	'se réunir'
[-ÍIHWEEZ-]	70	0,83	57,77	'analyser'
[-ÍIC-]	69	0,83	56,99	'tuer'
[-SOM-]	71	0,79	56,12	'lire'
[-ÁANK-]	71	0,76	54,11	'haïr'/'refuser'
[ANDIK-]	64	0,81	51,96	'écrire
[-RAMUK-]	69	0,74	50,93	'passer la nuit'
[-GUR-]	70	0,71	49,75	'acheter'

2.1.1.3.3. Les relations lexico-sémantiques

Nous distinguons au sein des nombreuses relations lexico-sémantiques (cf. Ménard 1988) entre les vocables verbaux, deux relations fondamentales en didactique des langues : la synonymie et l'antonymie.

2.1.1.3.3.1. La synonymie

Il n'est pas dans notre intention d'entreprendre ici une analyse poussée de cette notion. Le lecteur interessé pourra consulter les importants travaux de sémantique relatifs à cette question dont celui de Lyons (1995 : 60-65, 128-129) pour ne citer qu'un des plus récents.

Nous ne discutons donc pas non plus de l'opposition synonymie parfaite (totale) / partielle (Baylon & Fabre 1979 : 167-173); nous nous intéressons plutôt aux vocables verbaux substituables dans certains contextes et non dans d'autres et qui sont appelés tantôt « parasynonymes » tantôt « quasi-synonymes ».

Soulignons cependant que la synonymie connaît beaucoup de nuances. Comme le montre Ménard (1988), la relation d'équivalence concerne tantôt des unités permutables tantôt des unités substituables. Les premières comprennent les oppositions anaphores / cataphores et synonymes / coréférents lexicaux (ex. ZÈBRE / PRISONNIER). Les unités substituables sont quant à elles en rapport d'hyperonymie / hyponymie (ex. BÊTE / MOUTON) ou de base / base + modificateur (ex. FILLE / FILLETTE).

Dans la mesure où nous restons à l'intérieur des limites de notre corpus, ces nuances ne sont pas toutes exprimées et les oppositions sont moins nombreuses.

Ainsi, lorsque l'on parcourt la liste des vocables verbaux dont U ≥ 3 (cf. liste), on compte douze vocables verbaux quasi-synonymes. Nous les présentons par paire avec leurs indices de fréquence, de dispersion et d'usage.

(151)	*Vocable*	F_O	*D*	*U*	
	a. [-SHAAK-]	132	0,87	115,10	'vouloir *x*'
	b. [-GOMB-]	61	0,80	48,92	'souhaiter *x*'
	c. [-RWAAN-]	78	0,79	61,70	'se battre'
	d. [-HÁRANIR-]	34	0,79	26,99	'lutter pour *x* '
	e. [-ÍIC-]	69	0,83	56,99	'tuer'
	f. [-GANDAGUR-]	21	0,68	14,18	'assassiner'
	g. [-VYÁAR-]	26	0,71	18,55	'donner naissance à'
	h. [-ÍIBARUK-]	13	0,68	8,87	'enfanter' [+ soutenu]
	i. [-ÍIZIGIR-]	19	0,81	15,42	'espérer'/'avoir confiance'
	j. [-ÍIZEER-]	9	0,79	7,08	'mettre sa confiance en *x*'
	k. [-RÍ-]	47	0,66	31,12	'manger'
	l. [-FUNGUR-]	23	0,46	10,61	'manger' [+ soutenu]

Les unités (151 a, b) sont des auxiliaires modaux à sens volitif. Une étude syntactico-sémantique approfondie permettrait de saisir les nuances entre ces unités. Nous nous limitons ici à montrer qu'elles ne sont pas substituables dans tous les contextes.

(152) a. *twaashaaka kubíbasúbirira mwó* 'nous voulions vous le répéter' (w4)
b. *twaagomba kubíbasúbirira mwó* 'nous voulions vous le répéter'
c. *ivyó ushaaká ni bigirwé* 'que ta volonté soit faite' (w7)
d.**ivyó ugombá ni bigirwé*

Fournissons aussi quelques contextes pour les vocables verbaux restants :

(153) a. *twizigiye kó amacáakubíri agiiye guhéra*
'nous espérons que les divisions vont cesser' (w14).
b. *twizeeye kó amacáakubíri agiiye guhéra*
'nous espérons que les divisions vont cesser'.
c. *umuntu ntí yookwiizeera ivya gusa*
'l'individu ne devrait pas compter sur des gratuités'(w9).
d. **umuntu ntí yookwiizigira ivya gusa*

e. *eka barafungura barahaaga* 'et ils mangèrent, se rassasièrent' (w6)
f. *eka bararyá barahaaga* 'et ils mangèrent, se rassasièrent'
g. *kumenya akamaro ko gufungura néezá*
'connaître l'utilité de bien se nourrir' (w13)
h. **kumenya akamaro kó kuryá néezá*

i. *mu myáaka Afrika yaháranira ukwiíkuukira*
'pendant les années où l'Afrique luttait pour son indépendance' (w4)

j. *mu myáaka Afrika yarwáanira ukwiíkuukira*

'pendant les années où l'Afrique luttait pour son indépendance'

k. *ináama kaminúuza iharánira amajambere*

'Conseil suprême pour la révolution' (w11)

l.**ináama kaminúuza irwaanira amajambere*

L'unité (151 g) est polysémique et s'applique aussi bien aux plantes, aux animaux qu'aux humains. Elle contient donc le sens de (151 h) qui est monosémique. En commutant l'unité en (151 g) par celle en (151 h), on se rend compte que la commutation fournit tantôt un énoncé grammatical tantôt non, tel qu'illustré en (154 d, f).

(154) a. *Umwé mutóoyá aravyáara*

'la plus petite donna naissance à un enfant' (w16)

b. *Umwé mutóoyá aríibaruka* [3].

'la plus petite donna naissance à un enfant'.

c. *(...) kukó umugóre wíiwé atavyaará.*

'parce que sa femme est stérile' (w5).

d.**(...) kukó umugóre wíiwé atiibáruka.*

e. *igitooke kiravyáara* 'le bananier donne des rejets'.

f. **igitooke kiríibaruka*

Ainsi donc, les unités en (151 a) et (151 b), (151 c) et (151 d), (151 e) et (151 f), (151 k) et (151 l) ne sont pas substituables dans tous les contextes. Il en est de même de (151 g) et (151 h). Elles sont quasi-synonymes. Nous proposons au chapitre 5 des activités d'apprentissage du vocabulaire fondées sur les quasi-synonymes verbaux.

---

[3] Les exemples en (154 b) et (154 e) ne sont pas tirés du corpus. Nous les fournissons en tant que locuteur natif du kirundi.

2.1.1.3.3.2. L'antonymie

Les antonymes sont des unités lexicales de sens contraires. Cette définition lapidaire cache de nombreuses nuances. Ménard (1988) distingue par exemple des contrastes binaires et des contrastes non binaires.

Les contrastes binaires comprennent quatre sous-classes : les contrastes gradables (LENTEMENT / VIVEMENT), les contrastes non gradables (NAÎTRE / MOURIR), les contrastes converses (ÉLÈVE / PROFESSEUR), les contrastes en opposition directionnelle et antipodale (EN ARRIÈRE / EN FACE) ainsi que les oppositions tout / partie.

Quant aux contrastes non binaires, Ménard (1988) distingue des contrastes de même niveau taxonomique (ex. DIVAN / CHAISE), des contrastes sériés (jours de la semaine par exemple) et des contrastes en opposition cyclique et orthogonale (ex. EN ARRIÈRE / À CÔTÉ).

L'on comprendra que toutes ces nuances ne sont pas exprimées à l'intérieur de notre corpus. Les antonymes verbaux se retrouvent notamment au sein des verbes qui expriment le mouvement et la position. Ils correspondent majoritairement à des vocables en opposition directionnelle. Nous dressons la liste de ceux dont U ≥ 3 dans notre corpus:

(155)	[-gi-]	'aller'	~	[-vu-]	'quitter'
	[-shik-]	'arriver'	~	[-gend-]	'partir'
	[-taah-]	'rentrer'	~	[-shik-]	'il arrive'
	[-z-]	'venir'	~	[-gend-]	'partir'
	[-finjir-]	'entrer'	~	[-sohok-]	'sortir'
	[-iicar-]	's'asseoir'	~	[-hágarar-]	'se tenir debout' / 's'arrêter'

D'un point de vue pédagogique, la synonymie et l'antonymie constituent des relations lexico-sémantiques de première importance. Elles répondent aux exigences métalinguistiques inhérentes à la classe de langue exprimées par les questions du type « que signifie, que veut dire (un mot) ? » et par les réponses de l'enseignant : « ce mot signifie, veut dire (...) ».

Les relations de synonymie et d'antonymie sont également importantes dans la reformulation (orale ou écrite) des énoncés. N'exige-t-on pas aux apprenants de ne pas répéter les mêmes mots dans leurs textes ou de remplacer une forme affirmative par une forme négative équivalente (ex. *il est sourd / il n'entend pas*) ? En fait, connaître le sens d'un vocable c'est connaître aussi son (ses) antonyme(s) et son (ses) synonyme(s).

2.1.2. LES VOCABLES NOMINAUX À BASE NOMINALE

Dans une étude que nous avons faite (Ntirampeba 1993 : 94), nous sommes arrivé à la conclusion que la solution optimale pour la lemmatisation des substantifs est d'adopter comme vocable le substantif dans sa forme du singulier, en ramenant tous les dérivés préfixaux[1] à ce lemme.

En présence de variantes phonétiques / graphiques (*ifaranga / ifranga* 'franc' (monnaie) par ex.), le lemme est la variante qui respecte la structure de la langue (ici *ifaranga*); la séquence /f+r / > /fr/ n'étant pas attestée ailleurs en kirundi.

Lorsqu'une variante domine numériquement, elle est également considérée comme lemme (par exemple *ikáwa* est pris comme lemme dans la paire *ikáwa / akáwa* de fréquences respectives 46 et 7). Ces précisions étant données, examinons la part des vocables à base nominale dont U ≥ 3 dans le vocabulaire de base du kirundi écrit.

Les vocables nominaux à base nominale sont au nombre de 1 165 dans le corpus. Ils représentent 19 893 occurrences soit 19% du corpus. Au sein des 1 165 vocables à base nominale, ceux qui ont U ≥ 3 sont à 349 et représentent 18 047 occurrences soit 90% des occurrences des substantifs à base nominale de tout le corpus; les 10% restants étant couverts par les 816 dont U est inférieur à 3.

La fréquence moyenne des substantifs à base nominale dont U ≥ 3 est de 51, une moyenne inférieure à celle des verbes de la même tranche de vocables.

---

[1] Les dérivés préfixaux relèvent, rappelons-le, de la commutation des préfixes de classe pour un même radical.

Nous isolons au sein des vocables à base nominale des vocables métadiscursifs et analysons le reste en regard de l'opposition thématique / athématique.

2.1.2.1. LES VOCABLES MÉTADISCURSIFS

On peut relever parmi les substantifs à base nominale des vocables inhérents à l'articulation du discours. Ils le situent notamment dans le temps et servent dans la description et l'argumentation. Ce sont :

Vocable	F_O	*D*	*U*	*Glose*
UMUÚNSI	416	0,94	390,92	'jour'
IGIHE	282	0,93	293,96	'le temps de'
UMWÁAKA	336	0,88	296,16	'année'
UKWÉEZI	99	0,85	84,11	'mois'
ISÁHA	57	0,72	40,93	'heure'
AKARORERO	48	0,83	39,88	'exemple'
IJORO	33	0,68	22,35	'nuit'
UKUGÉNE	28	0,71	17,41	'la manière' / 'la façon'
UKUNTU	19	0,76	14,38	'la manière' / 'la façon'
INDWI	18	0,60	10,82	'semaine'
UMUNÓTA	15	0,52	7,83	'minute'

D'autres vocables sont plutôt reliés à la retranscription de dates en texte. Ce sont :

Vocable	*Fo*	*D*	*U*	*Glose*
IMIRONGO	662	0,89	587,73	'dizaines'
ICEÉNDA	648	0,83	540,20	'neuf' (chiffres)
IJANA	504	0,88	445,47	'cent' (chiffre)
IGIHUMBI	475	0,89	423,29	'mille' (chiffre)
ICÚMI	224	0,89	198,75	'dix'

INDWI	151	0,82	124,53	'sept'
UMUNAÁNI	132	0,87	114,63	'huit'

Le reste des substantifs à base nominale dont U ≥ 3 (332 vocables) sont soit thématiques soit athématiques. Comme ils constituent une liste très fournie, nous retiendrons, aux fins de notre présentation, ceux dont la fréquence est supérieure à 51 soit la fréquence moyenne des vocables à base nominale dont U ≥ 3. Présentons d'abord les vocables à base nominale thématiques.

2.1.2.2. LES VOCABLES THÉMATIQUES

La liste des vocables à base nominale thématiques est très riche. Nous indiquons entre parenthèses le thème auquel est relié le vocable lorsque la glose ne l'exprime pas de façon évidente. Nous les rangeons en ordre de fréquence pour la contraster avec la dispersion. L'on consultera l'index par fréquence décroissante pour se faire une idée de la dispersion normalement attendue pour les vocables de fréquence similaire. Les vocables nominaux thématiques à base nominale dont U ≥ 3 sont les suivants :

Vocable	F_O	*D*	*U*	*Glose*
UMUGAMBWE	412	0,65	268,55	'parti politique'
IMÁANA	252	0,58	146,59	'Dieu'
UBÚMWE	190	0,62	117,91	'unité nationale' (politique)
UMUSHIKIRANGANJI	144	0,63	91,06	'ministre'
EKLEZIYA	120	0,47	55,81	'église'
PAAPA	110	0,42	46,44	'pape'
PREZIDA	100	0,53	53,11	'un président'
UMUKRISTÚ	98	0,47	45,63	'un chrétien'
UMUPÁATÍRI	82	0,62	50,86	'prêtre'
URUBAÁNZA	77	0,69	52,88	'procès' (social)
IDIYOSEÉZE	66	0,38	24,93	'diocèse'
IPARUWAÁSE	66	0,50	32,77	'paroisse'

UMUGANGA	54	0,57	30,99	'médecin' (social)
URWÉEGO	54	0,50	27,03	'institution politique'
UMWEEPÍISKÓPI	53	0,37	19,63	'évêque'

On peut relever par exemple, pour le thème religieux les vocables thématiques suivants : *ubusáserdoóti* 'prêtrise', *umuséenyeri* 'évêque', *umweepískópi* 'évêque' *umusáserdoóti* 'prêtre' *umupatiri* 'prêtre', *idiyoseéze* 'diocèse', *iparuwaási* 'paroisse'. Notons au passage que ces vocables sont des emprunts intégrés au kirundi.

2.1.2.3. LES VOCABLES ATHÉMATIQUES

Hormis les vocables qui servent à la structuration du discours, ceux qui correspondent à des chiffres réécrits et les vocables thématiques, les vocables nominaux à base nominale dont U ≥ 3 comprennent des vocables athématiques.

Pour alléger notre présentation, nous ne retenons ici que ceux dont la fréquence dépasse 51 (fréquence moyenne des vocables nominaux à base nominale dont U ≥ 3). Ce sont :

VOCABLE	F_O	*D*	*U*	*Glose*
UMUNTU	838	0,91	758,58	'personne humaine'
IGIHÚGU	824	0,82	679,24	'pays'
UMURUÚNDI	343	0,79	270,52	'un Burundais'
UMWÁANA	245	0,79	194,46	'enfant'
IJAMBO	217	0,82	178,37	'parole'
IKINTU	193	0,90	173,63	'chose'
ISHUÚRE	179	0,82	147,31	'école' / 'classe'
UMUTÍMA	170	0,82	139,45	'cœur'
UMUGAÁMBI	165	0,79	130,51	'projet'
UMUSHÍINGANTAÁHE	152	0,85	129,92	'homme respectable'
UMUNYÁGIHÚGU	152	0,74	113,73	'citoyen'

UBÚRYO	124	0,90	112,10	'moyen de faire *x*'
IKÓMIÍNE	141	0,79	111,33	'commune'
IBANGA	128	0,85	108,62	'secret'/'profession'
ISHÍRAHÁMWE	138	0,79	108,42	'association'
IFARANGA	140	0,73	102,74	'monnaie'
INTÁARÁ	137	0,68	93,10	'province'
UMUVYÉEYI	119	0,77	91,61	'parent'
ISÍ	114	0,73	83,34	'terre'
UMURYANGO	108	0,71	76,65	'famille'
UKÚRI	85	0,89	75,64	'vérité'
UMURWI	94	0,80	74,86	'équipe' / 'groupe'
URUGÓ	92	0,79	72,26	'enclos'
NYAKUUBAHWA	84	0,82	68,68	'son Excellence'
UBUZIMA	90	0,76	68,47	'vie'
AKAZI	89	0,76	67,91	'travail rémunéré'
IZÍNA	79	0,85	67,50	'nom'
INZU	85	0,78	66,69	'maison'
INGINGO	80	0,83	66,27	'concertation'
INZIRA	74	0,88	64,91	'voie' / 'chemin'
UMUGÓRE	95	0,65	61,30	'femme'
ITEERAMBERE	76	0,74	56,51	'progrès'
AMAHÓRO	143	0,37	52,68	'paix'
UBWÓOKO	68	0,77	52,27	'ethnie'
IGICÉ	58	0,85	49,57	'moitié'
AMÁAZI	64	0,76	48,73	'eau'
AMAKUNGU	55	0,86	47,19	'pays étrangers'
IBARABARA	55	0,84	46,12	'route'
UMUHARI	63	0,69	43,63	'groupe' / 'association'
IBIRO	53	0,82	43,21	'bureau'
AMAJAMBERE	63	0,68	42,88	'progrès'
IKÁWA	53	0,51	26,92	'café'

Les vocables nominaux à base nominale athématiques sont très importants en enseignement du kirundi L2. Ce sont les noms les plus usuels de la langue. Ils assurent une fonction fondamentale de désignation. Signalons en passant la présence dans cette liste de l'hyperonyme *ikintu* 'chose'.

2.1.2.4. LES RELATIONS LEXICO-SÉMANTIQUES

2.1.2.4.1. La synonymie

Dans la liste des substantifs à base nominale dont U ≥ 3 figurent des quasi-synonymes dont le quadruplet *umugóre / umupfáasoni / umukényezi / umuruúndikazi*; leurs gloses sont fournies en (156). Leurs indices de fréquence, de dispersion et d'usage se présentent comme suit :

		F_O	*D*	*U*	*Glose*
(156)	a. *umugóre*	95	0,65	61,30	'femme'
	b. *umukényezi*	49	0,65	32,09	'femme'
	c. *umupfáasóni*	39	0,74	28,98	'femme de haut niveau social'
	d. *umuruúndikazi*	18	0,70	12,55	'femme burundaise'

La substitution entre (156 a) et (156 b) et (156 c) est partielle. L'unité en (156 a) est la plus générale alors que celle en (156 b) est connotée " recherché ". L'unité en (156 c) désigne une femme d'un haut niveau social. Nous avons relevé dans notre corpus quelques contextes que nous contrastons avec des énoncés de notre choix pour illustrer les cas de non substituabilité entre les vocables :

(157) a. *abwiira umugóre wíiwé ati* 'il dit à sa femme ceci' (w1)
b. *abwiira umukényezi wíiwé ati* 'il dit à sa femme ceci'
c. *Nyenícúubahiro Prezida (...) n'úmukényezi wíiwe* 'Son Excellence le Président et sa femme' (w11)
d. *Nyenícúubahiro Prezida (...) n'úmupfáasóni wíiwe* 'Son Excellence le Président et sa femme' (w11)
e. * *Prezida (...) n'úmugóre wíiwe*

Ces contextes montrent que tantôt la substitution n'altère pas le sens de l'énoncé (157 a et 157 b, 157 c et 157 d) tantôt elle l'altère (157 c, d et 157 e).

D'un point de vue diachronique, l'unité en (156 c) *umupfáasoni* signifie 'personnage respectable', 'prince de sang' ou 'beau-fils' (Rodegem 1970 : 449). Le sens en (156 c) est récent et a complètement éclipsé l'ancien sens qui ne figure d'ailleurs pas dans notre corpus.

L'unité en (156 d) est plus utilisée dans les textes politiques, tout comme le sont les vocables synonymes suivants :

	F_O	D	U	*Glose*
amajaambere	63	0,68	46,66	'progrès'
iteerambere	76	0,74	56,51	'progrès'

Ces deux vocables sont des néologismes probablement créés dans les années 60 pendant la décolonisation et qui servent encore aujourd'hui de slogans politiques. Le premier est toujours un pluriel et le second toujours un singulier.

Deux vocables servent à la structuration du discours. Ils signifient 'la manière de' / 'la façon de'. Ce sont *ukugéne* (F_O = 28, D = 0,71, U = 19,93) et *ukuntu* (F_O = 19, D = 0,76, U = 14,38).

L'on note également deux paires d'emprunts dont les vocables nous semblent des synonymes parfaits. Nous n'avons pu avoir aucun contexte où ils ne sont pas substituables mutuellement. Il s'agit des vocables suivants :

umuséenyeri 'évêque' (du français *Monseigneur*)
umweepískópi 'évêque' (du grec *episcopus* 'surveillant')
umusáserdoóti 'prêtre' (du latin *sacerdos* 'prêtre')
umupáatiri 'prêtre' (du latin *pater* 'père')

Signalons notre surprise de constater que le vocable *umukoóbwa* 'fille' est rare à l'écrit (F_O = 16; D = 0,46; U = 7,40) alors que c'est, intuitivement, celui qui est le plus utilisé à l'oral. À sa place se retrouve *umwiígeme* 'fille' (F_O = 40; D = 0,62; U = 24,65) avec une connotation « noble » (Rodegem 1970 : 106).

Signalons également un cas où un substantif à base nominale est quasi-synonyme d'un substantif à base verbo-nominale. Il s'agit de la paire *inzu* / *inyubákwa*.

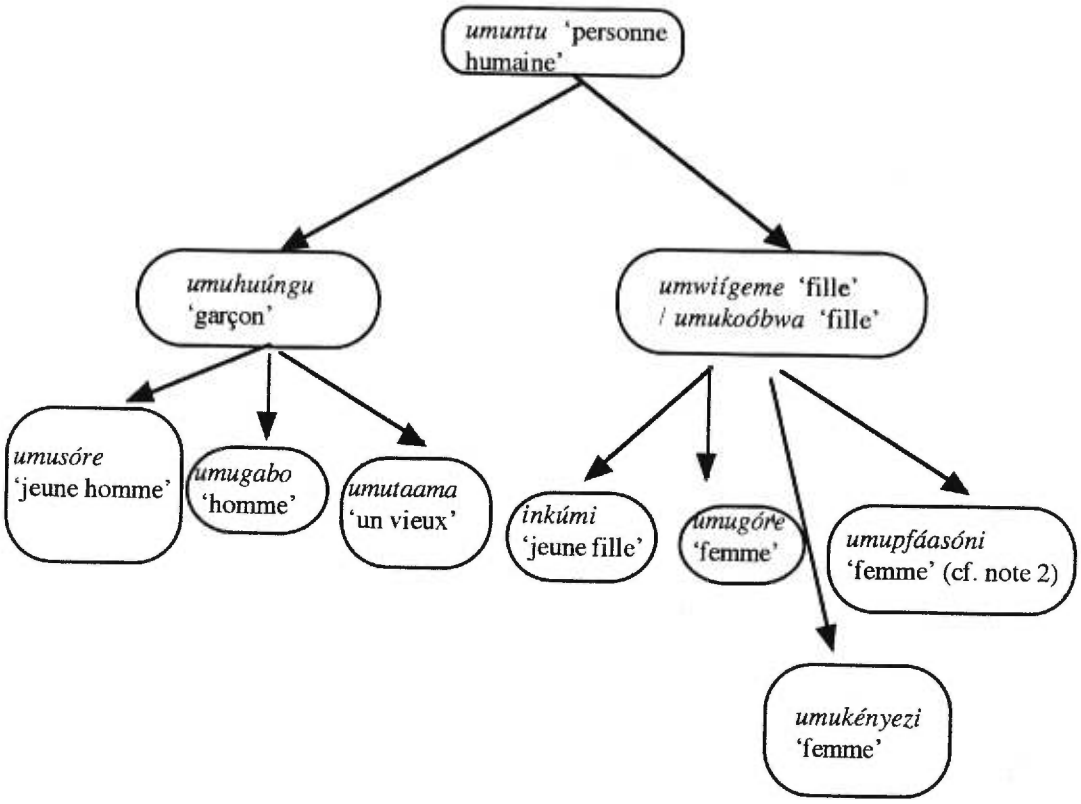
	F_O	D	U	*Glose*
inzu	85	0,78	66,69	'maison'
inyubákwa	18	0,70	15,48	'construction'

Les deux vocables sont substituables dans certains contextes (cf. 158 a, b) et pas dans d'autres (cf. 158 c, d). L'on a par exemple :

(158) a. *ayivugira (imísa) mu nyubákwa y'ábavyéeyi bíiwé* (w7) 'il la dit [la messe] dans la maison de ses parents'
b. *ayivugira (imísa) mu nzu y'ábavyéeyi bíiwé* 'il la dit [la messe] dans la maison de ses parents'
c. *gushika mu nyubákwa ya províinsi ya Bujuumbura* (w3) 'arriver dans le bâtiment de la province de Bujumbura'
d.**gushika mu nzu ya províinsi ya Bujuumbura*

2.1.2.4.2. Les relations d'hyperonymie / hyponymie

Les relations d'hyponymie et d'hyperonymie servent à la construction de relations hiérarchiques entre les vocables. Au sein des vocables dont U est ≥ 3, on peut établir par exemple la hiérarchie suivante[2] :



[2] Dans le schéma, le vocable UMUPFÁASÓNI 'femme' a une connotation " haut niveau social "

Nous ne pouvons inventorier tous les vocables en relations hiérarchiques. Une telle tâche exigerait d'abord de déterminer, par un travail taxonomique, des catégories hyperonymiques puis de proposer des sous-catégories. Nous reviendrons néanmoins au chapitre 5 § 2.4.3.3.2 sur l'utilisation des relations hiérarchiques entre les vocables en didactique des langues.

2.1.3. LES VOCABLES NOMINAUX À BASE VERBO-NOMINALE DONT U ≥ 3

Les vocables nominaux à base verbo-nominale faisant partie du vocabulaire de base du kirundi écrit sont au nombre de 1 371. Ils représentent 6 982 occurrences soit 6% de N. Parmi eux, ceux dont U ≥ 3 sont à 189 soit 13% des vocables nominaux à base verbo-nominale du corpus. Les 189 vocables nominaux à base verbo-nominale représentent 5 003 occurrences soit 71% des occurrences de tous les substantifs à base verbo-nominale; les 29% restants étant couverts par les 1 182 vocables dont U est inférieur à 3. La fréquence moyenne des vocables nominaux à base verbo-nominale dont U ≥ 3 est de 26.

C'est donc dire que les vocables substantifs à base verbo-nominale ont individuellement, en moyenne, une fréquence basse. Nous examinons ici ceux dont U est ≥ 3. Nous distinguons en leur sein des vocables thématiques et des vocables athématiques.

2.1.3.1. LES VOCABLES THÉMATIQUES

Les vocables nominaux à base verbo-nominale thématiques sont des substantifs caractérisés par un indice de fréquence élevé et un indice de dispersion faible par rapport à d'autres vocables de fréquence similaire. L'on pourra contraster les données fréquentielles avec celles sur la dispersion pour se faire une idée de la dispersion normalement attendue pour des vocables de fréquences égales. Nous fournissons la liste des vocables nominaux à base verbo-nominale thématiques pour U ≥ 3 et indiquons entre parenthèses les thèmes auxquels ces vocables sont liés. Le vocable nominal est présenté avec le radical sur lequel il est formé pour faciliter le repérage dans la liste. Pour simplifier la présentation, nous ne présentons que ceux dont F_O est ≥ 26.

Radical et vocable	F_O	*D*	*U*	*Glose*
[-síg-] UMUDASIGÁANA	101	0,70	70,53	'membre du parti Uprona' (politique)
[-seezer-] AMASEEZERANO	80	0,60	47,89	'pacte' / 'accord' (politique)
[-kóran -] IKÓRANIRO	60	0,59	35,70	'assemblée' (politique)
[-rer -] INDERO	48	0,65	31,17	'éducation' (politique)
[-túung-] UBUTÚUNZI	40	0,75	29,97	'richesse' (développement)
[-híinguur-] IHÍINGUURIRO	40	0,69	27,75	'usine' (développement'
[-tóor-] ITÓORA	38	0,57	21,65	'élection' (politique)
[-hung-] IMPUNZI	39	0,51	20,03	'réfugié' (politique)
[bwíiriz-] IBWIÍRIZWA	25	0,65	16,33	'loi' / 'règlement' (politique)

Les vocables nominaux à base verbo-nominale thématiques constituent des indicateurs sur les divers aspects de la vie nationale burundaise telle qu'exprimée à travers les deux journaux *Ndongozi* et *Ubumwe*. L'aspect politique est par exemple marqué par les vocables *umudásigáana* 'membre du parti UPRONA' (parti unique jusqu'en 1993), *amaseezerano* 'pacte' (expression du « pacte d'unité nationale » voté par référendum en 1992), etc.

2.1.3.2. LES VOCABLES ATHÉMATIQUES

Nous subdivisons les vocables nominaux athématiques en vocables généraux et en vocables métadiscursifs.

2.1.3.2.1. Les vocables généraux

Les vocables généraux ne sont à relier à aucun thème spécifique du corpus. Ils sont de première importance dans un vocabulaire de base, dans la mesure où ils se retrouvent généralement dans tous les types de texte.

Compte tenu de la longueur de la liste, nous ne présentons que les vocables nominaux à base verbo-nominale qui ont une fréquence supérieure ou égale à 26,

fréquence moyenne des vocables nominaux à base verbo-nominale dont U ≥ 3 (5 003 / 189). L'on consultera la liste (sur la disquette en annexe) pour le reste des vocables généraux nominaux à base verbo-nominale qui ne sont pas présentés ici. Voici la liste de ceux dont $F_0 \geq 26$.

Radical et vocable	F_O	*D*	*U*	*Glose*
[-náam-] INÁAMA	305	0,78	239,09	'réunion'
[-twáar-] INTWÁARO	193	0,76	147,51	'gouvernement'
[-kór-] IGIKÓRWA	171	0,86	146,49	'travail'
[-kór-] UMUKÓZI	131	0,78	102,83	'travailleur'
[-tégek-] ITÉGEKO	106	0,80	84,44	'loi'
[-aaruk-] URWAARUKA	80	0,78	62,75	'jeunesse' [+collectif]
[-goor-] INGOÓRANE	73	0,70	51,36	'difficulté'
[-íig-] INYÍGIISHO	62	0,81	50,06	'leçon'/'enseignement'
[-gáaniir-] IKIGÁANIIRO	59	0,74	43,85	'causerie'
[-rwáar-] INDWÁARA	57	0,73	41,41	'maladie'
[-gend-] URUGENDO	49	0,79	38,82	'voyage'
[-vúg-] UKUVÚGA	45	0,82	37,02	'fait de parler'/'dire'
[-ror-] IBIRORI	46	0,80	36,88	'spectacle'/'cérémonie'
[-báan-] UMUBÁANO	50	0,74	36,91	'cohabitation'
[-mar-] AKAMARO	46	0,73	33,71	'utilité'
[-ri-] UWURÍ	36	0,81	29,28	'celui qui est'
[-kúund-] URUKÚUNDO	42	0,63	26,65	'amour'
[-ri-] UWARÍ	30	0,83	24,97	'celui qui était'
[-rongoor-] UWURÓONGOOYE	33	0,72	23,75	'celui qui dirige'
[-hanuur-] IMPANUURO	32	0,70	22,37	'conseil'
[-shik-] UMUSHITSI	28	0,78	21,94	'visiteur'
[-áank-] UMWÁANSI	30	0,71	21,28	'ennemi'
[-táangur-] INTÁANGO	26	0,80	20,86	'commencement'
[-seruk-] UWUSÉRUKIRA	27	0,73	19,71	'celui qui représente *x*'
[-íikuukir-] UKWIÍKUUKIRA	29	0,67	19,44	'indépendance'
[-z-] AKAZÓOZA	27	0,68	18,31	'le futur'

[-íig-] UMWIÍGIISHA	24	0,69	16,57	'un enseignant'
[-íipfuuz-] ICÍIPFUUZO	22	0,73	16,14	'souhait'

On le voit, la liste des vocables nominaux à base verbo-nominale généraux dont U ≥ 3 est fournie. Ils sont importants en didactique de la langue car ils ne sont liés à aucun thème particulier et ont, de ce fait, des chances de se retrouver dans de nombreux textes.

2.2.3.2.2. Les vocables métadiscursifs

Les vocables métadiscursifs sont des vocables qui servent à parler du discours, des idées exprimées, de la façon de les exprimer, etc. Parmi les vocables nominaux à base verbo-nominale, on retrouve les suivants :

Radical et vocable	F_O	*D*	*U*	*Glose*
[-báz-] IKIBÁZO	160	0,83	132,52	'question'
[-túm-] IGITÚMA	133	0,90	119,95	'le pourquoi'
[-éerek-] IVYEÉREKEYE	132	0,82	108,23	'ce qui concerne *x*'
[-íiyumviir-] ICÍIYUMVIIRO	104	0,77	80,58	'idée'
[-raab-] IBIRÁABA	17	0,73	12,49	'ce qui concerne'
[-raab-] UKURAABA	6	0,65	3,89	'analyser' / 'regarder'

Les vocables métadiscursifs servent en didactique des langues notamment à exprimer certaines requêtes des apprenants et des enseignants, à formuler le libellé des questions et réponses, etc. Nous y reviendrons au chapitre 5 § 3.2.2. où la liste des vocables métalinguistiques sera enrichie par des vocables appartenant à d'autres catégories de vocables ou aux tranches dont les vocables connaissent un indice d'usage inférieur à 3.

2.1.3.3. VOCABLES NOMINAUX À BASE VERBO-NOMINALE ET RADICAUX VERBAUX

Lorsque l'on veut comprendre les rapports entre les substantifs à base verbo-nominale et les verbes construits sur un même radical, une question se pose. Tous ces substantifs à base verbo-nominale correspondent-ils à des radicaux verbaux fréquents qui figurent comme tels dans notre liste de vocable dont U ≥ 3 ?

La réponse est négative. Il y a dans notre liste, des substantifs à base verbo-nominale dont U ≥ 3 mais qui correspondent à des radicaux verbaux dont U est inférieur à 3. Ils ne sont cependant pas nombreux. Il s'agit des 15 substantifs suivants, ordonnés à partir de l'indice d'usage :

Vocable et radical	F_O	*D*	*U*	*Glose*
[-náam-] INÁAMA	305	0,78	239,09	'réunion'
[-aaruk-] URWAARUKA	80	0,78	62,75	'jeunesse' [+collectif]
[-gáaniir-] IKIGÁANIIRO	59	0,74	43,85	'causerie'
[-híinguur-] IHÍINGUURIRO	40	0,60	27,75	'usine'
[-cíir-] IGICÍIRO	18	0,68	12,19	'prix'/'valeur'
[-báash-] UBUBAÁSHA	17	0,68	11,48	'autorité'/'pouvoir'
[-óoroor-] UBWÓOROOZI	14	0,55	7,64	'élevage'
[-áamuuk-] AMÁAMUUKO	10	0,69	6,90	'origine'
[-gir-] INGIRO	11	0,60	6,56	'façon de faire'
[-búuran-] UMUBÚURANYI	14	0,46	6,41	'un plaignant'
[-cíir-] AGACÍIRO	9	0,71	6,41	'valeur'
[-voom-] IVOOMO	8	0,65	5,20	'source'
[-háruur-] IBIHÁRUURO	12	0,43	5,11	'calcul'
[-zírikan-] UMUZÍRIKANYI	8	0,62	4,97	'coeur
[-rog-] UMUROZI	15	0,31	4,58	'jeteur de mauvais sort'

Parmi les 189 vocables nominaux à base verbo-nominale dont U ≥ 3, il y en a donc 174 qui correspondent à des radicaux verbaux dont U ≥ 3 soit 92%. On peut donc

affirmer que, généralement, les substantifs à base verbo-nominale à haut indice d'usage correspondent à des radicaux verbaux fréquents.

Sur le plan pédagogique, cette information est importante. Si l'on se situe dans la morphologie du morphème (Corbin 1987 : 182-183) qui postule que les mots sont formés par concaténation, l'on peut estimer qu'il est plus économique pour l'apprenant d'apprendre des radicaux et des affixes plus fréquents et de former des mots en appliquant une règle que d'apprendre plusieurs mots différents.

Ainsi, un apprenant qui connaîtrait le radical [-rim-] 'cultiver', le suffixe agentif [-yi] et le préfixe de classe [-mu-] pourrait former *umurimyi* 'cultivateur' au lieu d'apprendre ce mot-forme prévisible comme un vocable à part.

Aussi, au sein des 189 substantifs à base verbo-nominale, on constate qu'il y en a plusieurs qui sont formés sur un même radical. L'on consultera l'index alphabétique des vocables (sur disquette) pour en avoir une idée exhaustive. Illustrons cependant le cas par le radical [-rim-] 'cultiver' qui entre dans la formation des substantifs suivants dont $U \geq 3$:

Vocable	F_O	*D*	*U*	*Glose*
UMURÍMA	26	0,72	18,69	'champ'
UMURIMYI	25	0,68	17,04	'cultivateur'
UBURIMYI	24	0,52	12,37	'agriculture'
UMURIMO	16	0,75	11,94	'travail'

Le fait que certains radicaux servent à former plusieurs substantifs contribue à accroître la facilité de reconnaissance des mots-formes. Cela devrait faciliter notamment la lecture (parce que les éléments formels à reconnaître sont redondants) et l'accès au sens du texte (parce que les dérivés partagent une composante sémantique commune).

2.1.4. LES VOCABLES ADJECTIVAUX DONT U ≥ 3

Les vocables adjectivaux de notre corpus sont au nombre de 77. Ils représentent 1 549 occurrences soit 1 % de N. De ces 77 vocables, 24 ont un indice d'usage supérieur ou égal à 3. Les 24 vocables représentent 1 330 occurrences soit 85% des occurrences des adjectifs du corpus. La fréquence moyenne des adjectifs dont U ≥ 3 est de 55, une fréquence moyenne relativement élevée si on la compare, pour la même tranche à celle des vocables verbaux (61), nominaux à base verbo-nominale (26) et nominaux à base verbo-nominale (51).

Le nombre de vocables adjectivaux est peu élevé. Nous montrons au § 4.2. que la langue dispose d'autres moyens pour exprimer les propriétés des personnes et des objets. Nous fournissons l'inventaire des 24 vocables adjectivaux dont U ≥ 3, vocables qui font partie du noyau du vocabulaire de base du kirundi écrit.

Radical	F_O	*D*	*U*	*Glose*
[-IÍNSHI]	386	0,89	341,99	'en grande quantité'/'nombreux'
[-KÚRU]	171	0,87	149,51	'grand'
[-KÉ]	125	0,85	106,34	'en petite quantité'
[-IIZA]	98	0,88	86,39	'beau'
[-SHAÁSHA]	66	0,73	48,18	'nouveau'
RUSAÁNGI	62	0,55	33,93	'démocratique'
[-BÍ]	41	0,81	33,04	'mauvais'
[-NÍNI]	36	0,80	28,83	'grand'
[-TÓ]	33	0,82	27,02	'petit'
KAMINÚUZA	52	0,51	26,38	'de très haut niveau'
[-KÚRU...KÚRU]	29	0,74	21,47	'essentiel'
[-TÓOYI]	21	0,73	15,27	'petit'
[-TÓ...TÓ]	21	0,69	14,50	'de moindre importance'
GATOORIKÁ	31	0,39	12,10	'catholique'
[-KÉEYÍ]	14	0,71	9,97	'en petite quantité'
MPUUZAMAKUNGU	23	0,34	7,92	'international'
[-EERANDA]	20	0,36	7,16	'saint'
NYAKÚRI	22	0,32	7,13	'véritable'
RUKRISTÚ	25	0,28	6,96	'chrétien'

NGENDÉRWAKÓ	21	0,30	6,29	'fondamental'
[-HIÍRE]	9	0,53	4,75	'heureux'
[-SHÁ]	9	0,52	4,70	'nouveau'
[-KÉEYÁ]	6	0,51	3,06	'en petite quantité'

De ces données, on peut extraire les oppositions lexicales suivantes :

(159) [-iínshi] *versus* [-ké] / [-kéeyi]
'en grande quantité'/'nombreux' 'en petite quantité'

[-kúru] / [-níni] *versus* [-tó] / [-tóoyi]
'grand' 'petit'

[-iizá] *versus* [-bí]
'beau' 'mauvais'

Les formes longues des radicaux adjectivaux (*-kéeyi, -tóoyi*) sont les variantes lexicales respectives de [-ké] 'en petite quantité' et de [-tó] 'petit'.

Quant à la paire de quasi-synonymes [-kúru] / [-níni] 'grand', seule une étude syntactico-sémantique permettrait de mieux cerner leur distribution.

La liste des vocables adjectifs dont U ≥ 3 contient également des adjectifs invariables qui ont surtout cours dans des textes à thématique politique. On les reconnaît dans la liste ci-dessus par la glose et par l'abscence de crochets autour du vocable.

2.1.5. LES VOCABLES GRAMMATICAUX DONT U ≥ 3

Les vocables grammaticaux de notre corpus sont au nombre de 608. Parmi eux, 447 sont flexionnels et 161 non flexionnels. Les vocables grammaticaux flexionnels représentent 19 301 occurrences soit 45% des mots-formes grammaticaux et 18% de N. Quant aux 161 vocables grammaticaux non fléchis, ils représentent 22 726

occurrences, soit 53% des occurrences des mots-formes grammaticaux et 21% de N. Nous les abordons dans cet ordre.

2.1.5.1. LES VOCABLES GRAMMATICAUX FLEXIONNELS

Les vocables grammaticaux flexionnels dont U ≥ 3 comprennent des allocutifs, des démonstratifs, des possessifs, des connectifs, des numéraux, des substitutifs et des interrogatifs. Rappelons que les unités appartenant à ces catégories connaissent des emplois pronominaux ou déterminatifs. Signalons aussi que quelques adverbes figurent au sein des vocables grammaticaux fléchis.

Notre étude n'étant pas syntaxique, nous n'avons pas fait de distinction entre des mots grammaticaux homographes appartenant à une même catégorie (celle des démonstratifs par exemple) mais ayant des emplois différents (pronominal ou déterminatif).

Voici comment se répartissent tous les vocables grammaticaux flexionnels du corpus selon U et selon le nombre de mots-formes :

	U≥3	*3<U≥ 0*	*U< 0*	**Total**
V	221	182	44	**447**
N	18 768	482	51	**19 301**
$\overline{X}$	84,92	2,64	1,15	**43,17**
σ	173,10	1,52	0,42	**128,47**

Tableau 18 - *Les vocables grammaticaux flexionnels selon* U *et* N

Au sein des 447 vocables grammaticaux flexionnels, ceux dont U ≥ 3 sont au nombre de 221. Nous résumons dans le tableau ci-dessous la représentation de chacune des sous-catégories morphosyntaxiques :

Classes de V grammaticaux	*V* $U \geq 3$	*N*	$\overline{X}$	σ
Connectifs	29	6 816	235,03	370,00
Démonstratifs	37	4 107	111,00	153,90
Indéfinis	37	1 961	53,00	67,86
Numéraux	37	1 799	48,62	52,69
Locatifs	5	1 846	369,20	196,28
Substitutifs	13	1 059	81,46	99,90
Possessifs	39	746	19,10	17,30
Allocutifs	6	204	34,00	8,17
Adverbes	14	478	34,14	55,43
Interrogatifs	3	85	28,33	34,44
Interpellatifs	1	10	10,00	0,00
Total	**221**	**19 160**	**86,30**	**173,94**

Tableau 19 - *Les classes des vocables grammaticaux flexionnels dont* $U \geq 3$

Comme on le voit, les 221 vocables grammaticaux flexionnels dont U ≥ 3 représentent 19 160 occurrences soit 99% (19 160 / 19 301) de toutes les occurrences des vocables grammaticaux fléchis; 1% étant comblé par les 226 vocables restants. L'on comprendra le rôle primordial que jouent ces unités en didactique de la langue. Nous présentons ces vocables selon leur sous-catégories morphosyntaxiques.

2.1.5.1.1. Connectifs

Les 29 connectifs dont U ≥ 3 représentent 6 816 occurrences soit 36% des occurrences des mots-formes grammaticaux fléchis. C'est dire leur important rôle syntaxique en kirundi.

Les connectifs correspondent à toutes les unités qui, dans la liste ci-dessous, ont la glose 'de'. Rappelons que nous les avons décrits au chapitre 3 § 3.3.1.2.3. et 3.3.1.2.4. Les connectifs sont au nombre de 64 soit 32 en emploi déterminatif et 32 en emploi pronominal. Ils sont construits sur les morphèmes [-a] ou [-ó]. Les connectifs en emploi pronominal diffèrent des connectifs en emploi déterminatif par une voyelle initiale.

Parmi les connectifs dont U ≥ 3, ceux construits sur [-a] ont des fréquences plus élevées que ceux construits sur [-ó] comme on peut le voir dans la liste que nous fournissons ci-dessous.

Signalons que le discriminant morphosyntaxique entre parenthèses (conn.) sert à lever une homographie avec d'autres vocables grammaticaux. Les gloses sont liées aux classes d'accord des vocables; c'est pourquoi elles sont tantôt au singulier tantôt au pluriel. Les gloses au singulier indiquent par exemple que le vocable connectif prend cette forme lorsqu'il connaît un emploi déterminatif avec un substantif qui appartient à la classe d'accord indiquée. Ainsi le vocable connectif *wa* est en emploi déterminatif avec des vocables nominaux dont les classes d'accord sont 3 et 4 tel qu'illustré ci-dessous :

(156) *umugabo wa* x 'le mari de *x*' cl.1.
umuríma wa x 'le champ de *x*' cl.3

Présentons maintenant la liste des vocables grammaticaux connectifs dont U ≥3. Il s'agit des vocables suivants :

Vocable	F_O	*D*	*U*	*Glose*
YA	1 579	0,93	1 474,57	'celui de' cl.4, 5, 9
WA	1 331	0,91	1 214,79	'celui de' cl.1, 3
RYAA	570	0,88	502,73	'celui de' cl.4
BAA	438	0,90	393,74	'ceux de' cl.2
CAA	399	0,91	364,59	'celui de' cl.7
BÓ	365	0,94	343,12	'ceux de' cl.7
VYA	338	0,91	306,97	'ceux de' cl.8

BWAA	259	0,90	230,46	'celui de' cl.14
ZAA	214	0,87	187,01	'ceux de' cl.10
YÓ (conn.)	156	0,90	140,24	'celui de' cl.4, 5, 9
KAA	150	0,92	137,28	'celui de' cl.12
WÓ	142	0,90	127,69	'de' cl.1,3
RWAA	128	0,93	118,74	'celui de' cl.11
KWAA	129	0,86	110,97	'celui de' cl.15
IVYA	76	0,83	62,89	'ceux de' cl.8
UWA	73	0,78	57,19	'celui de' cl.1, 3
BWÓ (conn.)	78	0,71	55,26	'celui de' cl.14
UBWAA	65	0,84	54,79	'celui de' cl.14
RYÓ (conn.)	61	0,82	49,91	'celui de' cl.4
ZÓ (conn.)	53	0,82	43,42	'ceux de' cl.10
VYÓ (conn.)	50	0,85	42,43	'ceux de' cl.8
CÓ (conn.)	32	0,82	26,34	'celui de' cl.7
IYA	28	0,75	20,87	'celui de' cl.4, 5, 9
ICAA	26	0,80	20,72	'celui de' cl.7
TWAA	19	0,83	15,77	'ceux de' cl.13
RWÓ (conn.)	18	0,78	13,97	'celui de' cl.11
IZAA	18	0,75	13,45	'ceux de' cl.10
AKAA (conn.)	13	0,80	10,36	'celui de' cl.12
URWAA	8	0,71	5,69	'de' cl.11

Ainsi sur 29 connectifs dont U ≥ 3, seuls 10 sont formés sur [-ó] alors que 19 sont construits sur [-a]. Cette sur-représentation des vocables connectifs construits sur [-a] est encore plus évidente pour les connectifs en emploi pronominal comme l'illustre la liste ci-dessous :

Vocable	F_O	*D*	*U*	*Glose*
IVYAA	76	0,83	62,89	'ceux de' cl.9
UWA	73	0,78	57,19	'celui de' cl.1, 3
UBWAA	65	0,84	54,79	'celui de' cl.14

ICAA	26	0,80	21,23	'celui de' cl.7
IYA	28	0,75	20,87	'celui de' cl.4, 9
IZAA	18	0,75	13,45	'ceux de' cl.10
AKAA (conn.)	13	0,80	10,36	'celui de' cl.12
URWAA	8	0,71	5,69	'celui de' cl.11

Comme on le voit, aucun connectif en emploi pronominal construit sur [-ó] ne figure dans la liste. On peut donc affirmer d'une part que les connectifs en emploi déterminatif ont des fréquences plus élevées que les correspondants en emploi pronominal; et d'autre part que le radical [-a] forme des connectifs plus fréquents que ceux formés sur le radical [-ó].

2.1.5.1.2. Démonstratifs

Les démonstratifs dont U ≥ 3 sont au nombre de 37. Ils représentent 4 107 occurrences soit 21% de toutes les occurrences des vocables grammaticaux flexionnnels.

Rappelons que nous avons distingué cinq groupes parmi les démonstratifs selon la distance qu'ils expriment (cf. chapitre 3 § 3.3.1.1.1). Dans la liste des vocables, nous marquons ces groupes par des chiffres romains. Nous indiquons à côté de la glose la classe d'accord.

Vocable	F_O	*D*	*U*	*Glose*	*Type de démonstratif*
AHO	556	0,93	515,03	'là' cl.16	II
IVYO	468	0,93	432,98	'ceux-là' cl.8	II
ICO	425	0,94	498,70	'celui-là' cl.7	II
UKO	453	0,94	424,07	'celui-là' cl.15	II
IYO (dém)	365	0,88	321,63	'celui-là' cl.4, 5 & 9	II
ABO	300	0,93	278,26	'ceux-là' cl.2	II
UWO	282	0,91	256,46	'celui-là' cl.1 & 3	II

IRYO	158	0,82	129,93	'celui-là' cl.5	II
AYO	111	0,87	96,05	'ceux-là' cl.6	II
UYO	97	0,86	83,15	'celui-là' cl.1 & 3	II
IRI	88	0,88	77,20	'celui-ci' cl.5	I
IYI	80	0,88	70,77	'celui-ci' cl.4, 5 & 9	I
UBWO (dém.)	58	0,81	46,93	'celui-là' cl.14	II
UYU	76	0,82	62,45	'celui-ci' cl.1 & 3	I
IKI (dém.)	64	0,83	53,10	'celui-ci' cl.7	I
IBI	46	0,83	38,36	'ceux-ci' cl.8	I
IZO (dém.)	55	0,82	45,06	'ceux-là' cl.10	II
NGAAHÁ	52	0,79	41,08	'ici' cl.16	déictique
URWO (dém.)	53	0,86	45,42	'celui-là' cl.11	II
AHA	41	0,86	35,12	'celui-ci' cl.16	I
HÁNO	37	0,71	26,10	'ce lieu dont on parle ' cl.16	V
UNÓ	30	0,72	21,60	'*x* dont on parle' cl.1 & 3	V
AKO	33	0,73	24,09	'celui-là' cl.12	II
IRYÁ	22	0,69	15,21	'celui-là' cl.5 & 9	III
NGAÁHO	25	0,66	16,46	'là' cl.16	déictique
HÁRIÍYA	17	0,66	11,24	'cela' cl.16	IV
AYA	22	0,67	14,75	'ceux-ci' cl.6	I
HÁRYA	17	0,52	8,78	'celui-là' cl.16	III
BÍRYA	12	0,56	6,67	'ceux-là' cl.8	III
BÁRYA	11	0,67	7,33	'ceux-là' cl.2	III
URYÁ	8	0,63	5,02	'celui-là' cl.1 & 3	III
KURYÁ	8	0,60	4,78	'celui-là' cl.15	II
UKWO (dém.)	8	0,69	5,56	'celui-là' cl.14	II
AKA	6	0,77	4,65	'celui-ci' cl.12	I
KÍRYA	4	0,49	3,01	'celui-là' cl.7	II
URU	3	0,56	3,33	'celui-ci' cl.11	I
UKU	3	0,58	3,74	'celui-ci' cl.15	I

Un premier coup d'œil sur le tableau permet de constater que les démonstratifs du groupe II sont les plus fréquents (17 / 37) suivis de ceux du groupe I (10 / 37). Les premiers expriment la proximité de l'objet par rapport au locuteur alors que les deuxièmes expriment un éloignement minime. Les démonstratifs exprimant un grand éloignement (type III, IV et V) sont nettement moins fréquents (respectivement 5 / 37, 1 / 37 et 2 / 37). Les démonstratifs à sens déictique sont peu nombreux : 2 sur 37.

Signalons une homonymie pour les démonstratifs des classes 1 & 3 et ceux des classes 4, 5 & 9 et la concurrence des formes démonstratives suivantes :

(157) *uwo* ($F_O = 282$) versus *uyo* ($F_O = 97$)
uko ($F_O = 453$) versus *ukwo* ($F_O = 8$)

2.1.5.1.3. Allocutifs

Le kirundi distingue deux types d'allocutifs : ceux à forme courte et ceux à forme longue. Nous les avons décrits au chapitre 3, § 3.3.1.3.1.

Parmi les allocutifs courts du kirundi, seuls deux d'entre eux ont un indice d'usage ≥ 3. Ce sont les suivants :

Vocable	F_O	*D*	*U*	*Glose*
MWÉ	38	0,83	31,51	'vous'
TWÉ	25	0,78	19,41	'nous'

Par contraste, quatre des six allocutifs longs du kirundi ont U ≥ 3. Ils sont repris ci-dessous :

Vocable	F_O	*D*	*U*	*Glose*
TWEEBWÉ	48	0,80	38,34	'nous'
MWEEBWÉ	32	0,80	25,46	'vous'
JEEWÉ	33	0,75	24,63	'moi'
WEWÉ	28	0,67	18,64	'toi'

2.1.5.1.4. Possessifs

Les possessifs sont au nombre de 192 soit 96 en emploi pronominal et 96 en emploi déterminatif. Ils ne sont pas homographes. Nous présentons d'abord ceux qui connaissent des déterminatifs. Ils sont au nombre de 35 sur 40 possessifs dont U ≥ 3, ce qui représente 87,5%. Il y a donc clairement une dominance des possessifs en emploi déterminatif. Ce sont les suivants :

Vocable	F_O	*D*	*U*	*Glose*
CÁACU	84	0,76	63,72	'notre' cl.7
WÍIWÉ	62	0,74	45,89	'sien' cl.1 & 3
YÁABO	46	0,87	40,19	'leur' cl.4, 5 & 9
WÁACU	56	0,66	37,12	'notre' cl.1 & 3
VYÁABO	31	0,84	26,16	'leur' cl.8
YÁACU	33	0,71	23,49	'notre' cl.4, 6 & 9
BWÁABO	32	0,72	23,17	'leur' cl. 14
YÍIWE	30	0,75	22,39	'sien' cl.4, 5 & 9
BÍIWE	26	0,83	21,54	'sien' cl.2
BÁABO	26	0,83	21,49	'leur' cl.2
BWÍIWÉ	22	0,74	16,22	'sien' cl.14
WÁABO	21	0,68	14,20	'leur' cl.1 & 3
ZÁABO	19	0,73	13,86	'leur' cl.10
BÁACU	17	0,75	12,83	'notre' cl.2
VYÁACU	20	0,65	12,91	'notre' cl.8
CÍIWE	15	0,76	11,36	'sien' cl.7
RYÁABO	14	0,72	10,02	'leur' cl.11
UTWAÁBO	13	0,69	8,97	'les leur' cl.12
BÁAHO	12	0,74	8,94	' appartenant à ce lieu' cl. 12
KÁABO	12	0,69	8,28	'leur' cl.12
CÁABO	15	0,54	8,06	'leur' cl.7
RYÍIWÉ	11	0,73	8,05	'sien' cl.5

BWÁACU	12	0,64	7,69	'notre' cl.14
WÁAYO	12	0,61	7,36	'sien' cl.1 & 3
KÍIWÉ	8	0,72	5,76	'sien' cl.12
YÁANYU	8	0,71	5,65	'votre' cl.1 & 3
YÁAYO	8	0,65	5,19	'sien' cl.5
VYÍIWÉ	7	0,59	4,12	'sien' cl.8
RWÁAYO	6	0,66	3,93	'sien' cl.11
WANJE	8	0,49	3,90	'mien' cl.1 & 3
RYÁACU	7	0,52	3,61	'notre' cl.11
RWÁABO	6	0,60	3,59	'leur' cl.11
KWÁABO	7	0,50	3,52	'leur' cl.15
ZÍIWÉ	7	0,49	3,45	'sien' cl.11
WAAWE	6	0,55	3,29	'tien' cl.1 & 3

Quant aux possesssifs qui connaissent des emplois pronominaux, seuls 4 d'entre eux ont U ≥ 3. Ce sont :

Vocable	F_O	*D*	*U*	*Glose*
UBWÍIWÉ	12	0,63	7,54	'les siens' cl.14
UTWAÁBO	13	0,69	8,97	'les leurs' cl.13
ABAÁBO	8	0,57	4,58	'les leurs' cl.2
ABÍIWÉ	6	0,51	3,05	'les leurs' cl.10

Il est frappant de constater que les possessifs dont U ≥ 3 connaissent presque exclusivement des emplois déterminatifs. Rappelons ici que leurs correspondants pronominaux ne leurs sont pas homographes et commencent par une des voyelles augments (a-, i-, ou u-) tels que décrits au chapitre 3 § 3.3.1.2.2.

Compte tenu de l'importance numérique des emplois déterminatifs des mots grammaticaux fléchis possessif, nous croyons qu'une pédagogie du kirundi devrait les prioriser par rapport aux emplois pronominaux.

2.1.5.1.5. Locatifs

La classe des mots-formes locatifs est limitée en kirundi à quatre éléments : *kó, hó, mwó, yó*. Ils ont tous un indice d'usage supérieur à 3. Le discriminant entre parenthèses permet de lever l'homographie avec d'autres vocables de notre corpus.

Vocable	F_O	*D*	*U*	*Glose*
hó	566	0,94	533,98	'y'
kó (loc.)	464	0,94	434,25	'sur'
mwó	380	0,95	360,13	'dans'
yó (loc.)	44	0,69	30,40	'en'

2.1.5.1.6. Indéfinis

Nous avons identifié trois morphèmes qui servent à la formations des indéfinis en kirundi : [-óóse] 'tous', [-mwé] ' certains' et [-ndi] 'autre' (cf. chapitre 3, § 3.3.1.1.2). Nous présentons ceux d'entre eux dont U ≥ 3 en séparant les données sur les trois radicaux par une ligne grasse.

Vocable	F_O	*D*	*U*	*Glose*
BÓÓSE	285	0,88	251,50	'tous'cl.2
YÓÓSE	173	0,88	151,46	'tous' cl.4, 5 & 9
VYÓÓSE	156	0,88	137,33	'tous'cl.8
WÉÉSE	90	0,87	78,51	'tout' cl.1 & 3
HÓÓSE	61	0,83	50,67	'partout'cl.16
TWÉÉSE	45	0,80	35,81	'nous tous'
ZÓÓSE	38	0,81	30,90	'tous' cl.10

CÓÓSE	32	0,81	25,80	'tout' cl.8
MWÉÉSE	31	0,81	25,16	'vous tous'
BWÓÓSE	32	0,77	24,58	'tout' cl.14
KWÓÓSE	14	0,72	10,06	'tout' cl.15
WÓÓSE	11	0,75	9,73	'tout' cl.1 & 3
KÓÓSE	11	0,52	5,75	'tout' cl.15
YÓÓMPI	8	0,50	4,00	'tous ensemble' cl.4
ABAÁNDI	266	0,92	245,64	'autres'cl.2
IBIÍNDI	129	0,85	109,99	'autres'cl.8
IKIÍNDI	73	0,86	63,11	'autres'cl.7
UWUÚNDI	84	0,65	54,82	'autre'cl.1 & 3
AYAÁNDI	58	0,82	47,68	'autres'cl.5
AHAÁNDI	46	0,83	38,34	'autre'cl.16
IYIÍNDI	45	0,83	37,40	'autre'cl.4, 5 & 9
IZIÍNDI	33	0,85	28,04	'autres'cl.10
UBUÚNDI	15	0,79	11,89	'autres'cl.14
UKUÚNDI	12	0,52	6,22	'autre' cl.15
IRIÍNDI	8	0,65	5,24	'autre'cl.5
RÍINDI	7	0,70	4,91	'autre'cl.5
AKAÁNDI	8	0,60	4,82	'autre'cl.12
UTUÚNDI	7	0,52	3,63	'autres'cl.13
BAMWÉ	63	0,89	56,13	'certains' cl.2
BAMWÉEBAMWÉ	38	0,75	28,61	'certains' cl.2
UMWÚUMWÉ	17	0,76	12,92	'l'un l'autre' cl.1 & 3
BIMWÉEBIMWÉ	17	0,70	11,87	'certains' cl.8
AMWÁAMWÉ	12	0,67	8,04	'certains' cl.6
ZIMWÉEZIMWÉ	7	0,47	3,28	'certains' cl.10

Dans l'ensemble, la catégorie des indéfinis est dominée par ceux construits sur [-ndi] 'autre' (13 / 37 vocables). Viennent ensuite ceux construits sur [-óóse] 'tous' (13 / 37 vocables), sur [-mwé]' 'certains' (7 / 37 vocables) et enfin sur [-óómpi] 'tous' (1 / 37 vocables).

Les indéfinis sont homonymes en emploi pronominal ou déterminatif. Il s'agit d'une homonymie fonctionnelle qui, rappelons-le, touche les vocables de même forme mais dont les fonctions syntaxiques sont différentes. À cette homonymie syntaxique s'ajoute une homonymie morphologique liée à la classe du substantif avec lequel ils s'accordent. Les classes concernées sont les classes 4, 5 & 9 et 1 & 3.

Signalons la présence dans la liste des indéfinis de deux vocables qui ne sont pas construits sur les radicaux identifiés ci-dessus. Il s'agit de KAANAAKÁ 'un tel' et NAAKÁ 'tel'.

2.1.5.1.7. Numéraux cardinaux

Nous avons identifé au chapitre 3, § 3.3.1.3.4, six radicaux (de 1 à 6) exprimant le numéral cardinal en kirundi. Les fréquences des numéraux cardinaux nous semblent liées notamment à la présence de nombreuses dates dans le corpus, dates que nous avons réécrites au long. Un corpus moins marqué du point de vue des datations pourrait être un meilleur indicateur de l'usage des numéraux en kirundi.

Il reste cependant intéressant de vérifier si un radical numéral domine sensiblement les autres. En additionnat les fréquences pour chaque radical, on obtient les résultats suivants :

F_O	*Radical*	*Glose*
500	[-biri]	'deux'
448	[-mwé]	'un'
291	[-tatu]	'trois'
243	[-taanu]	'cinq'
150	[-taandátu]	'six'
17	[-né]	'quatre'

Les radicaux correspondant aux trois premiers chiffres de la numération sont les plus fréquents. Il nous est difficile d'interpréter les fréquences de ces radicaux sans déborder le cadre de la linguistique.

On peut cependant faire remarquer que ces fréquences rejoignent pour 'deux' et 'trois' celles fournies par Greenberg (1963 : 42-43) pour l'anglais, l'espagnol, le français et l'allemand.

Rappelons que les vocables exprimant 'mille' 'neuf' 'cent' et 'dizaine' figurent dans la liste des vocables nominaux à base nominale (cf. § 2.1.3.1.).

2.1.5.1.8. Substitutifs

Nous avons distingué parmi les pronoms substitutifs ceux à forme courte et ceux à forme longue. Ces derniers ont des fréquences très basses; un seul (*yóoyó* 'lui' cl.4, 5 & 9) a un indice d'usage supérieur à 3. Nous présentons donc presque exclusivement des pronoms substitutifs à forme courte, qui, rappelons-le, sont au nombre de 16. Le discriminant (subst.) distingue les vocables d'autres qui leur sont homographes parmi les sous-catégories morphosyntaxiques.

Vocable	F_O	*D*	*U*	*Glose*
BÓ	365	0,94	343,12	'eux' cl.2
WÉ	302	0,81	243,42	'lui' cl.1
CÓ (subst.)	229	0,91	209,35	'lui' cl.7
YÓ (subst.)	166	0,88	146,74	'lui' cl.4, 5 & 9
VYÓ (subst.)	155	0,90	138,77	'lui' cl.8
KÓ (substit.)	92	0,81	74,90	'lui' cl.15
RYÓ (subst.)	37	0,78	28,70	'lui' cl.5
BWÓ (subst.)	21	0,59	12,36	'lui' cl.14
ZÓ (subst.)	16	0,75	11,97	'lui' cl.10
UTWÓ (subst.)	11	0,72	7,94	'lui' cl.12
RWÓ (subst.)	11	0,59	6,53	'lui' cl.11
KWÓ	8	0,55	4,41	'lui' cl.14
YÓOYÓ	6	0,57	3,44	'lui' cl.4, 5 & 9

2.1.5.1.9. Adverbes

Certains adverbes du kirundi sont flexionnels. Ceux dont U ≥ 3 sont les suivants :

Vocable	F_O	*D*	*U*	*Glose*
HAMWÉ	221	0,90	199,98	'ensemble'
GÚTE	42	0,75	31,47	'comment ?'
KUMWÉ (adv.)	32	0,78	25,09	'ensemble'
NKÓ	32	0,75	24,09	'comme'
GÚRTYO	32	0,72	23,19	'comme cela / ainsi'
IWAÁCU	15	0,69	10,37	'chez nous'
IWABO	10	0,69	7,11	'chez eux'
BÓONYÈNE	10	0,72	7,16	'eux seuls' cl.2
RIMWÉ NA RIMWÉ	7	0,71	4,97	'des fois'
YÓONYÉNE	6	0,70	4,23	'lui seul' cl.4, 5 & 9
WÉENYÉNE	6	0,65	3,89	'lui seul' cl.1
IWÉ	6	0,57	3,43	'chez lui'

À part ces adverbes flexionnels, le kirundi connaît d'autres adverbes non flexionnels. Ils constituent une catégorie numériquement importante parmi les vocables grammaticaux non flexionnels. Nous y reviendrons.

Pour l'instant, résumons. Les vocables grammaticaux flexionnels sont dominés par six sous-catégories morphosyntaxiques qui sont dans l'ordre : les connectifs, les démonstratifs, les indéfinis, les numéraux, les locatifs et les substitutifs.

Les connectifs les plus fréquents sont ceux construits sur le radical [a] alors que les démonstratifs sont essentiellement ceux de types 1 et 2. Les connectifs en emploi déterminatif sont de loin plus fréquents que ceux en emploi pronominal. Les indéfinis les plus fréquents sont construits sur les radicaux [-ndi] 'autre' et [-óóse] 'tous'. Les

numéraux sont dominés par les trois premiers chiffres. Les substitutifs les plus fréquents sont des substitutifs courts. Il en est de même des allocutifs. Les locatifs ont tous un indice d'usage élevé. Signalons enfin que les possessifs dont U ≥ 3 connaissent presque exclusivement des emplois déterminatifs.

2.1.5.2. LES VOCABLES GRAMMATICAUX NON FLEXIONNELS

Les vocables grammaticaux non flexionnels sont au nombre de 161. Ils représentent 22 726 occurrences soit 53% des mots-formes grammaticaux du corpus.

Ils appartiennent aux catégories des prépositions, des prédicatifs nominaux, des conjonctions et des adverbes. Nous y avons aussi retenu une classe de vocables divers composés d'interjections et d'onomatopées.

Les vocables grammaticaux non flexionnels dont U ≥ 3 sont au nombre de 99 et représentent 22 591 occurrences soit 99% de toutes les occurrences des vocables grammaticaux non flexionnels. Ils représentent donc, en termes d'occurrences, l'essentiel des mots-formes grammaticaux non fléchis. L'on comprendra donc que nous les privilégions dans notre analyse.

Voici comment se répartissent tous les vocables grammaticaux non flexionnels du corpus selon U et selon le nombre de mots-formes :

	U≥3	*3>U≥ 0*	*U< 0*	**Total**
V	99,00	47,00	15	**161,00**
N	22 591,00	120,00	15	**22 726,00**
$\overline{X}$	228,19	2,55	1	**141,15**
σ	620,80	1,63	0	**498,22**

Tableau 20 - *Les vocables grammaticaux non flexionnels selon* U *et* N

On peut résumer la représentation des différentes catégories morphosyntaxiques comme suit :

Classes de V grammaticaux	*V dont U ≥ 3*	*N*	*X*	*σ*
Adverbes	49	935	80,30	129,05
Conjonctions	33	386	314,72	855,43
Prépositions	8	6 170	771,25	1079,71
Prédicatifs nominaux	5	2 047	409,40	584,12
Autres	4	53	13,25	5,96
Total	**99**	**22 591**	**228,19**	**620,80**

Tableau 21 - *Les classes des vocables grammaticaux non flexionnels dont U ≥3*

Nous présentons ci-dessous les vocables grammaticaux non flexionnels par sous-catégorie morphosyntaxique en indiquant chaque fois les indices de fréquence, de dispersion et d'usage. Il s'agit bien entendu des vocables dont U ≥ 3.

2.1.5.2.1. Prépositions

Nous avons dressé au chapitre 3 § 3.3.2.2. un inventaire des prépositions du kirundi en nous fondant sur notre corpus. Nous dressons la liste de ceux dont U ≥ 3.

Vocable	F_O	*D*	*U*	*Glose*
MU	3 177	0,97	3 079,50	'dans'
KU	1 375	0,92	1 266,80	'sur'
MURI	846	0,92	778,86	'dans'
I	392	0,98	346,21	'à'
KURÍ	199	0,91	180,47	'sur'
KUVA	161	0,92	147,39	'depuis'
NAYÓ	10	0,67	6,75	quant á
KÍRETSE	12	0,76	9,12	sauf
KUGEZA	10	0,38	3,83	jusqu'à

L'on peut constater que les prépositions à forme courte (*mu* et *ku*) ont des indices plus élévés que ceux de leurs correspondants à forme longue (*murí* et *kurí*).

Signalons que le vocable *mu* 'dans' a la deuxième fréquence la plus élevée de l'ensemble de notre corpus, la première étant occupée par la conjonction de coordination *na* 'et'.

2.1.5.2.2. Prédicatifs nominaux

Les prédicatifs nominaux constituent une classe peu fournie en kirundi. Ils font tous partie des vocables dont U ≥ 3. Les données suivantes montrent que le prédicatif nominal *ni* 'c'est' domine au niveau de la fréquence.

Vocable	F_O	*D*	*U*	*Glose*
NI	1 439	0,94	1 354,7	'c'est'
NTAA	264	0,91	241,41	'il n'y a pas'
ATÁA	228	0,90	205,90	'sans'
SI (préd.)	101	0,86	86,98	'ne...pas'
NYAA	15	0,52	7,78	'*x* dont il est question'

2.1.5.2.3. Conjonctions

2.1.5.2.3.1. Conjonctions de coordination

Les conjonctions de coordination relient des phrases ou des membres de phrases. Voici la liste de ceux qui figurent dans notre corpus et dont U ≥ 3 :

Vocable	F_O	*D*	*U*	*Glose*
NA	4 876	0,96	4 698,60	'et'
KAÁNDI	424	0,91	385,40	'de plus'
CANKÉ	275	0,88	241,48	'ou'
HANYUMA	79	0,84	65,98	'ensuite'
MUGÁBO	77	0,83	64,05	'mais'
YAMARÁ	32	0,86	27,44	'mais'
BÉ	35	0,72	25,25	'et'

KANÁTSINDA	17	0,67	11,45	'du reste'
NKAKÓ	15	0,67	9,98	'de fait'

Il nous faut signaler ici que la conjonction de coordination *na* 'et' est le vocable le plus fréquent de notre corpus.

2.1.5.2.3.2. Conjonctions de subordination

La liste des conjonctions de subordination est très fournie. Le discriminant entre parenthèses indique que le vocable est homographe à un autre vocable d'une autre catégorie morphosyntaxique. Les conjonctions de coordination dont U ≥ 3 sont les suivantes :

Vocable	F_O	*D*	*U*	*Glose*
KÓ (conj.)	1 212	0,95	1 151,4	'que'
NGO	758	0,86	653,45	'que'
REERÓ	396	0,92	364,42	'alors'
KUKÓ	375	0,93	348,31	'parce que'
NÓ	260	0,90	235,27	'et'
ARÍKO	265	0,86	227,41	'mais'
IYÓ (conj.)	189	0,81	152,40	'si'
KUGÍRA	154	0,89	137,55	'pour que'
KUGÍRANGO	150	0,88	131,65	'pour que'
NOONÉ	138	0,86	119,00	'alors'
KUBÉERA	121	0,87	105,85	'à cause de'
YÚUKO	109	0,81	88,07	'que'
NÁAHÓ	100	0,86	85,93	'même si'
MÁZE	53	0,81	42,84	'dès lors que'
KUBÉERAKO	29	0,81	23,50	'parce que'
AHUÚBWO	27	0,82	22,09	'plutôt'
NOONEHÓ	18	0,80	14,32	'd'ailleurs'

2.1.5.2.4. Adverbes

Les adverbes, tout comme les conjonctions, constituent une classe fournie en kirundi. Ils ont des fréquences élevées. Nous distinguons des sous-catégories parmi les adverbes selon le sens qu'ils expriment : la manière, la comparaison, le lieu, le temps, etc.

2.1.5.2.4.1. Adverbes de manière

Les vocables grammaticaux non flexionnels dont U ≥ 3 comprennent sept adverbes de manière. Ce sont :

Vocable	F_O	*D*	*U*	*Glose*
INGÉNE	198	0,90	177,37	'comment'
NÉEZÁ	195	0,86	168,49	'bien'
ICEÉSE	42	0,72	30,07	'publiquement'
NÁABI	34	0,81	27,42	'mal'
NANTÁARYÓ	13	0,64	8,31	'constamment'
IDÓ N'ÍIDÓ	12	0,59	7,12	'en détails'
BUHÓRO	7	0,72	5,03	'lentement'

2.1.5.2.4.2. Adverbes de comparaison

Les adverbes de comparaison dont U ≥ 3 sont à trois. Les deux premiers signifient 'comme' : *nká* 'comme' (F_O = 443; D = 0,96; U = 424,57) et *nkó* 'comme' (F_O = 32, D = 0,75, U = 24,09). Le troisième est *gusumba* 'plus que' (F_O = 12, D = 0,60; U = 7,16).

L'adverbe *nkó* 'comme' est une variante contextuelle de *nka* 'comme' utilisée avant une préposition, tel qu'illustré en (158) où (w4) indique le sous-corpus de référence et le mot souligné la préposition :

(158) *ibihúgu* ***nka*** *Malawi* 'des pays comme le Malawi' (w4)
nkó *murí Liberiya* 'comme au Libéria' (w4)

2.1.5.2.4.3. Adverbes de lieu

Les vocables dont U ≥ 3 comprennent treize adverbes de lieu dont un, *imbere* 'avant' se démarque nettement. Il exprime l'antériorité spatiale et temporelle. Ce sont :

Vocable	F_O	*D*	*U*	*Glose*
IMBERE	461	0,91	417,77	'devant'/'avant'
INYUMA	111	0,89	98,61	'derrière'/'après'
HAGÁTI	75	0,86	64,71	'au milieu'
HAÁFI	36	0,80	28,77	'près de'
HEEJURU	31	0,86	26,51	'au-dessus'
INÓ	29	0,74	21,43	'de ce côté-ci'
HAASÍ	25	0,73	18,19	'par terre'
KURE	23	0,77	17,75	'loin'
MUHÍRA	22	0,70	15,39	'à la maison'
HANZÉ	19	0,60	11,48	'dehors'
IRUHAÁNDE	8	0,76	6,05	'à côté'
HAÁKURYA	7	0,59	4,11	'de ce côté-là'
EPFÓ	6	0,62	3,74	'en bas'

2.1.5.2.4.4. Adverbes de temps

Sept adverbes de temps figurent dans la liste des vocables dont U ≥ 3. Ce sont :

Vocable	F_O	*D*	*U*	*Glose*
UBU	207	0,93	191,96	'maintenant'
KEÉRA	68	0,78	52,74	'autrefois'
KARE	45	0,84	37,58	'tout-à-l'heure'
EJOBUÚNDI	25	0,77	19,21	'demain'/'hier'
ÉJO	22	0,71	15,64	'dernièrement' / 'bientôt'
VUBÁ	11	0,67	7,39	'bientôt'
HAMBERE	10	0,60	6,05	'autrefois'

2.1.5.2.4.5. Adverbes d'intensité

Trois adverbes d'intensité se retrouvent parmi les vocables dont U ≥ 3. Ce sont :

Vocable	F_O	*D*	*U*	*Glose*
CAANECAANE	150	0,90	134,42	'surtout'
CAANE	147	0,83	122,09	'beaucoup'
RWÓÓSE	41	0,74	30,40	'très'

2.1.5.2.4.6. Adverbes d'affirmation et de négation

Huit adverbes d'affirmation et de négation se retrouvent dans la liste des vocables dont U ≥ 3. Il s'agit des vocables suivants :

Vocable	F_O	*D*	*U*	*Glose*
NYÉNE	243	0,93	225,62	'bien sûr'
EKA	67	0,85	56,62	'non'
EREGA	39	0,66	25,64	'bien sûr'
SI (adv.)	30	0,61	18,42	'ne ...pas'
NAMBA	14	0,82	11,43	'rien'
OYA	8	0,62	4,93	'non'
EEGÓ	7	0,59	4,14	'oui'
KUMBÚRE	6	0,67	4,02	'peut-être'

Signalons que la particule dicto-modale *nti* 'ne..pas' dont la valeur affecte toute la phrase a été comptée parmi les adverbes de négation. Sa fréquence est de 637. Ses indices de dispersion et d'usage sont respectivement de 0,91 et 578,71.

Le locuteur du kirundi peut être frappé par l'absence des vocables *eegóme* 'oui' et *oyayé* 'non', variantes longues de *eegó* 'oui' et *oyá* 'non'. Ces variantes longues ont un indice d'usage inférieur à 3 (cf. liste).

Nous pensons que ces vocables caractérisent la communication orale et beaucoup moins l'écrit. Cette hypothèse serait à vérifier sur un corpus oral.

2.1.5.2.4.7. Adverbes d'interrogation

Six adverbes d'interrogation ont un indice d'usage ≥ 3. Ce sont les suivants :

Vocable	F_O	*D*	*U*	*Glose*
MBEÉGA	38	0,68	25,74	'est-ce-que ?'
RYÁARÍ	25	0,83	20,81	'quand ?'
MBEÉ	21	0,63	13,26	'est-ce-que ?'
KUKÍ	18	0,69	12,38	'pourquoi ?'
UBWO (interr.)	14	0,63	8,76	'vraiment ?'
HÉ	9	0,50	4,46	'où ?'

Les variantes *mbeé* et *mbeéga* n'ont pas, à notre connaissance, de contrainte de distribution qui les distinguent entre elles. La substitution opère dans tous les cas tel qu'illustré à travers les contextes ci-dessous où l'indication entre parenthèses réfère au sous-corpus.

(159) a. *mbeéga ivyo ni ibikí?* (w1)
mbeé ivyo ni ibikí?
'ça c'est quoi?'

b. *mbeé umúti w'Íkiruúndi urafíse akamaro?* (w13)
mbeéga umúti w'ikiruúndi urafíse akamaro?
'est-ce qu'un médicament traditionnel a de l'efficacité?

Avant de passer à l'analyse des deux autres tranches de vocables (dont 3 <U ≥0 et dont U < 0), signalons que notre corpus renferme peu d'interjections et d'onomatopées dont U ≥ 3. Seule l'interjection *hingé* 'minute!' satisfait à cette condition (F_O = 9, D = 0,42; U = 3,81).

2.2. LES VOCABLES DONT U EST COMPRIS ENTRE 3 ET 0

Les vocables dont 3 > U ≥ 0 sont au nombre de 1 905 et représentrent 47% du vocabulaire de notre corpus. Ce sont essentiellement des verbes (327 / 1 905), des substantifs à base nominale (578 / 1 905) et à base verbo-nominale (766 / 1 905) ainsi qe des vocables grammaticaux (229 / 1 905). Le tableau suivant fournit la représentation des différentes catégories.

	V	*N*		
Classes de V	*3 < U ≥ 0*	*3 < U ≥ 0*	$\bar{X}$	σ
Verbaux	327	971	2,97	1,77
Nominaux à base nominale	578	1 542	2,66	2,08
Nominaux à base verbo-nominale	766	1 506	1,97	1,44
Adjectivaux	5	28	5,60	7,02
Grammaticaux	229	602	2,62	1,54
Total	1 905	4 649	2,44	1,81

Tableau 22 - *Les classes de vocables dont* 3 > U ≥ 0

Les 1 905 vocables représentent 4 649 occurrences soit 4% du corpus, un pourcentage peu élevé.

La fréquence moyenne de ces vocables est de 2,44. Compte tenu de cette faible fréquence, nous ne nous attarderons pas sur leur description.

Il importe cependant de noter que les vocables dont U est compris entre 0 et 3 partagent certaines caractéristiques de mots rares tel que définis par Ménard (1978 : 33-43). Ce dernier a constaté pour le français que, d'un point de vue morphologique, les vocables rares sont généralement des vocables longs.

D'un point de vue sémantique, les vocables rares constituent des créations nouvelles ou des emprunts. D'autres doivent leur présence à des choix stylistiques des auteurs (Ménard 1978 : 37-38). Qu'en est-il de notre corpus ?

D'un point de vue morphologique et en fonction des catégories morphosyntaxiques, on peut relever un nombre considérable de vocables longs.

D'un point de vue sémantique, de nombreux vocables sont des créations nouvelles répondant à des besoins nouveaux de désignation alors que d'autres sont des emprunts. Nous fournissons quelques exemples pour les différentes catégories de vocables.

2.2.1. VERBES

Les vocables verbaux dont U est compris entre 3 et 0 sont morphologiquement caractérisés par un plus grand nombre de radicaux trisyllabiques et tétrasyllabiques. L'on compte 32 radicaux trisyllabiques (contre 4 parmi les vocables verbaux dont U ≥ 3 dont $F_O \geq 60$, leur fréquence moyenne). Quant aux radicaux tétrasyllabiques, on en compte 1 contre 1 pour les vocables dont U ≥3.

Nous présentons ceux d'entre eux dont la fréquence est supérieure ou égale à 2,44 (soit la fréquence moyenne des vocables dont U est compris entre 0 et 3) selon l'ordre décroissant de leur indice d'usage.

Vocable	F_O	*D*	*U*	*Glose*
[-ÍIHEBUUR-]	5	0,58	2,92	'se décourager'
[-HÚUNGABAN-]	8	0,33	2,66	'être troublé'
[-HÉEZAGIR-]	6	0,44	2,65	'bénir'
[-SÓBAANUR-]	5	0,47	2,35	'démêler'/'expliquer'
[-KÚRAKUR-]	4	0,59	2,34	manquer de maturité
[-DÚRUMBANY-]	4	0,53	2,11	'semer le trouble'
[-SÁAGIRIZ-]	4	0,46	1,82	'cerner *x*'
[-ÍIKANGUR-]	5	0,36	1,80	'se réveiller'

[-ÍIHEREER-]	4	0,40	1,60	'se retirer pour faire *x*'
[-ZÍIMANGAN-]	4	0,40	1,60	'disparaître'
[-KÁYANGAN-]	4	0,39	1,57	'briller'
[-GENDAGEND-]	4	0,38	1,54	'déambuler'/'se promener'
[-ÍIBUNGENG-]	5	0,15	0,70	'être enceinte'

2.2.2. NOMS À BASE NOMINALE

Les vocables nominaux à base nominale dont U est compris entre 3 et 0 ont deux traits particuliers : certains sont morphologiquement longs, ce sont généralement des noms composés; d'autres sont des emprunts plus ou moins intégrés au kirundi.

2.2.2.1. NOMS COMPOSÉS

Les noms composés dont il est question ici sont soit des unités qui appartiennent au fond traditionnel de la langue, soit des formations récentes créées pour des besoins de désignations modernes. Nous présentons à titre d'exemple (selon l'ordre décroissant de U) ceux dont la fréquence dépasse 2,44 (fréquence moyenne des vocables dont U est compris entre 0 et 3). Ce sont :

Vocable	F_O	*D*	*U*	*Glose*
UBUDUMBIDUMBI	5	0,53	2,64	'groupes / groupuscules'
INGIRAKAMARO	4	0,56	2,24	'*x* utile'
IKIMÉNYAMÉNYA	4	0,55	2,21	'preuve / signe'
IMPFAAGUSA	4	0,46	1,85	'*x* inutile /*x* bon à rien'
IZAABÚKURU	4	0,44	1,77	'troisième âge'
UBUBÉGITÓ	4	0,44	1,75	'ignorance / pusillanimité'
AMASIGARACIICARO	4	0,43	1,70	'promesses non tenues'
NYENÚRUGÓ	4	0,41	1,62	'chef de ménage'
UMUHÉKEHÉKE	4	0,31	1,26	'graines éparpillées'
AMABÉEREBÉERE	6	0,17	1,03	'lait maternel'

Quant aux formations récentes, elles ont été créées pour satisfaire des besoins de désignation modernes. Sans pouvoir dater exactement l'apparition de ces vocables, on peut penser qu'ils datent des années 50, période de la création de l'État burundais de type moderne et de ses nouvelles institutions telles le parlement, la magistrature, l'armée, etc.

On peut citer parmi ces créations récentes les vocables suivants :

Vocable	F_O	*D*	*U*	*Glose*
UBURÉENGANZIRA	8	0,31	2,45	'autorisation'
INTWAAZANGABO	4	0,54	2,18	'un officier'
UMUSHÍKIRIZAMAÁNZA	6	0,29	1,75	'substitut du procureur'
IKIRÁANGAMIÍNSI	7	0,22	1,57	'calendrier'
UMUNYAPOLITIÍKE	4	0,26	1,05	'un politicien'
UMUVÚUKAGÍHUGU	4	0,23	0,91	'citoyen'

2.2.2.2. EMPRUNTS

De nombreux vocables nominaux à base nominale sont empruntés, esssentiellement au français et au swahili. Nous les présentons dans l'ordre.

2.2.2.2.1. Emprunts au français

Le kirundi a beaucoup emprunté au français. Nous l'illustrons par les vocables dont la fréquence est supérieure à 2,44, fréquence moyenne des vocables dont U est compris entre 0 et 3. Ce sont :

Vocable	F_O	*D*	*U*	*Glose*
UMULIYAARÍDI	5	0,48	2,40	'milliard'
UMUPOROTISAÁNTI	7	0,33	2,30	'un protestant'
UMUZAYIRUWA	5	0,44	2,21	'un Zaïrois'
IKÍRO	4	0,55	2,20	'kilo'
IBIÍSI	8	0,26	2,07	'bus' (autobus)
ITEREFOÓNE	4	0,50	2,01	'un téléphone'
IDEFILE	5	0,38	1,90	'défilé'

IKAMYO	5	0,37	1,87	'camion'
IPOROJE	5	0,36	1,81	'projet'
ISAKARAMENTU	4	0,45	1,80	'sacrement'
IHEGÍTAARI	5	0,35	1,77	'hectare'
UMUKOMITE	8	0,22	1,76	'un membre d'un comité'
IKILOMEETÉRO	4	0,40	1,62	'kilomètre'
ISIMA	12	0,09	1,04	'ciment'
AGASHO	4	0,24	0,98	'cachot
UMUSHÍNWA	6	015	0,93	'chinois'
IKAAZIYE	4	0,23	0,92	'casier' (bouteilles)
ITOÓNI	4	0,22	0,86	'une tonne'
KONTAÁBURE	4	0,19	0,74	'comptable' (une personne)
IMÉETÉRO	8	0,09	0,71	'mètre'
INÓTA	6	0,10	0,60	'une note' (résultat)

2.2.2.2.2. Emprunts au swahili

Les emprunts au swahili sont nombreux. Nous ne présentons ici que ceux dont la fréquence est supérieure à 2,44. Il s'agit des emprunts suivants :

Vocable	F_O	*D*	*U*	*Glose*	*Mot en swahili*
ICUÚMBA	4	0,49	2,44	'chambre'	*cumba* 'chambre'
ISÁBUNÍ	4	0,56	2,26	'savon'	*sabuni* 'savon'
IKIRAATO	12	0,16	1,98	'soulier'	*kiato* 'soulier'
IKIZUNGU	9	0,20	1,78	'manières européennes'	*kizungu*
INÚSU	5	0,40	1,61	'moitié'	*nusu* 'moitié'
UMUFUÚNDI	5	0,28	1,42	'maçon'	*mfundi* 'artisan'
IGIKWEÉMBE	4	0,23	0,92	'pagne'	*kikwembe* 'pagne'

2.2.2.2.3. Emprunts à diverses langues

D'autres langues fournissent aussi des emprunts, par exemple l'anglais, l'espagnol, etc. Les vocables visés sont les suivants : *ishaáti* (de l'anglais *shirt* 'chemise', *uburengeeti* (de l'anglais *blanket* 'couverture'), *ibéenderá* (de l'espagnol *bandera* 'drapeau').

Signalons en passant que les vocables nominaux à base verbo-nominale dont U est compris entre 3 et 0 ne contiennent pas d'emprunts, ni d'unités particulièrement longues. Nous ne nous y attarderons donc pas. L'on peut penser qu'il s'agit d'un hasard statistique.

2.2.3. VOCABLES ADJECTIVAUX

Les vocables adjectivaux dont U est compris entre 0 et 3 sont au nombre de cinq. Ce sont :

Vocable	*Fo*	*D*	*U*	*Glose*
MFÁTAKÍBAÁNZA	18	0,13	2,42	'transitoire'
[-TOÓTO]	3	0,50	1,51	'jeune'
[-TAAGATIFÚ]	4	0,21	0,86	'saint'
[-NÍNI...NÍNI]	2	0,28	0,56	'de grande taille'
[-BÍSI]	4	0,09	0,09	'cru'

2.2.4. VOCABLES GRAMMATICAUX

Nous distinguerons les vocables grammaticaux fléchis des non fléchis. Les vocables grammaticaux fléchis dont U est compris entre 0 et 3 sont au nombre de 229 dont 182 fléchis et 47 non fléchis. Les vocables grammaticaux fléchis sont surtout des possessifs (85), des démonstratifs (22) des indéfinis (21), des interrogatifs (11) et des substitutifs (8). Les autres sous-catégories des vocables grammaticaux flexionnels se partagent les vocables restants.

Quant aux vocables grammaticaux non flexionnels dont U est compris entre 0 et 3, ils sont au nombre de 47 dont 20 adverbes, 13 conjonctions et 6 interjections.

Des 20 adverbes, ceux dont la fréquence est ≥ 3 (leur fréquence moyenne arrondie) sont les suivants :

Vocable	F_O	*D*	*U*	*Glose*
NÍNGOGA	7	0,42	2,91	'rapidement'
HARUGURU	5	0,54	2,70	'en amont'
RUNTU	8	0,25	1,96	'humainement'
BWEÉGU	4	0,47	1,89	'définitivement'
MUÚNSI	5	0,43	1,71	'sous'
RUGURU	4	0,52	1,35	'à l'intérieur du pays'/'en amont'
ICÁARIMWÉ	4	0,20	0,81	'ensemble'

2.3. LES VOCABLES DONT U< 0

Les vocables dont U < 0 sont au nombre de 832. Ils représentent 1 124 occurrences soit 1% de N pour 20% de V. Leur fréquence moyenne est de 1. L'on comprendra donc que la majorité d'entre eux sont des hapax. Les vocables dont U < 0 se répartissent comme suit :

	V	*N*			
Classes de V	*U < 0*	*U < 0*	$\bar{x}$	σ	*Hapax*
Verbaux	71	90	1,27	1,06	63
Nominaux à base nominale	238	304	1,27	0,71	196
Nominaux à base verbo-nominale	416	473	1,14	0,48	375
Adjectivaux	48	191	3,97	3,76	17
Grammaticaux	59	66	1,11	0,37	53
Total	832	1124	1,35	1,26	704

Tableau 23 - *Les classes de vocables dont* U < *0*

Les vocables hapax sont les plus nombreux dans la catégorie des vocables nominaux à base verbo-nominale, puis dans celle des vocables nominaux à base nominale, dans celle des vocables verbaux et enfin dans celle des vocables grammaticaux. Ils représentent 84% des vocables dont U est inférieur à 0.

Quels sont ces hapax? Tout comme pour les vocables dont U est compris entre 0 et 3, nous les abordons d'un point de vue morphologique et sémantique. Procédons catégorie par catégorie.

2.3.1. VOCABLES VERBAUX

Les vocables verbaux dont U est inférieur à 0 sont au nombre de 71. Parmi eux, 130 sont des hapax.

D'un point de vue morphologique, on constate que les 63 hapax comprennent 3 radicaux tétrasyllabiques et 15 radicaux trisyllabiques.

D'un point de vue sémantique, les 63 vocables verbaux hapax contiennent bon nombre de vocables utilisés à des fins stylistiques. Nous en fournissons quelques exemples en indiquant les vocables de la langue commune qui auraient pu être utilisés à leur place.

Vocable	*Glose*	*Vocable*	*Glose*
[-SÉSEREZ-]	'faire très mal'	[-BÁBAZ-]	'faire mal'
[-ÍIDAGADUR-]	'se défouler'	[-NÉEZEERW-]	'fêter'
[-TUBIRAN-]	'rencontrer subitement Y'	[-HÚUR-]	'rencontrer Y'
[-SHANGASHIRW-]	'se fâcher très fort'	[-SHÁVUR-]	'se fâcher'
[-ZÚUNGURUK-]	'déambuler'	[-TÉEMBEER-]	'se promener'
[-TÚURAGAR-]	'durcir à cause de la chaleur'	[-ÚUM-]	'sécher'

Signalons quelques emprunts au sein des vocables verbaux hapax. Nous indiquons leurs gloses et les vocables d'origine. Il s'agit des vocables suivants :

[-KOP-]	'donner à crédit' (du swahili *mkopo* 'dette')
[-CÁAGUR-]	'choisir' (du swahili *kucagua* 'choisir'
[-ZÚUNGURUK-]	'déambuler[1]' (du swahili *kuzunguka* ' tourner autour')

[1] Cf. *Le Petit Robert* 1976.

2.3.2. VOCABLES NOMINAUX À BASE VERBO-NOMINALE

Les vocables nominaux à base verbo-nominale dont U est inférieur à 0 sont au nombre de 416, soit la moitié de tous les vocables de cette tranche. Parmi eux, 375 sont des hapax.

D'un point de vue morphologique, les substantifs à base verbo-nominale sont complexes. Leur formation requiert plusieurs morphèmes, contrairement aux substantifs à base nominale qui n'en exigent que deux (le troisième, l'augment, étant prévisible cf. chapitre 3). Nous formulons l'hypothèse que le nombre élevé de morphèmes (des points de vue syntagmatique et paradigmatique) fait que les noms à base verbo-nominale sont plus nombreux que les noms à base nominale dans la langue, où ils expriment une multitude de sens liés à leurs morphèmes; ce qui réduit les possibilités de redondance des vocables nominaux à base verbo-nominale.

Voici par exemple quelques vocables longs parmi eux :

UWUZÓOGUSHÍGIKIRA	'celui qui te soutiendra'
UWUTÚUBAHIRIZA	'celui qui ne fait pas respecter Y'
UWURÉENGANIJWE	'celui qui subit une injustice'
UWÚTAÁYITEEGEEREYE	'ce qui ne l'a pas compris'
IKIZÓOGUSÍGARANIRA	'celui qui gardera *x* pour toi'

Les vocables à base verbo-nominale hapax sont aussi marqués par un nombre élevé d'infinitifs substantivés longs.

On a par exemple :

UKWEÉMERERWA	'fait d'être accepté par Y'
UKUBÚUMVISHA	'fait de leur faire comprendre'
UGUKÓRANIRIZA	'fait de rassembler Y [PRÉP] Z'

2.3.3. VOCABLES NOMINAUX À BASE NOMINALE

Les vocables nominaux à base nominale dont U est inférieur à 0 ont deux caractéristiques : certains sont morphologiquement longs, ce sont généralement des composés; d'autres sont des emprunts plus ou moins intégrés au kirundi.

2.3.3.1. NOMS COMPOSÉS

Les noms composés dont il est question ici sont soit des unités qui appartiennent au fond traditionnel de la langue, soit des formations récentes.

Nous illustrons les premiers par les exemples suivants :

IGIHENDABAJA	'lumière du soleil couchant'
IGISHÍTSIISHÍTSI	'souche de végétal'
IKIGANOGANO	'tige de blé'
IVYAÁKAATSI	'herbes médicinales'
IBITÉERASÓNI	'actions immorales'
UMWUÚZUKURUZA	'arrière-petit-enfant'

Quant aux formations récentes, on peut citer :

IVYÚUNYUNYU	'sels minéraux'
UMUGÍRANÉEZÁ	'membre d'un organisme humanitaire'
IKIMÉNYESHAMAKURÚ	'journal'
IMVÁAMAHAÁNGA	'un étranger'
UMUSHÍIGWAMAÁNZA	'juge'
UMUMÉNYESHAMÁANA	'un catéchiste'

2.3.3.2. EMPRUNTS

Les vocables nominaux à base nominale comptent de nombreux emprunts au français et au swahili. Citons pour les premiers :

IDISPANSÉERI	'dispensaire'
YUUBÍLE	'jubilé'
ITELEVIZIYO	'télévision'
UMUFÁRIZÁYO	'un pharisien'
IFWAYE	'un foyer social'
IMÓOTÉERI	'moteur'

Parmi les emprunts au swahili dont U est inférieur à 0, on peut citer :

IGITÁARÁ	'boisement' (de *kitaa* 'boisement')
IDIRÍSHA	'fenêtre' (de *dirisha* 'fenêtre')
IGISWÁAHÍLI	'langue swahili' (de *swahili* 'swahili')
UMWISHO	'fin du mois' (*mwisho* 'fin')

Signalons l'existence de vocables en voie d'intégration morphologique et qui, dans notre corpus sont caractérisés par des morphèmes appartenant au kirundi et au français. Nous les reprenons tels qu'ils figurent dans le corpus. Ce sont par exemple :

UMU-SÉNÉGALAIS	'un Sénégalais'
UMU-MINISTRE	'un ministre'
AMA-STOCK	'les stocks'
AMA-ROBINETS	'les robinets'
I-CENTRE	'un centre'

2.3.4. VOCABLES ADJECTIVAUX

Les vocables adjectivaux dont U < 0 sont au nombre de 48. Vingt d'entre eux constituent des adjectifs invariables. Nous les présentons exhaustivement; ils attestent en effet d'une certaine créativité de la langue et de ce fait présentent un intérêt lexicologique certain. L'on a :

CAÁMI	'royal'
KAAMÁ	'traditionnel'
KAÁRUHARIWE	'incomparable'
KAMARAMAZINDA	'qui fournit la preuve'
KAVÁAMAHAÁNGA	'étranger'
KAVUUKÍRE	'national'
KIJAMBERE	'moderne'
KIMARARUNGU	'qui met fin à l'ennui'
KIVÁANDIMWÉ	'amical'
KIVÚUKANYI	'amical'
KWAADARAÁTO	'carré'
MASABANO	'emprunté'
MBÁRIRANO	'entendu / raconté'
MBÚUMBARÚGO	'vivrier'
MPÁNAVYÁAHA	'pénal'
MPANUUZWAJAMBO	'consultatif'
MPARANIRAMAJAMBERE	'de perfectionnement'
MPUUZAMIGAMBWE	'inter-partis'
MVÁAMAHAÁNGA	'importé'
NDÁANGAMÍCO	'culturel'
NDEMAMUBIRI	'protéiné'
NDIMWÁ	'arable'
NGÁRUKIRAGÌHUGU	'qui sauve le pays'
NGIRAKAMARO	'utile'
NGOROORERAMAJAMBERE	'de perfectionnement'

NJÁBUKAMÁAZI	'exporté'
NKÁRISHABWÈENGE	'de perfectionnement'
NSHÍINGAMÁTEEKA	'législatif'
NSHÍINGWABÌKORWA	'exécutif'
NSHÍMIKIRO	'fondamental'
NTAHINYÚZWA	'suprême'
NTAYEGAYEZWA	'inamovible'
NTUNGANYAMUGAMBWE	'pour l'organisation du parti'
NYAMUKURÚ	'principal'
NYÚNGAANIRANGO	'd'appui aux familles'
RUBAÁMBA	'criminel'
RUHÓONYANGAÁNDA	'de destruction massive'
RURANGÍRANWA	'célèbre'
SAHWANYA	'synodal'

2.3.5. VOCABLES GRAMMATICAUX

Les vocables grammaticaux fléchis dont U est inférieur à 0 sont au nombre de 59. Parmi eux, 44 sont fléchis et 15 ne sont pas.

Nous distinguerons ici aussi les vocables grammaticaux fléchis des non fléchis.

Les 44 vocables grammaticaux non fléchis 38 sont des hapax. Ce sont essentiellement des possessifs (21), des démonstratifs (6) et des indéfinis (6).

Quant aux vocables grammaticaux non fléchis dont U est inférieur à 0, ils sont au nombre de 15, tous des hapax dont 7 adverbes.

2.4. LES « AUTRES » MOTS-FORMES DU CORPUS

La catégorie des « autres » mots-formes du corpus est composée essentiellement de noms propres, d'abréviations et de divers mots-formes appartenant à des langues autres que le kirundi. Cette catégorie représente un peu plus de 7 400 mots-formes, soit 7% du corpus.

2.4.1. LES NOMS PROPRES

Les noms propres du corpus sont essentiellement des anthroponymes, des toponymes et des noms de mois.

2.4.1.1. LES ANTHROPONYMES

Les noms de personnes présents dans le corpus sont généralement des noms d'autorités politiques (ex. présidents, ministres) ou religieuses (pape, évêque). Signalons ici qu'à la saisie, quand un nom est suivi ou précédé d'un prénom, nous les avons liés; cela nous permet de les compter comme une seule unité; ainsi *Peetéro Buyóya* est réécrit *Peetéro-Buyóya* 'Pierre Buyoya'. Parfois, le nom précède le prénom comme dans : *Ndadaye Melchior* (F_O = 6) où le prénom a une variante orthographique (dans *Melkiyori Ndadaye*). Lorsque le nom n'est pas suivi du prénom, il apparaît comme tel dans l'index. Pour dégager le nombre d'occurrences d'un nom propre, nous cumulons les fréquences pour tous ces emplois. Nous avons par exemple :

Nom	F_O	*Fonction*
Buyóya	77	Président du Burundi (1987-1993, 1995...)
Yeézu	53	Jésus
Ndadáye	42	Président du Burundi (3.7.1993 -21.10.1993)
Yohaáni Paulo	37	Jean-Paul (II) [le pape]
Rwagasóre	18	Héros de l'indépendance du Burundi
Ntibantuúnganya	10	Président du Burundi (1994-1996)
Bagaza	13	Président du Burundi (1976-1987)
Bududira	9	Évêque
Ntaamwáana	8	Évêque

Une analyse des collocatifs de ces noms propres pourrait déboucher sur une étude intéressante de lexicologie politique burundaise. Mais on comprend que ce n'est pas le but de la présente recherche.

2.4.1.2. LES TOPONYMES

Les toponymes les plus fréquents dans le corpus dépouillé sont des noms de pays et de centres urbains, administratifs ou religieux. Comme pour les anthroponymes, les toponymes connaissent des variantes graphiques : ex. *Burundi* et *Uburundi*. En cumulant les fréquences, on obtient les résultats suivants :

Nom	F_O	*Glose*
Uburuúndi *et* Buruúndi	505	Burundi
Bujumbura	128	Bujumbura (capitale du Burundi)
Afrika	77	Afrique
Rwanda *et* Urwanda	62	Rwanda
Ngoozi	34	Ngozi (province)
Muyiínga	32	Muyinga (province)
Gitéga	30	Gitega (province)
Zayíire *et* Zaïre	28	Zaïre
Róma	16	Rome

Ces données fréquentielles sur les toponymes pourraient aider notamment à l'évaluation de la couverture médiatique du pays.

2.4.1.3. LES MOIS

Les noms de mois sont considérés dans le corpus comme des noms propres. À ce titre, ils commencent par une majuscule.

Les mois servent comme marqueurs temporels pour situer les événements relatés par les journaux. Leurs fréquences seraient en partie à relier aux événements politiques. Ainsi *Munyonyó* 'onzième mois' est le mois au cours duquel le Burundi est passé de la monarchie à la république (coup d'État du 28 novembre 1966), de la

première à la deuxième république (coup d'État du premier novembre 1976); ce qui expliquerait une plus grande récurrence de ce nom dans le corpus. Nous avons les fréquences suivantes pour les mois de l'année :

Nom	F_o	*Glose*
Munyonyó	43	'onzième mois'
Gitugútu	39	'neuvième mois'
Mukákaro	35	'septième mois'
Ntwaaránte	30	'troisième mois'
Nyakaánga	27	'dixième mois'
Kigarama	25	'douzième mois'
Ruheéshi	24	'sixième mois'
Ruhuhúma	24	'deuxième mois'
Ndamukiza	22	'quatrième mois'
Rusaamá	22	'cinquième mois'
Nzéro	20	'premier mois'
Myandagaro	19	'huitième mois'

2.4.2. LES ABRÉVIATIONS

Les abréviations utilisées dans les journaux dépouillés désignent des organisations internationales ou nationales, des partis politiques et organisations affiliées ou des associations civiles. Parmi les premières, on peut citer :

Abréviation	F_O	*Au long*
S.O.S	14	(Nom propre d'une école)
O.N.U.	8	Organisation des Nations Unies
S.R.D.	6	Société Régionale de Développement
B.R.B.	4	Banque de la République du Burundi

Quant aux partis politiques et aux organisations qui leur sont affiliées, on peut citer :

Nom	F_o	*Au long*
U.PRO.NA.	100	Union pour le progrès national
FRO.DE.BU.	36	Front pour la démocratie au Burundi
P.R.P.	11	Parti pour la réconciliation du peuple
J.R.R.	8	Jeunesse révolutionnaire Rwagasore
U.F.B.	8	Union des femmes burundaises
PA.LI.PE.HUTU	7	Parti pour la libération du peuple hutu

2.4.3. LES MOTS-FORMES APPARTENANT À D'AUTRES LANGUES

Nous rangeons sous ce titre divers mots-formes du corpus appartenant essentiellement au français. Ces mots-formes dénotent des réalités nouvelles; ils font concurrence à leurs équivalents en kirundi. Nous fournissons quelques exemples en indiquant l'équivalent en kirundi correspondant.

maternité	iyaakiiriro ry'abavyéeyi
défilé	idefile
coup d'État	kudeta
minorités	imicé y'ábantu baké
chorale	umurwi w'ábarírimyyi

2.5. FRÉQUENCE DES CATÉGORIES GRAMMATICALES

Si l'on se reporte aux données du tableau 16 (p.159), on constate que la fréquence des mots-formes dans le corpus s'établit comme suit :

Classes de mots-formes	F_O	%
Mots-formes verbaux	25 651	24,76 %
Mots-formes substantifs	26 875	25,95 %
Mots-formes adjectifs	1 549	1,49%
Mots-formes grammaticaux	42 027	40,58%
Total	**96 102**	**92,79%**
N	103 561	100%

Tableau 24 - *Fréquence des catégories grammaticales*

Ces données sont-elles similaires à celles disponibles pour d'autres langues ? Nous allons comparer nos résultats à ceux obtenus dans d'autres études sur la fréquence des mots à l'écrit. Il s'agit de l'étude de Brunet (1981) pour le français et de celle de Johansson & Hofland (1989 : 15) fondée sur un corpus de textes informatifs anglais. Le tableau 25 regroupe les résultats de ces études.

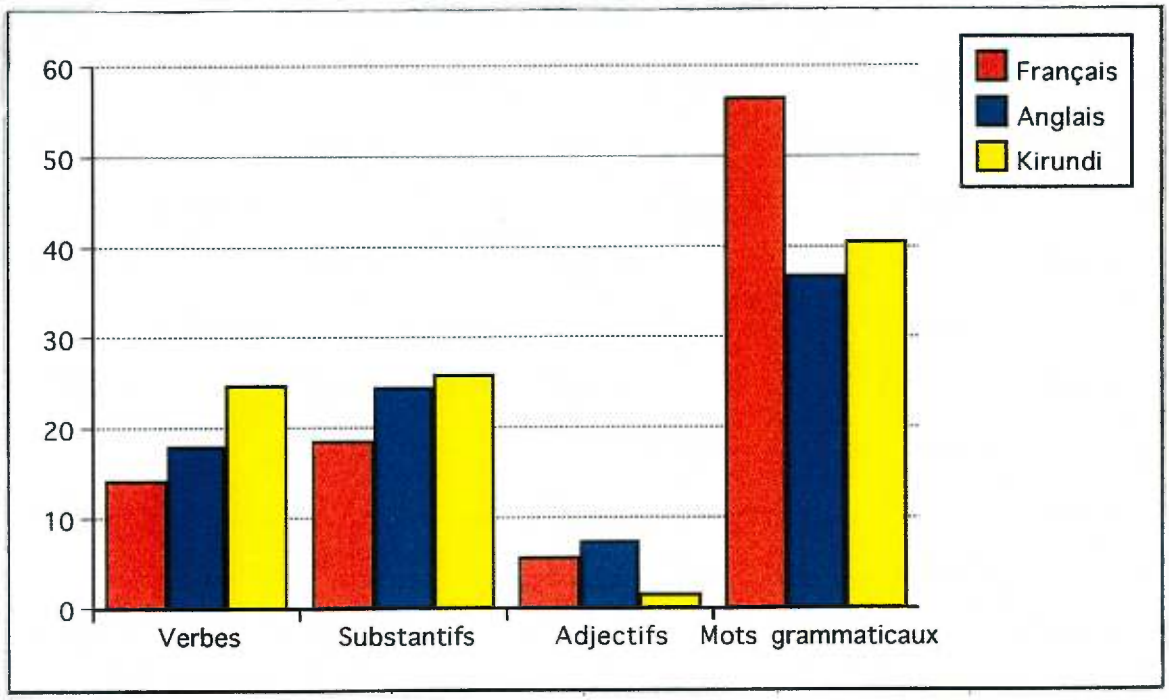
	Français		*Anglais*		*Kirundi*	
Catégorie	Brunet (1981)		Johansson & Hofland (1989)		Ntirampeba	
	N	%	N	%	N	%
Verbes	9 918 948	14,1 %	179 975	17,9 %	25 651	24,76 %
Substantifs	12 949 711	18,4 %	245 992	24,5 %	26 875	25,95 %
Adjectifs	3 837 060	5,5 %	73 546	7,3 %	1 549	1,49 %
MG	37 704 848	56,6 %[1]	368 666	36,8 %[2]	42 027	40,58%
Total	**70 273 552**	**94,6%**	**1 000 000**	**86,5%**	**103 561**	92,79%

Tableau 25 - *Fréquence des catégories grammaticales en français, en anglais et en kirundi*

[1] Ce chiffre est établi à partir des données du tableau 76 de Brunet (1981 : 298).

[2] Nous établissons ce chiffre à partir des données de Johansson & Hofland (1989 : 15), tableau - 5.

Reprenons ces données sous forme d'histogramme :



Graphique 2 - *Fréquence des catégories grammaticales en français, en anglais et en kirundi*

Pour le français, les unités les plus fréquentes sont les mots grammaticaux, puis les substantifs, les verbes et les adjectifs. Cet ordre vaut aussi pour l'anglais et le kirundi.

Lorsque l'on compare les chiffres sur les fréquences des catégories grammaticales en français, en anglais et en kirundi, trois conclusions se dégagent :

- les verbes sont plus fréquents en kirundi que dans les deux autres langues où les fréquences sont fort proches;
- les adjectifs sont nettement moins fréquents en kirundi que dans les deux autres langues;
- les mots grammaticaux ont des fréquences variables dans les trois langues.

La haute fréquence des verbes peut s'expliquer par le fait que la syntaxe du kirundi use beaucoup d'auxiliaires modaux; ainsi, rares sont les phrases qui contiennent moins de deux verbes.

Quant à la petite fréquence des adjectifs, elle s'explique notamment par le fait que nous avons restreint la définition de l'adjectif en kirundi par des critères morphologiques, en l'occurrence la présence d'un préfixe identique à celui du substantif avec lequel l'adjectif s'accorde.

Le kirundi dispose d'autres moyens pour exprimer les propriétés et les qualités des personnes ou des objets. On peut l'illustrer à partir des dénominations des couleurs, exprimées grâce à des mots-formes verbaux comme en (160 a, b) et à des syntagmes verbaux entiers comme en (160 c, d).

		Adjectif		*Glose*
(160)	a.	noir	*ciirábura* [ki]préf.pron. [íirabur]rad. [a]asp. 'il' 'être noir' 'inaccompli'	'qui est noir'
	b.	blanc	*ceerá* [ki]préf.pron. [éer]rad. [a]asp. 'il' 'être blanc' 'inaccompli'	'qui est blanc'
	c.	vert	*gisá na icaátsi kibisi* 'il ressemble' 'à' 'herbe' 'vert'	'qui est vert'
	d.	jaune	*gisá na umuhondo* 'il ressemble' 'à' 'colostrum'	'qui est jaune'

Si nous avions tenu compte de ces autres moyens d'expression, le nombre d'adjectifs dans le corpus se serait trouvé accru. Cela aurait cependant compliqué les analyses quantitatives pour deux raisons :

- la longueur des expressions déboucherait sur des index à entrées phraséologiques comme en (160 c)
- l'on serait obligé d'accorder le statut de mot à des séquences non lexicalisées comme en (160 c, d).
- l'on devrait procéder à la désambiguïsation de certaines formes verbales (160 a et 160 b par exemple) qui expriment une qualification pour les distinguer des autres formes verbales qui leur sont homographes.

Nous avons considéré les mots-formes en (160 a, b) comme des mots-formes verbaux et les séquences en (160 c, d) comme des syntagmes verbaux.

Le fait que l'adjectif soit exprimé à travers des formes verbales et des syntagmes verbaux est, sur le plan pédagogique, de première importance. La didactique des adjectifs du kirundi doit être abordée en rapport avec certains verbes et certains substantifs à l'intérieur du syntagme verbal.

À l'intérieur du kirundi, on observe une tendance paritaire entre les verbes et les substantifs avec respectivement des moyennes de 24% et 25%. Elle pourrait s'expliquer par la structure syntaxique de la langue.

En effet, comme l'ordre syntaxique du kirundi est SVO, chaque phrase met généralement en jeu deux substantifs, ce qui revient, pour une phrase, à contenir deux verbes (un modal et un lexical) et deux substantifs. Il faudrait mener la recherche sur d'autres types de corpus pour voir si cette tendance se maintient. Cette tendance paritaire des verbes et substantifs est moins prononcée en français et en anglais écrit.

Un autre aspect à analyser pour les rapports entre les catégories grammaticales est celui de la stabilité de leurs fréquences. Nous l'abordons au § 2.6.

2.6. STABILITÉ DES FRÉQUENCES DES CATÉGORIES GRAMMATICALES

La stabilité des fréquences des catégories grammaticales varie d'une langue à l'autre. En français par exemple, Gougenheim *et al.* (1964 : 117) a montré que, d'un corpus à l'autre, le pourcentage des verbes varie moins que celui des mots grammaticaux qui, à son tour, varie moins que celui des adjectifs. Qu'en est-il du kirundi écrit ?

L'analyseur morphologique nous a permis de dégager, pour chacun des seize sous-corpus, le nombre de verbes, de substantifs (à base nominale et à base verbo-nominale) et d'adjectifs. Les mots grammaticaux constituaient un fichier à part de même que les « mots divers », constitués des noms propres, des mots appartenant à d'autres langues et de quelques résidus. Rappelons également que la dénomination des sous-corpus « w1 » à « w16 » répond à un souci de conformité avec le langage de *WordCruncher* : « work 1 » à « w16 » signifie littéralement « tâche 1 » à « tâche 16 » et a pour sens « sous-corpus 1 » à « sous-corpus 16 ». Nous présentons un tableau-synthèse de la fréquence de ces catégories grammaticales où SBN = substantifs à base nominale, SBVN = substantifs à base verbo-nominale, Adj. = adjectifs et MG = mots grammaticaux.

Sous-corpus	*Verbes*	*SBN*	*SBVN*	*MG*	*Adj.*	*Autres*	*Total*
w1	792	536	205	1 218	48	65	**2 864**
w2	1 338	1 084	324	2 224	42	256	**5 268**
w3	1 216	1 041	413	2 162	54	133	**5 019**
w4	2 789	1 982	816	4 724	239	1 655	**12 020**
w5	765	468	228	1 056	37	47	**2 601**
w6	2 078	1 593	571	3 094	126	162	**7 624**
w7	1 270	1 008	357	2 136	46	197	**5 014**
w8	2 377	2 262	810	4 508	306	1 397	**11 660**
w9	557	554	173	1 094	39	138	**2 555**
w10	1 392	971	313	2 033	41	183	**4 933**
w11	1 124	1 116	396	2 142	49	161	**4 988**
w12	2 934	2 919	878	6 058	255	2 618	**15 662**
w13	1 099	716	273	1 581	75	57	**3 801**
w14	2 068	1 605	470	3 259	64	342	**7 808**
w15	984	754	309	1 683	60	103	**3 893**
w16	2 868	1 286	356	3 055	68	217	**7 850**
Total	**25 651**	**19 893**	**6 982**	**42 027**	**1 549**	**7 459**	**103 561**
%	**24,76**	**19,23**	**6,74**	**40,48**	**1,40**	**7,20**	**100%**

Tableau 26 - *Fréquences absolues des catégories grammaticales*

Peut-on affirmer que dans des corpus similaires on aurait à peu près ou autant de verbes, de substantifs, etc.?

Si l'on veut analyser la stabilité des fréquences des différentes sous-catégories morphosyntaxiques dans les 16 sous-corpus, il faut calculer un coefficient de variation pour chacune d'elles. On a les résultats suivants :

Catégories morphosyntaxiques	$\bar{X}$	σ	v
Mots grammaticaux	2 626,7	126,10	0,05
Verbes	1603,2	371,05	0,23
Tous les substantifs	1674,2	719,42	0,43
À base verbo-nominale	430,75	80,34	0,19
À base nominale	1 243,40	311,61	0,25
Adjectifs	96,81	44,27	0,46
Autres	471,63	465,10	0,99

Tableau 27 - *Stabilité des fréquences des catégories grammaticales*

Comme on peut le constater, les mots grammaticaux ont le coefficient de variation le plus faible suivis par les substantifs à base verbo-nominale, les verbes et les substantifs à base nominale.

Que la classe des mots grammaticaux ait un coefficient de variation faible ne surprend guère dans la mesure où, généralement, le nombre de mots grammaticaux n'est pas fonction d'un type de texte précis.

Les substantifs et les verbes auraient un coefficient de variation faible pour des raisons syntaxiques. L'ordre syntaxique du kirundi est SVO, chaque phrase met généralement en jeu deux substantifs, ce qui revient pour une phrase à avoir deux verbes (un modal et un lexical) et deux substantifs.

Notons cependant un détail important du tableau 26 : les mots-formes substantifs à base nominale sont systématiquement supérieurs en nombre aux substantifs à base verbo-nominale. Comment interpréter ce résultat?

Selon une hypothèse formulée pour l'anglais par Clark (1993 : 122-124), hypothèse partagée par Geeraert *et al.* (1994 : 134-146), les unités lexicales formellement simples sont, en langue générale, plus fréquentes que les unités dérivées ou composées.

Dès lors, on peut prédire que l'unité lexicale non dérivée *inzu* 'maison' est plus fréquente que son quasi-synonyme *inyubákwa* 'construction', qui, lui, est un dérivé construit sur le radical [-úubak-] 'construire'. Leurs fréquences respectives sont de 80 et 18.

Quant aux adjectifs et à la catégorie « autres » leurs coefficients de variation sont les plus élevés. Ces deux catégories seraient plus sensibles au type de texte.

2.7. LES 50 VOCABLES LES PLUS FRÉQUENTS

La comparaison des 50 vocables les plus fréquents en français, en anglais et en kirundi permet notamment de se faire une idée de la représentativité des différentes catégories grammaticales dans les zones de hautes fréquences.

Par souci d'adéquation, nous nous servons des études portant exclusivement sur des corpus écrits. Il s'agit, pour le français, du dictionnaire des fréquences du CNRS (1971), de l'étude de Johansson & Hofland (1989) pour l'anglais et de nos résultats pour le kirundi. Initialement, la liste de l'anglais n'est pas lemmatisée; nous avons donc lemmatisé les 50 mots-formes de la liste anglaise pour opérer les comparaisons voulues.

Dans le tableau ci-dessous, où SBVN = substantifs à base verbo-nominale, SBN = substantifs à base nominale, A = adjectifs et MG = mots grammaticaux, nous fournissons le nombre de vocables appartenant aux différentes catégories grammaticales parmi les 50 vocables les plus fréquents.

	Verbes	SBVN	SBN	A	MG
Kirundi	10	0	9	1	30
Français[1]	5	0	0	1	44
Anglais[2]	3	0	0	0	47

[1] Données du dictionnaire des fréquences du CNRS (1971).

[2] Données déduites du tableau 8 de Johansson & Hofland (1989 : 19-20), colonnes 1& 2.

Les 50 vocables les plus fréquents sont dominés et en français et en anglais et en kirundi écrit par les mots grammaticaux. C'est donc une caractéristique partagée par les trois langues. Il y en a cependant un peu moins en kirundi. Comment interpréter cette différence?

On peut raisonnablement penser qu'elle tient, d'une part à la structure de chaque langue particulière, et d'autre part à la définition du « mot grammatical ».

Compte tenu du fait que les études ne désambiguïsent pas les mots grammaticaux, il est difficile de décider de la représentivité des différentes catégories morphosyntaxiques (pronoms, prépositions, conjonctions, etc.) à travers la catégorie très large de « mots grammaticaux ». Quant aux substantifs, ils ne sont pas représentés parmi les 50 vocables les plus fréquents du français et de l'anglais et les verbes le sont proportionnellement très peu.

En ce qui concerne les verbes, les cinq plus fréquents du français sont, dans l'ordre : *avoir, être, pouvoir, faire* et *dire*; les trois plus fréquents de l'anglais sont : *be* 'être', *have* 'avoir' et *will* (l'auxiliaire).

Les 10 vocables verbaux les plus fréquents du kirundi sont dans l'ordre :

(161)

1. [-ri]	'être'		6. [-bón-]	'voir'
2. [-bá-]	'être' / 'habiter'		7. [-ti	'répliquer'
3. [-cí-]	'faire consécutivement *x*'		8. [-vúg-]	'parler'
4. [-gi-]	'aller'		9. [-kór]	'travailler'
5. [-gir-]	'faire'		10.[-mar-]	'finir'

Sept de ces verbes sont des auxiliaires modaux soit le n°1, le n°2, le n°3, le n°4, le n°5, le n°7 et le n°4. L'on trouvera leurs indices de fréquence, de dispersion et d'usage au § 2.1.1.2.3 de ce chapitre. Seuls trois vocables sont lexicaux (les n°6, 9 et 10). Ainsi donc pour le français, l'anglais et le kirundi, les verbes de hautes fréquences sont presque tous des auxiliaires.

L'on note une présence importante des substantifs du kirundi dans les 50 vocables les plus fréquents. Ce sont dans l'ordre :

(162)

1. *umuntu*	'personne humaine'	6. *umuúnsi*	'jour'
2. *igihúgu*	'pays'	7. *igihumbi*	'mille'
3. *imirongo*	'dizaine'	8. *igihe*	'le moment'
4. *iceénda*	'neuf'	9. *umwáaka*	'année'
5. *ijana*	'cent'		

Des neufs substantifs, quatre sont liés à la réécriture des dates (les n°3, 4, 5 et 7); trois situent le discours dans le temps *umwaáka* 'année', *igihe* 'le moment', *umuúnsi* 'jour'; un est lié au discours politique (*igihúgu* 'pays') et *umuntu* 'personne humaine' constitue un hypéronyme des 'humains'.

On n'observe aucun substantif à base verbo-nominale parmi les 50 vocables les plus fréquents. Cette absence de substantifs à base verbo-nominale s'explique par la complexité de ce type de substantifs; l'on croit qu'en général la fréquence est liée à la simplicité des formes. Nous avons par ailleurs montré que les substantifs à base verbo-nominale étaient dans l'ensemble moins fréquents que les substantifs à base nominale.

Voilà donc la structure lexicale du vocabulaire de base du kirundi écrit.

Nous avons estimé que, pour une langue agglutinante, les résultats lexicaux quoique intéressants, devaient être complétés par des données morphologiques fondés sur des index de certains morphèmes, mots-formes sélectionnés en fonction des objectifs pédagogiques qui nous guident. Nous abordons au § 3 cet aspect important de notre étude.

3. LES STATISTIQUES MORPHOLOGIQUES

Avant de poursuivre avec les données de morphologie telles que fournies par l'analyseur morphologique, revenons à notre objectif de départ; nous pourrons par la suite préciser les limites de ce volet de notre étude et les résultats attendus.

Notre but était de mettre au point un vocabulaire de base du kirundi écrit. Pour une langue agglutinante, il nous fallait un lemmatiseur qui ramenerait, automatiquement, les mots-formes fléchis d'un même vocable à ce dernier. Nous l'avons mis au point grâce au concours d'un informaticien et le soutien financier de notre codirecteur de recherche.

Le kirundi étant une langue agglutinante où le mot-forme atteint jusqu'à 14 positions paradigmatiques, il nous a fallu opérer un choix dans les morphèmes à compter. Nous les présentons ci-dessous.

3.1. LES MORPHÈMES OBJETS DE COMPTAGES

La description des mots-formes (cf. chapitre 3) a permis d'isoler de nombreux morphèmes qui participent à leur formation. Ils ne présentent pas tous un intérêt égal pour notre recherche.

Nous éliminons de nos comptages tous les morphèmes dont la présence est fortement liée à la situation de communication dans la mesure où nous ne cherchons pas à quantifier un type de texte particulier mais à dégager un vocabulaire de base. Nous ne retenons donc pas pour fins d'analyse :

- les préfixes verbaux personnels;
- les interjections;
- les onomatopées.

Nous ignorons aussi dans nos comptages les morphèmes dont le nombre d'occurrences est entièrement prévisible. Il s'agit :

- de l'augment pour les substantifs à base verbo-nominale et à base nominale (entièrement prévisible à partir du préfixe de classe);
- du préfixe adjectival (il est identique au préfixe de classe du substantif avec lequel l'adjectif s'accorde).

Nous ne tenons pas compte non plus des morphèmes strictement régis par la syntaxe, soit :

- les préfixes verbaux sujets pour les mots-formes verbaux;
- les morphèmes objets incorporés dans les mots-formes verbaux;
- les prédicatifs verbaux;
- le morphème réfléchi.

Finalement, nous ne quantifions pas la combinatoire des morphèmes. La multiplicité des combinaisons rend extrêmement difficile, dans une étude prospective comme la nôtre, leur traitement. Nous excluons de ce fait :

- les combinaisons de prédicatifs verbaux et de particules adverbales;
- les combinaisons des suffixes de dérivation.

Nous résumons dans le tableau 28 les morphèmes que nous retenons aux fins de la quantification.

Radicaux	*Morphèmes propres au verbe*	*Morphèmes propres au substantif*	*Morphèmes communs au verbe et au substantif*
1. Radicaux verbaux	Morphèmes aspectuels	1. Préfixes de classe	Suffixes de dérivation
2. Radicaux des substantifs à base verbo-nominale		2. Dérivatifs thématiques nominaux	
3. Radicaux des substantifs à base nominale			
4. Radicaux adjectivaux			

Tableau 28 - *Les morphèmes objets de comptages*

3.2. LES LIMITES DE CE VOLET DE RECHERCHE

Les limites de notre recherche sur la morphologie quantitative du kirundi écrit sont inhérentes à la performance de l'analyseur morphologique dont nous nous servons.

Rappelons qu'il reçoit en entrée un texte de N mots-formes, crée une base de données où chaque mot est représenté par une fiche (cf. tableau 10, p. 91), base de données qui est corrigée manuellement et qui est à la base de nos statistiques morphologiques. Ces dernières portent exclusivement sur la fréquence des morphèmes indiqués ci-dessus.

Les limites de l'analyseur morphologique sont liées aux faits linguistiques qu'il ne prend pas en compte. Il s'agit de la cooccurrence des morphèmes, de l'ordre des affixes, du déplacement tonal et de la quantification des règles (combien de fois une règle est appliquée). L'amélioration de l'analyseur devrait porter sur ces éléments.

Il reste cependant que l'analyseur morphologique nous a permis de réaliser un travail autrement impossible manuellement.

Le traitement de la morphologie du kirundi par ordinateur pose des problèmes liés notamment à la structure de la langue. Nous les abordons au § 3.3.

3.3. LES PROBLÈMES EN MORPHOLOGIE COMPUTATIONNELLE DU KIRUNDI

Nous avons vu tout au début de cette recherche (cf. Introduction § 2.4.) que les principaux problèmes posés par le traitement automatique de langues agglutinantes sont : le choix du lemme et la complexité morphophonologique. Nous les abordons dans l'ordre.

3.3.1. LE CHOIX DU LEMME

Le choix du lemme est crucial dans l'élaboration des vocabulaires de base car la lemmatisation est incontournable. Il nous a donc fallu décider à quel lemme ramener les différents types de mots-formes du kirundi. Nous avons décidé, rappelons-le, de ramener :

- les mots-formes verbaux au radical;
- les mots-formes substantifs à la forme du singulier dans la classe neutre;
- les mots-formes adjectivaux au radical.

Rappelons également que les mots-formes grammaticaux flexionnels ont été lemmatisés selon leur forme. C'est grâce aux concordances réalisées avec *WordCruncher* que nous avons départagé les fréquences des vocables grammaticaux homographes.

Parmi les vocables lexicaux, deux cas ont nécessité un traitement spécifique : il s'agit des mots-formes résultant de la réduplication du radical et de ceux mettant en jeu des radicaux dont l'un constitue une variante de l'autre.

3.3.1.1. LA RÉDUPLICATION DU RADICAL

Lorsqu'un radical est réduplqué, nous le considérons comme un autre lexème. Nous fondons cette décision sur l'écart sémantique entre le verbe à radical non réduplqué et celui avec radical réduplqué.

Ainsi l'infinitif *kugendagenda* 'marcher lentement' qui résulte, d'une réduplication du radical [-gend-] 'marcher' qu'on retrouve dans l'infinitif *kugenda* 'marcher', constitue un lexème différent de ce dernier.

Signalons que pour les mots grammaticaux nous soudons les mots-formes issus de la réduplication où l'on a par exemple *bamwé* 'les uns' versus *bamwéebamwé* 'certains'.

Rappelons aussi que dans l'attribution de la sous-catégorie morphosyntaxique, les substantifs composés et désadjectivaux sont rangés avec les substantifs à base nominale.

3.3.1.2. LES VARIANTES LEXICALES

Lorsque les mots-formes résultent de deux radicaux qui sont des variantes lexicales l'un de l'autre, il faut choisir le radical qui constitue le lemme. C'est le cas pour les mots-formes suivants, où c'est nous qui soulignons les segments qui changent de forme :

(163)	*zirahérekeranya*	versus	*zirahérekezanya*	'ils se suivent'
	kashaagirije	versus	*kasaagirije*	'il cerne'
	ingwáara	versus	*indwáara*	'maladie'

À la suite de Rodegem (1970), nous retenons comme lemmes les radicaux [-hérekez-] 'accompagner', [-sáagiriz-] 'cerner', [-rwáar] 'tomber malade'.

Les règles phonologiques présentées au chapitre 3 permettent d'expliquer le passage des radicaux aux mots-formes. Ainsi, par assimilation progressive de /r/ après la nasale /n/, [-rwáar] 'tomber malade' forme *indwáara* 'maladie' (cf. chapitre 3 § 3.1.4) qui se retrouve parfois sous la forme écrite *ingwáara* .

La reconnaissance automatique des radicaux verbaux et des substantifs a été rendue difficile par la complexité morphophonologique des mots-formes. Nous illustrons un certain nombre de ces difficultés dans les lignes qui suivent.

3.3.2. LA COMPLEXITÉ MORPHOPHONOLOGIQUE

On peut illustrer la complexité morphophonologique du kirundi avec le verbe *gusasa* 'faire le lit'. Au passé accompli, à la troisième personne du singulier, on a le mot-forme suivant :

(164) *yaashashe* [a]$_{\text{préf.pers.}}$ [a]$_{\text{préd.v.}}$ [sas]$_{\text{rad.}}$ [ye]$_{\text{asp.}}$ 'il a fait le lit'

Ce mot-forme met en jeu les changements morphophonologiques suivants que nous présentons, par commodité, de gauche à droite à l'exemple (164) :

a. consonantisation du préfixe verbal de personne /a/ qui change en semi-voyelle [j]
b. allongement de la voyelle /a / qui suit [j];
c. palatalisation de /s/ en initiale de radical qui devient [sh];
d. palatalisation de /s/ en finale de radical qui devient [sh] au contact de [j] du morphème aspectuel;
e. effacement du [j] du morphème aspectuel.

Lorsque dans un mot-forme interviennent plusieurs règles morphophonologiques, comme dans l'exemple ci-dessus, sa reconstruction prend souvent beaucoup de temps, fournit de mauvaises analyses ou échoue, nécessitant par là une intervention manuelle.

Outre ces difficultés, auxquelles se confronte toujours le traitement automatique de la morphologie des langues agglutinantes (cf. Introduction § 2.4.), le traitement automatique du kirundi bute sur des problèmes spécifiques à la langue. Il s'agit de l'existence dans le dictionnaire de l'analyseur de paires de radicaux dont l'un ressemble formellement à un dérivé de l'autre et de l'homographie des préfixes d'accord.

3.3.3. PAIRES DE RADICAUX DONT L'UN RESSEMBLE FORMELLEMENT À UN DÉRIVÉ DE L'AUTRE

Soit les mots-formes *bakorana* 'ils travaillent ensemble' et *bakorana* 'ils se réunissent'. Ces deux mots-formes sont construits sur deux radicaux différents du dictionnaire de l'analyseur tel qu'illustré en (165) :

(165) a. *bakorana* 'ils travaillent ensemble' [ba]$_{\text{préf.pers.}}$ [kór]$_{\text{rad.}}$[an]$_{\text{suff.ass.}}$ [a]$_{\text{asp.}}$
b. *bakorana* 'ils se réunissent' [ba]$_{\text{préf.pers.}}$ [kóran]$_{\text{rad.}}$ [a]$_{\text{asp.}}$

L'existence de nombreuses paires de radicaux dont l'un ressemble formellement à un dérivé de l'autre pose des problèmes de reconnaissance. L'analyseur morphologique ne distingue pas ces radicaux. En conséquence, l'on doit procéder à une désambiguïsation des mots-formes et opter pour un traitement manuel des mots-formes désambiguïsés. La tâche est lourde si l'on se réfère aux fréquences de quelques-uns parmi les radicaux concernés :

Radical	*Glose*	F_O
[-kór-]	'travailler'	329
[-kóran-]	'se réunir'	55
[-saang-]	'surveiller'	260
[-saangir-]	'partager'	81
[-tég-]	'tendre un piège'	32
[-téguur-]	'préparer'	95

Nous avons désambiguïsé les radicaux les moins fréquents en leur adjoignant un discriminant sémantique. Cela limite les interventions manuelles mais elles restent importantes compte tenu du nombre élevé de radicaux impliqués.

Quant à l'homographie des préfixes d'accord et des problèmes qu'elle pose pour le traitement automatique de la morphologie du kirundi, nous en avons parlé au chapitre 3 § 3.1.3.

Nous présentons les résultats de morphologie quantitative selon l'ordre du tableau 28, à savoir : les morphèmes propres au verbe, les morphèmes propres au substantifs et les morphèmes communs aux verbes et aux substantifs à base verbo-nominale.

3.4. FRÉQUENCE DES MARQUEURS ASPECTUELS DU VERBE

Le tableau 29 présente les fréquences des morphèmes aspectuels. La colonne *Sans* couvre les verbes défectifs qui ne manifestent pas de paradigme aspectuel. Pour des raisons de lisibilité, cette colonne est placée entre celles de [-ye] et de [-e], ce qui permet d'obtenir des fréquences décroissantes de gauche à droite.

Sous-corpus	*a*	*ye*	*Sans*	*e*	***Total verbes***
w1	441	177	93	81	**792**
w2	660	372	154	152	**1 338**
w3	671	307	130	108	**1 216**
w4	1 279	709	487	314	**2 789**
w5	431	183	86	65	**765**
w6	1 163	494	237	184	**2 078**
w7	690	325	157	98	**1 270**
w8	1 104	569	375	329	**2 377**
w9	297	171	60	29	**557**
w10	789	348	138	117	**1 392**
w11	512	341	157	114	**1 124**
w12	1 315	838	495	286	**2 934**
w13	651	247	115	86	**1 099**
w14	1 125	544	229	170	**2 068**
w15	543	247	120	74	**984**
w16	1 606	695	418	149	**2 868**
Total	**13 277**	**6 567**	**3 451**	**2 356**	**25 651**
$\bar{X}$	**829,81**	**410,43**	**215,68**	**147,25**	**1603,18**
σ	**191.32**	**50,88**	**50,73**	**29,99**	
v	**0,22**	**0,12**	**0,25**	**0,20**	

Tableau 29 - *Fréquences absolues des morphèmes aspectuels*

Ces données sont tirées de sous-corpus inégaux. Il nous faut donc, pour apprécier la fréquence, la dispersion et l'usage des différents marqueurs aspectuels, calculer leurs fréquences théoriques et leur coefficient de variation, éléments qui permettront de calculer l'indice de dispersion et d'usage. Nous procédons de la même manière que pour les vocables, tel que décrit au chapitre 1 § 3. Nous avons les résultats suivants :

Morphèmes aspectuels	F_O	*v*	*D*	*U*
a	13 277	0,22	0,94	13 392,0
ye	6 567	0,12	0,97	6 848,8
Sans	3 451	0,25	0,94	3 039,4
e	2 356	0,20	0,95	2 232,1

Tableau 30 - *Les marqueurs aspectuels selon* v, D *et* U

Les morphèmes aspectuels fréquents sont dans l'ordre :

[-a] 'action qui se déroule encore' (imperfectif)
[-ye] 'action qui s'est totalement déroulée' (perfectif)
[-e] 'action qui ne s'est pas encore déroulée' (imperfectif inchoatif)

L'absence de morphème aspectuel (*Sans*) concerne essentiellement les verbes défectifs construits sur les radicaux [-ri] 'être', [-ti] 'dire' et [-zi] 'savoir'. Rappelons que [-ri] 'être' a la fréquence la plus élevée parmi les vocables verbaux. Les verbes défectifs sont, de façon générale, plus fréquents que les verbes avec le morphème [e], du moins dans notre corpus.

Généralement, dans les langues du monde, l'aspect inaccompli est non marqué alors que l'accompli est marqué et les éléments non marqués sont en général plus fréquents que les éléments marqués (Mel'cuk 1994 : 13). La grande fréquence de l'aspect inaccompli du kirundi relèverait donc d'une caractéristique plus générale des langues naturelles.

Pour juger de la stabilité des différents morphèmes aspectuels dans les 16 sous- corpus, nous avons calculé leur coefficient de variation (cf. *v*).

On observe que le morphème aspectuel accompli [-ye] est celui dont le coefficient de variation est le plus bas. Il est donc le plus stable. Les trois autres ont des coefficients très proches.

Quant à l'indice d'usage, le plus élévé appartient au marqueur de l'inaccompli suivi de celui de l'accompli et enfin de celui de l'imperfectif inchoatif.

3.5. LES MORPHÈMES PROPRES AU SUBSTANTIF

3.5.1. LES PRÉFIXES DE CLASSE

Nous présentons ci-dessous les fréquences des préfixes de classe telles que fournies par l'analyseur morphologique. L'on constatera que nous n'avons pas cherché à désambiguïser les préfixes des classes 3 et 4 ([-mu-] et des classes 9 et 10 [-n-]. La classe « autre » groupe des substantifs qui n'ont pas de préfixe de classe et des substantifs relatifs formés avec les morphèmes [uwu-], [iyi-] et [izi-], tenant lieu de préfixes de classe. Nous les exemplifions plus loin.

	mu	*ba*	*mi*	*ri*	*ma*	*ki*	*bi*	*n*	*ru*	*ka*	*tu*	*bu*	*ku*	*ha*	*autre*	*total*
w1	101	98	35	99	58	50	67	58	15	18	3	39	37	4	49	731
w2	171	223	99	213	108	126	54	131	40	41	2	96	31		56	1 391
w3	187	189	75	182	155	82	87	132	23	40	7	96	105		72	1 432
w4	303	392	253	278	240	286	187	221	65	74	8	222	93	13	208	2 823
w5	91	63	23	75	60	73	55	71	24	35	7	30	47	2	35	691
w6	248	308	125	317	180	162	166	200	88	70	7	122	55	6	107	2 161
w7	202	146	81	238	91	106	69	102	32	37	5	83	86	3	72	1 353
w8	371	410	198	545	184	245	167	218	27	76	2	241	134	7	322	3 147
w9	109	76	60	95	66	64	68	79	6	9	3	43	27	2	14	721
w10	206	182	82	134	113	126	78	117	26	30	2	79	25	4	65	1 269
w11	253	186	87	152	123	152	100	124	14	43	3	106	46	7	71	1 467
w12	404	382	327	478	302	380	277	351	77	109	13	398	104	10	356	3 928
w13	136	111	65	108	79	98	78	124	20	24	3	59	36	7	32	980
w14	289	222	148	357	213	210	134	180	48	55	4	103	38	5	58	2 064
w15	111	104	87	144	75	92	134	118	13	27	4	79	25	12	33	1 058
w16	432	171	69	201	97	124	57	171	46	92	8	54	36	4	77	1 639
Total	3 614	3 263	1 814	3 616	2 144	2 376	1 778	2 397	564	780	81	1 850	925	86	1 627	26 875
$\bar{X}$	225,8	203,9	113,3	226	134	148,5	111,1	149,8	35,2	48,7	5,06	115,6	57,8	17,2	101,6	
σ	841	633	440	803	464	427	666	444	252	241	59	450	522	77	476	
v	0,29	0,22	0,23	0,29	0,23	0,16	0,29	0,17	0,46	0,24	0,51	0,34	0,46	0,61	0,60	

Tableau 31 - *Fréquences absolues des préfixes de classes*

À partir de ces fréquences, nous calculons le coefficient de variation, l'indice de dispersion et d'usage des préfixes de classes. Les résultats sont les suivants :

Préfixe de classe	F_O	*v*	*D*	*U*
ri	3 616	0,29	0,94	3 409,3
mu	3 614	0,29	0,94	3 407,4
ba	3 263	0,22	0,95	3 101,4
n	2 397	0,17	0,95	2 288,6
ki	2 376	0,16	0,95	2 266,6
ma	2 144	0,23	0,95	2 028,7
bu	1 850	0,34	0,93	1 726,7
mi	1 814	0,23	0,94	1 696,4
bi	1 778	0,29	0,91	1 618,9
autres	1 627	0,60	0,91	1476,6
ku	925	0,46	0,87	803,2
ka	780	0,24	0,92	718,3
ru	564	0,46	0,89	499,5
ha	86	0,61	0,79	68,3
tu	81	0,51	0,84	67,8

Tableau 32 - *Les préfixes de classes selon* v, D, *et* U.

Le tableau montre que les préfixes de classes les plus stables au niveau des fréquences sont :

[ki] v = 0,16
[n] v = 0,17
[ba] v = 0,22
[ma] v = 0,23
[mi] v = 0,23
[ka] v = 0,24

Les préfixes de classe les plus fréquents et à indice d'usage les plus élevés sont dans l'ordre :

[- ri-]	cl.5
[- mu-]	cl.1 & cl.3
[- ba-]	cl.2
[-n-]	cl.9 & cl.10
[-ki-]	cl.7
[-ma-]	cl.6
[-bu-]	cl.14
[-mi-]	cl.4
[-bi-]	cl.8

Vient ensuite la classe « *autre* » qui renferme des substantifs sans préfixe de classe dont certains mots de parenté comme :

Substantif	F_O	*D*	*U*	*Glose*
sé	31	0,52	16,10	'son père',
soókuru	14	0,67	9,32	'grand-père'
daatá	8	0,51	4,07	'papa'

La classe « autre » renferme également des substantifs formés avec les morphèmes [uwu-], [iyi-] et [izi-], tenant lieu de préfixes de classe :

Substantif	F_O	*D*	*U*	*Glose*
uwuróongooye	33	0,72	23,75	'celui qui dirige'
uwusháaka	8	0,68	5,46	'celui qui veut'

Les hautes fréquences des préfixes de classes ne se prêtent pas à une explication sémantique parce que de toutes les classes, seules les classes 1 (singulier) et 2 (pluriel) sont motivées comme nous le verrons au § 3.5.1.2.

3.5.1.1. LE PRÉFIXE DE CLASSE [-ri-]

Selon Mel'cuk & Bakiza (1987), les substantifs de la classe 5 n'ont pas d'unité sémantique. L'on peut cependant distinguer :

- des substantifs à base verbo-nominale (noms d'action)

Substantif	F_O	*D*	*U*	*Glose*
igenekerezo	110	0,85	93,53	'date'
iseezerano	80	0,60	47,89	'pacte'
ihiganwa	21	0,45	9,38	'compétition'

- des noms abstraits, utilisés souvent en politique :

Substantif	F_O	*D*	*U*	*Glose*
ijambo	217	0,82	178,37	'parole'
ibanga	128	0,85	108,62	'secret'
ishírahámwe	138	0,79	108,42	'une organisation'
itégeko	106	0,80	84,44	'loi'
ishaka	13	0,57	7,42	'volonté'

- des néologismes, empruntés aux langues étrangères :

Substantif	F_O	*D*	*U*	*Glose*
ishuúre	179	0,82	147,31	'école' de l'anglais *school*
ikómiíne	141	0,79	111,33	'la commune'
ibiro	53	0,82	43,21	'bureau'
iparuwaáse	66	0,50	32,77	'paroisse'
iraadiyo	19	0,69	13,10	'radio'

- des substantifs qui sont toujours au singulier :

Substantif	F_O	*D*	*U*	*Glose*
iteerambere	76	0,74	56,51	'développement'
itéeká	46	0,72	33,22	'droit'
izúuba	23	0,58	13,38	'soleil'

3.5.1.2. LE PRÉFIXE DE CLASSE [-mu-]

Le préfixe [-mu-] est morphologiquement ambigu : il correspond à la classe 1 et à la classe 3. Il s'agit d'un cas d' homonymie morphologique.

La classe 1 est sémantiquement cohérente : elle ne contient que des noms humains. Cependant, beaucoup d'autres noms humains appartiennent à d'autres classes; par exemple *umusóre* 'jeune homme' (cl.3) forme son pluriel dans la classe 4 et donne *imisóre* 'jeunes hommes' et *inkúmi* 'jeune fille' (classe 9) forme son pluriel en classe 10 et donne *inkúmi* 'jeunes filles'.

La classe 3 n'a pas d'unité sémantique; on peut juste dire qu'elle renferme des noms de gros arbres, qui ne sont pas très nombreux.

La fréquence du préfixe [-mu-] nous semble plus attribuable à la classe 1 qu'à la classe 3. En témoignent les substantifs de la classe 1 et de fréquence supérieure ou égale à 100 :

Substantif	F_O	*D*	*U*	*Glose*
umuntu	838	0,91	758,58	'personne humaine'
umuruúndi	343	0,79	270,52	'citoyen du Burundi'
umukúru	297	0,77	227,69	'le chef'
umwáana	245	0,79	194,46	'enfant'
umushíingantahe	152	0,85	129,92	'adulte mâle'
umukózi	131	0,78	102,83	'travailleur'
umugabo	148	0,59	87,59	'homme mâle'

umuvyéeyi	119	0,77	91,61	'parent'
umunyágihúgu	152	0,75	113,73	'autochtone'
umudásigáana	101	0,70	70,53	'membre du parti Uprona'

Ces dix substantifs représentent à eux seuls 2 506 occurrences du préfixe de classe [mu-] soit 63% de tous les substantifs de préfixe [-mu-].

Les substantifs de la classe 3 ayant une fréquence supérieure ou égale à 100 sont moins nombreux que ceux de la classe 1. Ce sont :

Substantif	F_O	*D*	*U*	*Glose*
umuúnsi	416	0,94	390,92	'jour'
umugambwe	412	0,65	268,55	'parti politique'
umwáaka	336	0,88	296,16	'année'
umutíma	170	0,82	139,45	'cœur'
umugaámbi	165	0,79	130,51	'projet'
umuryango	108	0,71	76,65	'famille'

Il y aurait lieu de pousser plus loin la recherche, de manière à désambiguïser le morphème [-mu-] et à affiner les résultats en tenant compte de tous les substantifs ayant [-mu-] comme préfixe de classe.

Dans les limites que nous nous sommes fixées, il nous a paru peu pertinent de nous astreindre à retourner dans les index, à désambiguïser à la pièce chaque mot-forme ayant [-mu-] comme préfixe de classe et à ramener son allomorphe [-mw-] au morphème [-mu-] (ex. *umwáana* 'enfant').

Quoique approximative, nous estimons en effet suffisante la donnée sur les rapports de fréquence entre les deux classes de substantifs.

2.5.1.3. LE PRÉFIXE DE CLASSE [-ba-]

Le préfixe [-ba-] caractérise les substantifs appartenant à la classe 2. Elle est sémantiquement homogène, composée de noms de personnes. Sa haute fréquence est à relier à celle de la classe 1.

2.5.1.4. LE PRÉFIXE DE CLASSE [-n-]

Le préfixe de classe [-n-] est ambigu : la même forme vaut pour le singulier et le pluriel, soit les classes 9 et 10. Sa fréquence relative de 9,16% n'est donc pas très élevée si on la compare à celle des classes 6, 8, 14 et même 4.

Pour désambiguïser le morphème [-n-], il faudrait recourir à une concordance qui fournirait le contexte linguistique permettant de décider si l'on a affaire à un mot-forme au singulier ou au pluriel.

Les classes 9 et 10 sont sémantiquement assez homogènes. Selon Mel'cuk (1987 : 330), elles comprennent notamment deux groupes sémantiques : celui des animaux sauvages et domestiques et celui des objets fabriqués. Parmi les noms d'animaux on relève dans notre corpus :

Substantif	F_O	*D*	*U*	*Glose*
inká	37	0,63	23,37	'vache'
intaama	13	0,67	8,77	'mouton'
impéne	15	0,58	8,77	'chèvre'
imbwá	6	0,15	0,91	'chien'
inkóko	3	0,21	0,64	'poule'/ 'coq'

Parmi les objets fabriqués on peut citer :

Substantif	F_O	*D*	*U*	*Glose*
inzu	85	0,78	66,69	'maison'
isáha	57	0,72	40,93	'montre'
imódoka	14	0,56	7,86	'voiture'

La classe 9 renferme de nombreux emprunts. *Isáha* 'heure' / 'montre' vient par exemple du swahili *saa* 'heure' / 'montre' et *imódoka* 'voiture' est emprunté à l'anglais *motocar* 'voiture'.

2.5.1.5. LE PRÉFIXE DE CLASSE [-ki-]

La classe 7 comprend des noms d'objets de type varié (territoires, objets fabriqués, éléments végétaux, etc.) et des noms de langues. Citons :

Substantiif	F_O	*D*	*U*	*Glose*
igihúgu	824	0,82	679,24	'pays'
igisagára	51	0,81	41,14	'ville'
igíti	35	0,77	36,92	'arbre'
ikiruúndi	36	0,66	23,75	'le kirundi'
ikiraato	12	0,16	1,98	'soulier'

2.5.1.6. LE PRÉFIXE DE CLASSE [-ma-]

Les substantifs appartenant à la classe 6 (*ma*) représentent 8% de tous les substantifs de notre corpus. Mel'cuk & Bakiza (1987 : 328-329) distinguent pour cette classe de substantifs quatre groupes sémantiques :

- substantifs désignant des liquides :

Substantif	F_O	*D*	*U*	*Glose*
amáazi	64	0,76	48,73	'eau'
amaráso	20	0,76	15,15	'sang'
amáta	12	0,45	5,42	'lait'

- des substantifs désignant des phénomènes sociaux :

Substantif	F_O	*D*	*U*	*Glose*
amahóro	143	0,37	52,68	'paix'
amajambere	63	0,68	42,88	'progrès'
amahera	26	0,70	18,07	'argent'

- des noms de moments ou de périodes :

Substantif	F_O	*D*	*U*	*Glose*
amaryáama	1	-0,05	-,08	'moment où l'on va se coucher'
amatúruka	2	-0,05	-0,10	'moment où le bétail va au pâturage'
amataaha	4	-0,05	-0,20	'moment où *x* rentre'

- des noms abstraits :

Substantif	F_O	*D*	*U*	*Glose*
amagára	51	0,68	34,70	'santé'
amacáakubíri	47	0,54	25,22	'les divisions'
amahíirwé	1	-0,02	-0,02	'le bonheur'

On peut ajouter à ces quatre groupes sémantiques celui des parties du corps représenté dans notre corpus par les substantifs suivants que nous présentons dans leur forme du singulier :

Substantif	F_O	*D*	*U*	*Glose*
ijíisho	44	0,75	32,86	'œil'
ukubóko	29	0,78	22,51	'bras'
ukuguru	21	0,68	14,20	'jambe'
urushí	11	0,75	8,24	'paume'
igúfa	5	0,47	2,37	'os'

2.5.1.7. LE PRÉFIXE DE CLASSE [-bu-]

Les substantifs de la classe 14 (*bu*) représentent 6% des substantifs de tout le corpus. La classe 14 comprend notamment :

- des noms collectifs

Substantif	F_O	*D*	*U*	*Glose*
ubúmwe	190	0,62	117,91	'unité'
ubwóoko	68	0,77	52,27	'ethnie'

- des noms de qualités :

Substantif	F_O	*D*	*U*	*Glose*
ubwéenge	36	0,74	26,66	'intelligence'
ubuntu	29	0,74	21,39	'humanité'/'bonté'
ubushíingantaáhe	15	0,75	11,20	'droiture'
ubutwáari	10	0,78	7,75	'courage'

2.5.1.8. LE PRÉFIXE DE CLASSE [-mi-]

Les substantifs de la classe 4 (mi) représentent 6% des substantifs du corpus. L'allomorphe du morphème [-mi-] est [-my-].

Tout comme ceux de la classe 3 (sa correspondante au singulier), la classe 4 n'est pas sémantiquement homogène. Les substantifs les plus fréquents dans notre corpus sont :

Substantif	F_O	*Glose*
imirongo	644	'dizaines'
imyáaka	156	'années'
imigambwe	149	'partis politiques'
imigaámbi	105	'projets'
imiínsi	86	'jours'

2.5.1.9. LE PRÉFIXE DE CLASSE [-bi-]

La classe 8 renferme de nombreux noms collectifs non comptables que Mel'cuk (1987 : 330) appelle des *pluralia tantum* (= toujours au pluriel) Son allomorphe est [-vy-].

On peut citer :

Substantif	F_O	*D*	*U*	*Glose*
ivyeérekeye	132	0,82	108,23	'ce qui concerne *x*'
ibirori	46	0,80	37,02	'spectacle'
ibiráaba	17	0,73	12,49	'ce qui concerne *x*'

La classe 8 renferme également des substantifs désignant des entités comptables dont nous fournissons la forme au singulier. Citons :

Substantiif	F_O	*D*	*U*	*Glose*
ikintu	193	0,90	173,63	'chose'
icíiyumviiro	104	0,77	80,58	'idée'
igíti	35	0,77	36,92	'arbre'

Les autres classes enregistrent des scores relativement bas (moins de 5%). Nous les considérons comme peu fréquentes.

Dans l'ensemble, parmi les paires de classes singulier /pluriel, les morphèmes du pluriel sont plus fréquents[1]. Le morphème [-ri-] (classe 4) est plus fréquent que le morphème [-ma-] (classe 6), le morphème [-ba-] (classe 2) est plus fréquent que le morphème [-mu-] (classe 1) et [-mi-] (classe 4) est plus fréquent que [-mu-] (classe 3). Seul [-ki-] (classe 7) est plus fréquent que [-bi-] (classe 8). Nous formulons l'hypothèse que ce fait serait lié au corpus écrit que nous utilisons; le pluriel est en effet plus anonyme que le singulier. L'hypothèse serait à vérifier sur un corpus oral.

---

[1] Il faut garder à l'esprit que le préfixe de classe [-mu-] est homographe pour les classes 1 et 3.

3.5.2. LES DÉRIVATIFS THÉMATIQUES NOMINAUX

Nous présentons ci-dessous les fréquences absolues des dérivatifs thématiques nominaux. La colonne « *Sans* » fournit les données sur les substantifs formés à partir de verbes défectifs qui n'acceptent pas de dérivatif thématique nominal.

Sous-corpus	*a*	*o*	*yi*	*Sans*	*e*	*i*	*u*	***Total***
w1	99	46	30	15	15	0	0	**205**
w2	116	121	25	32	26	4	0	**324**
w3	207	100	55	19	30	2	0	**413**
w4	242	306	103	79	44	1	1	**776**
w5	93	52	44	22	15	1	1	**228**
w6	216	185	61	68	37	3	1	**571**
w7	177	113	30	27	8	2	0	**357**
w8	340	246	80	39	62	2	1	**778**
w9	78	59	17	11	7	1	0	**173**
w10	103	116	29	42	22	1	0	**313**
w11	156	121	68	24	20	6	1	**396**
w12	361	302	119	64	110	2	0	**948**
w13	127	67	28	33	15	3	0	**273**
w14	201	140	55	31	39	3	1	**470**
w15	135	72	29	45	27	0	1	**309**
w16	116	127	52	37	18	4	2	**356**
Total	**2 767**	**2 173**	**825**	**588**	**495**	**35**	**9**	**6 892**
$\overline{X}$	172,9	135,8	51,5	36,7	30,9	2,1	0,5	**430,7**
σ	48,19	30,10	12,61	16,0	11,9	1,87	0,62	
v	0,26	0,20	0,24	0,46	0,38	0,85	1,1	

Tableau 33 - *Fréquences absolues des dérivatifs thématiques nominaux*

À partir de la fréquence des morphèmes, nous avons calculé un coefficient de variation pour chacun d'eux, l'indice de dispersion et d'usage. Nous obtenons les résultats suivants :

Dérivatifs thématiques nominaux	F_O	*v*	*D*	*U*
a	2 767	0,26	0,93	2 817,90
o	2 173	0,20	0,95	2 226,70
yi	825	0,24	0,94	772,89
Sans	588	0,46	0,88	492,83
e	495	0,38	0,90	445,85
i	35	0,85	0,78	27,28
u	9	1,10	0,72	6,44

Tableau 34 - *Les dérivatifs thématiques nominaux selon* v, D *et* U

Les dérivatifs thématiques nominaux dont les fréquences sont les plus stables à travers les 16 sous-corpus sont dans l'ordre :

[o] : v = 0,20
[yi] : v = 0,24
[a] : v = 0,26
[e] : v = 0,38

Mais selon la fréquence et l'indice d'usage, [a] vient en premier, suivi de [o] et de [yi]. Les données montrent que les dérivatifs thématiques nominaux [-i] et [-u] sont peu mis à contribution dans la formation des substantifs à base verbo-nominale. Ceux n'ayant pas de dérivatif thématique nominal (= *Sans*) sont construits sur des radicaux défectifs. Les dérivatifs thématiques nominaux les plus utilisées sont donc dans l'ordre :

[-a] 'résultat de l'action exprimée par le radical'

[-o] 'résultat, lieu, instrument, etc. de l'action exprimée par le radical'

[-yi] 'agent de l'action exprimée par le radical'.

Comment expliquer ces hautes fréquences ? Selon Carstairs-McCarthy (1994 : 2553-2557), les catégories sémantiques exprimées par les morphèmes dérivationnels sont notamment les catégories de « résultat » et d' « agent ». Les fréquences de [-yi] 'agent' et de [-a] 'résultat' s'expliqueraient dans ce contexte. Dans la langue, il faut fondamentalement des « agents », qui agissent et produisent des « résultats ». Les hautes fréquences relèveraient donc d'une caractéristique générale des langues naturelles.

Quant à la grande fréquence de [-o], elle s'expliquerait par la multiplicité de ses sens (cf. page 134).

3.6. LES SUFFIXES DE DÉRIVATION

Le verbe du kirundi a en commun avec le substantif à base verbo-nominale les suffixes de dérivation dont nous avons fait l'inventaire au chapitre 3, § 3.2.1.3.

Nous avons montré au chapitre 4, § 1.3.1.2.3, que de manière générale, les substantifs à base verbo-nominale fréquents correspondent à des radicaux verbaux fréquents.

L'on peut cependant s'interroger sur la nécessité de la quantification des affixes dérivationnels. Elle est liée à la notion de productivité.

La productivité des unités linguistiques est généralement étudiée sous deux aspects :

- dans le lexique conventionnel, existant, réel, déjà en usage dans une communauté linguistique; Clark (1993 : 126-127) parle dans ce cas de productivité passée;
- dans le lexique potentiel qui sert dans l'innovation lexicale; Clark (1993 : 126-127) parle de productivité présente.

Notre étude qui vise à élaborer un vocabulaire de base du kirundi écrit se fonde sur un corpus écrit. Sous cet éclairage, les données sur la productivité des morphèmes relèvent du lexique conventionnel qui a cours pendant les deux

synchronies couvertes par le corpus qui vont de 1975 à 1979 inclusivement pour la première et de 1990 à 1994 inclusivement pour la seconde.

Dans quelle mesure pouvons-nous extrapoler les résultats tirés du lexique conventionnel sur le lexique potentiel?

Sans être clairement mesurable, le lien entre les données fréquentielles obtenues sur le lexique conventionnel et le lexique potentiel est évident. Les locuteurs semblent utiliser davantage dans l'innovation lexicale les affixes les plus productifs du lexique conventionnel.

> *« Are these frequencies in the established lexicon paralleled by people's preferences for on form over the others in novel words ? Judgement studies show that people more often choose the suffixe identified, on the basis of existing words, as more productive »* (Clark 1993 : 129).

Nous passons à l'analyse de la fréquence des suffixes de dérivation. Nous présentons d'abord la fréquence des suffixes dans les mots-formes verbaux puis dans les substantifs à base verbo-nominale.

3.6.1. LES SUFFIXES DE DÉRIVATION DANS LES MOTS-FORMES VERBAUX

Le tableau 35 fournit les données sur la fréquence des suffixes de dérivation inventoriés au chapitre 3. Nous avons optimisé la lecture du tableau de manière à le rendre lisible en présentant les données selon des fréquences décroissantes.

La colonne « *Sans* » correspond au nombre de mots-formes verbaux qui ne sont pas suffixés.

	Sans	*ir*	*u*	*i*	*an*	*ish*	*ik*	*ar*	*uur*	*uuk*	*agir*	*ang*	*agur*	*agar*	***Total***
w1	558	93	45	45	24	17	5	0	1	2	2	0	0	0	**792**
w2	816	156	110	109	80	36	16	1	8	6	0	0	0	0	**1 338**
w3	855	140	99	63	30	17	1	0	4	5	0	0	2	0	**1 216**
w4	1 850	298	265	186	97	35	26	4	14	6	2	2	3	1	**2 789**
w5	513	96	57	26	40	14	6	0	10	1	0	1	1	0	**765**
w6	1493	172	164	97	90	35	12	1	10	2	0	1	1	0	**2 078**
w7	839	167	98	66	54	24	15	0	0	5	1	0	1	0	**1 270**
w8	1 539	309	183	170	87	55	19	4	5	4	1	1	0	0	**2 377**
w9	390	49	47	26	19	11	11	0	2	2	0	0	0	0	**557**
w10	1 020	116	106	69	41	22	9	2	4	2	0	0	1	0	**1 392**
w11	719	160	86	90	38	19	7	1	2	1	0	0	0	1	**1 124**
w12	1 686	409	341	254	133	54	27	2	9	13	3	0	3	0	**2 934**
w13	750	121	89	68	30	22	9	1	4	1	0	0	4	0	**1 099**
w14	1 496	171	149	115	81	37	10	2	2	3	1	0	1	0	**2 068**
w15	721	91	56	50	31	17	10	2	0	5	0	0	1	0	**984**
w16	2 356	181	82	109	87	22	12	0	6	8	0	1	3	1	**2 868**
Total	17 601	2 729	1 977	1 543	962	437	195	20	81	66	10	6	21	3	25 651
$\overline{X}$	1100,1	170,5	123,5	96,44	60,1	27,3	12,1	1,25	5,06	4,13	0,63	0,38	1,31	0,19	1 603,18
σ	385,65	21,83	25,77	13,10	14,1	7,35	4,01	0,97	3,46	2,14	0,74	0,54	1,20	0,38	
v	0,35	0,13	0,21	0,14	0,23	0,27	0,33	0,78	0,68	0,52	1,18	1,45	0,91	2,02	

Tableau 35 - *Fréquences absolues des suffixes de dérivation dans les mots-formes verbaux*

Ces données nous ont permis de calculer pour chaque suffixe de dérivation un coefficient de variation, un indice de dispersion et un indice d'usage. Nous avons les résultats suivants :

Suffixe de dérivation	F_O	*v*	*D*	*U*
sans	17 601	0,35	0,91	16 016,91
ir	2 729	0,13	0,97	2 647,13
u	1 977	0,21	0,95	1 878,15
i	1 543	0,14	0,96	1 481,28
an	962	0,23	0,94	904,28
ish	437	0,27	0,93	406,41
ik	195	0,33	0,92	175,50
uur	81	0,68	0,82	66,42
uuk	66	0,52	0,87	57,42
ar	20	0,78	0,80	15,96
agur	21	0,91	0,76	16,05
agir	10	1,18	0,70	7,00
ang	6	1,45	0,63	3,78
agar	3	2,02	0,48	1,44

Tableau 36 - *Les suffixes de dérivation dans les mots-formes verbaux selon* v, D *et* U

On peut constater que 68% des mots-formes verbaux du corpus (soit 17 601 / 25 651) n'ont pas de suffixes de dérivation. Les 32% restants sont essentiellement couverts par cinq suffixes de dérivation qui sont dans l'ordre :

[-ir-] 'applicatif'
[-u-] 'passif'
[-i-] 'causatif'
[-an-] 'associatif'
[-ish-] 'causatif indirect'

En observant le coefficient de variation de ces morphèmes, on constate que ceux dont les fréquences sont les plus stables à travers les 16 sous-corpus sont dans l'ordre :

[ir]	:	v = 0,13
[i]	:	v = 0,14
[u]	:	v = 0,21
[an]	:	v = 0,23
[ish]	:	v = 0,27

3.6.2. LES SUFFIXES DE DÉRIVATION DANS LES MOTS-FORMES SUBSTANTIFS À BASE VERBO-NOMINALE

Le tableau 36 présente les fréquences absolues des différents suffixes de dérivation dans les 16 sous-corpus. La colonne « *sans* » fournit le nombre de substantifs à base verbo-nominale sans suffixe de dérivation.

	sans	*u*	*ir*	*an*	*i*	*ish*	*ik*	*uuk*	*ar*	*uur*	*agir*	*ang*	*agur*	*Total*
w1	147	22	14	10	9	3	0	0	0	0	0	0	0	205
w2	204	41	22	22	15	12	5	1	1	0	0	0	1	324
w3	277	29	46	28	17	11	1	4	0	0	0	0	0	413
w4	494	71	71	84	30	16	5	3	1	1	0	0	0	776
w5	135	23	27	24	9	4	3	1	0	2	0	0	0	228
w6	361	70	57	43	20	15	3	1	0	1	0	0	0	571
w7	224	21	32	20	36	18	6	0	0	0	0	0	0	357
w8	408	119	80	79	45	35	9	1	0	1	0	1	0	778
w9	99	21	25	16	7	3	2	0	0	0	0	0	0	173
w10	230	14	25	25	6	10	0	2	0	1	0	0	0	313
w11	255	47	30	35	16	12	0	0	0	1	0	0	0	396
w12	545	96	107	120	47	17	7	11	0	0	0	0	0	948
w13	169	54	18	14	10	5	1	0	1	1	0	0	0	273
w14	317	48	44	30	10	19	1	1	0	0	0	0	0	470
w15	175	51	23	23	15	18	1	0	0	3	0	0	0	309
w16	280	23	16	16	11	9	0	0	0	0	1	0	0	356
Total	4 322	750	637	589	303	207	44	25	3	11	0	1	1	6 982
X	270	46	39	36	18	12	2	1,5	0,1	0,6	0,3	0,6	0,6	436,37
σ	44	18	11	13	7	6	2	2	0,9	0,4	0,2	0,2	0,2	
v	0,17	0,39	0,28	0,36	0,42	0,50	0,77	1,38	1,45	2,12	3,86	3,70	3,96	

Tableau 37 - *Fréquences absolues des suffixes de dérivation dans les mots-formes substantifs à base verbo-nominale*

Avec ces données, nous calculons pour chaque morphème son coefficient de variation à travers les 16 sous-corpus, son indice de dispersion et son indice d'usage. Nous obtenons les résultats suivants :

Suffixe de dérivation	F_O	*v*	*D*	*U*
Sans	4 320	0,17	0,96	4 147,20
u	750	0,39	0,90	675,00
ir	637	0,28	0,93	592,41
an	589	0,36	0,91	535,99
i	303	0,42	0,89	269,67
ish	207	0,50	0,87	180,09
ik	44	0,77	0,80	35,20
uuk	25	1,38	0,64	16,00
uur	11	1,45	0,63	6,93
ar	3	2,12	0,45	1,35
ang	1	3,70	0,04	0,04
agir	1	3,86	0,00	0,00
agur	1	3,96	-0,02	-0,02

Tableau 38 - *Les suffixes de dérivation dans les mots-formes substantifs à base verbo-nominale selon* v, D *et* U

Près de 62% des substantifs à base verbo-nominale n'ont pas de suffixe de dérivation (4 320 sur 6 982). Le reste des substantifs à base verbo-nominale est quasiment couvert par les cinq suffixes suivants :

[-u-]	'passif'
[-ir-]	'applicatif'
[-an-]	'associatif'
[-i-]	'causatif direct'
[-ish-]	'causatif indirect

Par rapport à la stabilité des fréquences de ces suffixes, stabilité évaluée par le coefficient de variation, on constate que les plus stables sont dans l'ordre :

- les substantifs à base verbo-nominale sans suffixe de dérivation
- les substantifs avec le suffixe [-ir-] 'applicatif'
- les substantifs avec le suffixe [-an-] 'réciproque'.

Le fait que près de 62% des substantifs à base verbo-nominale utilisés dans le corpus écrit que nous avons dépouillé soient sans suffixe de dérivation nous semble confirmer à suffisance que le kirundi écrit préfère les formes courtes et simples aux formes longues et complexes.

Lorsque l'on compare la présence des suffixes dans les mots-formes verbaux et dans les substantifs à base verbo-nominale, on obtient les résultats suivants :

Dans les verbes		%	*Dans les substantifs à base verbo-nominale*		%
[-ir-]	'applicatif'	10%	[-u-]	'passif'	10%
[-u]	'passif'	7%	[-ir-]	'applicatif'	8%
[-i-]	'causatif direct'	5%	[-an-]	'associatif'	8%
[-an-]	'associatif'	3%	[-i-]	'causatif direct'	4%
Sans		70%	*Sans*		62%

Dans l'ensemble, les mots-formes sans suffixe de dérivation dominent, que ce soit pour les verbes ou pour les substantifs à base verbo-nominale.

On voit que la formation des verbes et celle des substantifs à base verbo-nominale recourt aux mêmes suffixes. Mais comment expliquer la prédominance des quatre types de suffixes?

Nous pensons que la haute fréquence du causatif en kirundi est liée à l'importance de la causation dans la vie humaine. Selon Mel'cuk (1994 : 319), le causatif se retrouve dans presque toutes les familles de langues connues.

Nous pensons que la haute fréquence du morphème passif [-u-] est liée au fait qu'il exprime la voix, donc à son rôle grammatical même s'il est situé paradigmatiquement parmi les morphèmes de dérivation.

L'applicatif du kirundi a un sens général; Mel'cuk (1994 : 334) préfère parler « d'applicatif général ». C'est que l'action exprimée par le radical implique « de façon quelconque » le patient. L'action est faite pour x, à la place de x, en un lieu x, etc. Cette généralité du sens de l'applicatif expliquerait sa haute fréquence[2] .

Quant au réciproque [-an-], l'action exprimée par le radical s'effectue selon une interaction entre les participants à l'action. Sa haute fréquence nous semble liée à la communication en général et à l'activité sociale. Sa fréquence serait liée de ce fait à la nature du corpus.

Tels sont les résultats quantitatifs morphologiques auxquels aboutit ce chapitre de notre étude. Ils ont été obtenus grâce à un analyseur morphologique qui reçoit en entrée des fichiers de textes (notre corpus réparti en 16 sous-corpus) et fournit en sortie un index des morphèmes objets de comptages tel que inventoriés au § 3.1.

Ces résultats morphologiques viennent compléter les résultats lexicaux pour caractériser le vocabulaire de base du kirundi écrit.

Mais sélectionner un vocabulaire de base est une chose, en faire bon usage en est une autre. Nous abordons avec le chapitre 5 quelques applications pédagogiques du vocabulaire de base.

---

[2] Cette haute fréquence de l'applicatif est également signalée pour le wolof par Diop *et al.* (1975).

CHAPITRE 5

QUELQUES APPLICATIONS PÉDAGOGIQUES DU VOCABULAIRE DE BASE DU KIRUNDI ÉCRIT

1. INTRODUCTION

Une liste de vocables de base constitue une donnée brute. L'on peut en faire plusieurs usages. Les listes des vocables de base servent notamment à l'évaluation du vocabulaire des ouvrages didactiques et du niveau du vocabulaire des apprenants (Nation 1992, McCullough & Chacko 1976). Elles sont aussi utiles dans l'analyse de la lisibilité des textes (Beccaria 1976, Conquet & Richaudeau 1976). Elle servent aussi et surtout dans la didactique des langues.

Nous nous limiterons ici aux applications qui ont trait à la didactique des langues. Dans ce domaine, les applications ne vont pas toujours de soi et nombre de chercheurs se sont demandé ce qu'il faut ou ce que l'on peut faire de telles listes. Galisson (1971 : 15), évoquant la difficulté à utiliser la liste des vocables de base du français dont fait partie le vocable *vache*, avoue par cette boutade qu'« on ne sait que faire de la vache fondamentale ».

Nous estimons donc nécessaire d'aller au-delà de la liste pour en proposer des applications dans la didactique de la langue. Ce chapitre ne constitue cependant pas une recherche sur les multiples aspects de la didactique du vocabulaire du kirundi (écoute, expression orale, lecture, écriture, littérature, etc.); une telle étude ferait l'objet d'une ou de plusieurs autres recherches. De plus, nous ne traitons pas de la problématique de l'évaluation de l'acquisition du vocabulaire, problématique qui va généralement de pair avec toute réflexion didactique sur le vocabulaire. Elle constituerait à elle seule l'objet d'une autre recherche. Le lecteur intéressé pourra à ce sujet consulter, par exemple, Nation (1992).

Notre propos sera succinct et se limitera dans un premier temps, à présenter les aspects généraux de l'enseignement du vocabulaire. Nous formulerons ensuite des propositions sur l'utilisation du vocabulaire de base du kirundi écrit dans

l'enseignement du kirundi L1 dans le primaire, à l'école secondaire et en alphabétisation des adultes. Nous présenterons enfin quelques pistes pour l'enseignement du kirundi L2.

2. ASPECTS GÉNÉRAUX DE L'ENSEIGNEMENT DU VOCABULAIRE

Les aspects généraux de l'enseignement du vocabulaire concernent essentiellement deux questions : quel vocabulaire enseigner et comment l'enseigner ? soit une question de contenu et une autre de méthode (Nemni 1986 : 155-180).

La question du vocabulaire à enseigner est déjà réglée, du moins pour l'essentiel; c'est celui que nous avons sélectionné. Mais il faut considérer aussi le fait que l'utilisation d'un vocabulaire fondamental en didactique ne fait pas l'unanimité. Nous verrons au § 2.1. les critiques qu'on lui adresse généralement.

Reste la question de la méthode ou « comment enseigner le vocabulaire ? ». Nous présentons au § 2.4. un éventail de propositions méthodologiques et d'activités pour un enseignement systématique du vocabulaire.

Cette question de méthode est elle-même liée à celle de l'objet d'apprentissage. En effet, « qu'est-ce que connaître / apprendre / enseigner un vocable ? ». Nous répondons à cette question au § 2.2.

2.1. POUR ET CONTRE L'UTILISATION DES LISTES DE FRÉQUENCE EN DIDACTIQUE DES LANGUES

De manière générale, on reconnaît à l'utilisation des listes de fréquence en didactique des langues les principaux mérites suivants :

- elles permettent d'éviter la subjectivité dans le choix du vocabulaire à enseigner et de distinguer les vocables usuels des non usuels;
- elles permettent une limitation de la quantité des vocables à enseigner, le vocabulaire d'une langue étant trop vaste pour être abordé sans être partitionné;
- elles aident à opérer une gradation et un enseignement progressif du vocabulaire.

L'utilisation des listes de fréquence dans la didactique des langues fait cependant l'objet de critiques diverses, la question étant de savoir si ces listes constituent des outils fiables pour la didactique de la langue. Nous passons en revue les principales critiques formulées à leur endroit et fournissons notre appréciation de la portée de la critique.

2.1.1. LES VOCABLES FRÉQUENTS SONT PAUVRES EN INFORMATION

Dire que les vocables fréquents sont pauvres en information (Carter & McCarthy 1988 : 5) amène à s'interroger sur la notion d'information en linguistique.

Selon la théorie de l'information (Greimas & Courtés 1993 : 188), l'information est tout élément susceptible d'être codé. Il s'agit de rendre compte des modalités de transfert des messages (en tant que signaux organisés selon un code) d'un émetteur à un récepteur en se fondant uniquement sur le signifiant. Dans ce contexte, plus une unité (ici un vocable) est prévisible, moins elle est informative (Richards *et al.* 1985 : 141). Les vocables grammaticaux, qui sont généralement les plus fréquents dans les langues naturelles, tomberaient sous cette loi.

Sauf que le vocable n'est pas que « signifiant ». Il a aussi un signifié et un syntactique. Il porte des informations phonétiques, phonologiques, morphologiques, syntaxiques et sémantiques, informations qui n'ont pas encore fait l'objet de mesures quantitatives. Nous croyons donc que la critique ne tient pas; elle réduit le vocable au seul signifiant, donc aux informations phonético-graphiques.

2.1.2. LES VOCABULAIRES DE BASE CONSTITUENT DES LISTES DE FORMES ET PAS DE SENS

Le fait que les listes de vocables soient des listes de formes et non de sens (Wallace 1982 : 14) constitue une critique majeure à l'endroit des listes de vocables. L'utilisation de ces listes en classe de langue devient problématique, particulièrement pour ce qui est de l'enseignement des vocables polysémiques.

Mais selon Parisi & Castelfranchi (1988 : 135-136), on peut distinguer trois éléments qui aident le locuteur, l'enseignant et l'apprenant à contourner la polysémie d'un vocable d'une liste. Primo, la fréquence permet de privilégier d'emblée le sens le plus fréquent dans la langue. Secundo, le thème (ce dont on parle) sélectionne un sens déterminé. Tertio, le contexte syntaxique informe sur des cooccurrents et de ce fait élimine le ou les sens inadéquat(s).

À ces facteurs de type linguistique s'ajoutent les interactions entre tous les intervenants dans un contexte didactique. En classe de langue par exemple, le sens des vocables s'appréhende par un jeu de négociation enseignant <-> apprenants et apprenants <-> apprenants (Bérard 1991 : 57-59). L'enseignant ou les apprenants sont souvent amenés à expliquer un vocable, à en expliciter le sens ou à le reformuler. Même des moyens non linguistiques peuvent être mis à contribution comme la mime, le dessin, etc.

2.1.3. LES LISTES DIFFÈRENT POUR UNE MÊME LANGUE

Bien souvent, les listes pour une même langue diffèrent les unes des autres; les vocables les plus fréquents ne couvrent pas tous les textes (Nation 1992 : 21) et certains vocables disponibles sont absents des listes (Carter & McCarthy 1988 : 5).

Ces trois critiques tiennent aux corpus qui servent à la constitution des différentes listes et aux buts que poursuivent les chercheurs qui les mettent au point. Il est clair qu'une étude fondée sur un vocabulaire d'enfants (Préfontaine & Préfontaine 1968 par exemple) fournira une liste différente de celle obtenue en dépouillant un corpus d'adolescents (Fortier 1993) ou des corpus écrits, comme Baudot (1992) et Juilland *et al.* (1970), ou même des corpus mixtes (oral et écrit) comme Beauchemin *et al.* (1992).

Les différences entre les listes sont donc normales dans la mesure où chaque liste répond à des besoins précis. Mais il faut apporter une nuance : exception faite des dictionnaires de spécialité, les listes partagent malgré tout beaucoup de ressemblances. En témoignent les résultats de l'étude de Martel (1984 : 45) sur la comparaison des 54 vocables les plus fréquents du français fondamental (Gougenheim *et al.* 1964) et du québécois fondamental (Beauchemin & Martel 1979). Il ressort de cette étude, que seuls les vocables « *aller, autre, bien, là* [...] *puis* » manifestent des écarts significatifs.

2.1.4. LA FRÉQUENCE EST UN CRITÈRE INSUFFISANT

Selon Nation (1992 : 20-21), certains mots de haute fréquence ne sont pas nécessaires aux apprenants qui débutent et l'ordre des fréquences des mots n'est pas celui de leur enseignement.

Le critère de fréquence n'est pas le seul à déterminer l'ordre d'apprentissage des vocables. L'enseignant opère d'autres choix fondés notamment sur les autres critères que sont la répartition, la disponibilté et l'usage (cf. chapitre 2) ainsi que sur des considérations qualitatives comme les besoins linguistiques des apprenants. Ces derniers sont fort variés. McCarthy (1990) distingue quatre types de besoins lexicaux chez l'apprenant : il a besoin d'un vocabulaire pour parler du monde qui l'entoure, d'un vocabulaire d'orientation, d'un vocabulaire métalinguistique et d'un vocabulaire relié à ses intérêts académiques.

Plus spécifiquement, l'enfant burundais qui entre à l'école primaire connaît bon nombre de vocables de la langue. Il a d'abord besoin d'apprendre à lire et à écrire, puis à enrichir son vocabulaire.

L'étudiant du secondaire maîtrise mieux la langue orale et écrite. Mais il a besoin de vocabulaire technique et métalinguistique. Un futur technicien agricole doit par exemple connaître le vocable pour désigner « l'érosion des sols ».

L'adulte qui participe à l'alphabétisation connaît bien sa langue orale. Selon le programme du Ministère des Affaires Sociales, il a besoin d'apprendre à lire et à écrire en même temps qu'il a besoin de certains vocables techniques usuels dans l'administration et les divers milieux techniques.

L'étranger qui apprend le kirundi comme langue seconde a besoin d'un vocabulaire général pour s'exprimer à l'oral et à l'écrit mais aussi pour lire le journal et écouter ses interlocuteurs.

L'exploitation de la liste du vocabulaire de base en classe de kirundi tiendra donc compte de ces besoins. Nous reviendrons plus loin sur les critères qui, à l'intérieur de la liste, permettent d'opérer des choix quant à l'ordre de présentation des vocables en classe de langue. Précisons d'abord ce qu'on peut entendre par « connaître » un vocable.

2.2. QU'EST-CE QUE CONNAÎTRE UN VOCABLE ?

La connaissance d'un vocable implique plusieurs composantes. Selon Wallace (1982 : 27), connaître un vocable c'est notamment :

- le prononcer et l'écrire correctement;
- le rappeler (réutiliser) quand c'est nécessaire;
- l'utiliser dans la bonne forme grammaticale;
- l'utiliser en respectant ses collocatifs;
- en connaître le (ou les) sens et relier chaque sens à un objet ou à un concept;
- connaître son registre et l'utiliser à ce registre;
- en connaître les connotations et les associations lexico-sémantiques.

La connaissance d'un vocable implique donc des connaissances phonético-graphiques, morpho-sémantiques, sémantiques, référentielles et culturelles. Corrolairement, Wallace (1982 : 9-13) énumère quelques signes d'un mauvais apprentissage du vocabulaire, dont :

- le vocable est mal prononcé ou mal orthographié;
- l'apprenant a un problème d'accès lexical; il est incapable d'utiliser un vocable quand il en a besoin. Il recourt alors à des procédés de compensation dont le plus connu est la paraphrase;
- le vocable est utilisé dans un registre qui ne convient pas.

Comme on le voit, ces symptômes d'un mauvais apprentissage du vocabulaire relèvent tous de la production. Wallace (1982) laisse de côté la reconnaissance, qui se manifeste par la capacité d'un apprenant à décoder le sens d'un vocable entendu ou lu. Dans le cas d'un vocable entendu par exemple, on peut distinguer schématiquement à la suite de Richards *et al.* (1985 : 73) trois temps :

- l'apprenant garde le vocable dans la mémoire à court terme;
- il l'analyse en unités plus petites (morphèmes);
- il en déduit le sens en regard de la situation de communication.

L'on comprendra qu'à chacun de ces trois moments, la compréhension peut être défectueuse.

La connaissance d'un vocable a donc plusieurs facettes. Un mauvais apprentissage du vocabulaire se répercutera sur toutes ces facettes. Nous fournissons ci-dessus quelques pistes pour un enseignement systématique du vocabulaire du kirundi à partir de la liste du vocabulaire de base.

2.3. L'ENSEIGNEMENT SYSTÉMATIQUE DU VOCABULAIRE

Le point de départ d'un enseignement systématique du vocabulaire est un contexte linguistique authentique corrélé à un thème centralisateur (Nemni 1986 et Ligier & Varagnolo 1986). Ces deux éléments fournissent un cadre à la présentation des vocables.

Nous les abordons aux §§ 2.3.1 et 2.3.2. Quant à la présentation des vocables, elle recourt à un ensemble de démarches méthodologiques que nous présentons au §2.4.3.

2.3.1. LE CONTEXTE LINGUISTIQUE AUTHENTIQUE

Selon Bérard (1991 : 49-55) le caractère authentique d'un contexte didactique peut être apprécié en regard de deux éléments :

- les documents de travail introduits en classe sont proches de ceux que l'on retrouve dans la vie quotidienne;
- les activités proposées aux apprenants se rapprochent des types d'échanges qui existent dans la vie quotidienne.

Il existe un éventail de contextes linguistiques authentiques dans lesquels l'enseignant peut tirer des textes qui servent de base à l'enseignement du vocabulaire. L'on peut citer :

- les films, les émissions de télévision ou de radio;
- les pièces de théâtre;
- les livres de littérature;
- la littérature orale : la poésie, les chansons, les contes, les discours de circonstance;
- les journaux écrits;
- les interviews et les discours politiques, etc.

2.3.2. LES THÈMES

L'enseignement du vocabulaire est articulé autour d'un thème. Le choix est fait en fonction notamment de l'intérêt et des besoins des apprenants. L'on consultera Bérard (1991: 33-41) sur l'analyse des besoins comme outil d'élaboration des programmes d'enseignement.

L'inventaire des thèmes susceptibles d'être exploités en classe de vocabulaire dépend de la culture et ne peut donc être généralisé d'un milieu à un autre.

Tout en reconnaissant ces limites, nous fournissons à titre de suggestions quelques thèmes autour desquels pourraient s'articuler l'enseignement des vocables qui figurent dans le vocabulaire de base (cf. annexe 4). Nous nous inspirons d'une typologie réalisée par le Ministère de l'éducation de l'Ontario (Canada). Nous utiliserons un de ces thèmes pour montrer comment opérer le choix des vocables.

Prenons par exemple le thème de « l'individu » et le sous-thème « personnes et lieux » dans une leçon d'enrichissement du vocabulaire destiné à des enfants de sixième année d'école primaire. On présentera d'abord les vocables suivants, connus des élèves, qui ont tous un indice d'usage élevé, c'est-à-dire supérieur ou égal à 3 (cf. liste) :

(164)	*umuntu*	'personne'	*umuryango*	'famille'
	umuruúndi	'Burundais'	*ubwóoko*	'ethnie'
	umuvyéeyi	'parent'	*umwáana*	'enfant'

ikibondo	'enfant'[1]	*umugabo*	'homme'
umusóre	'jeune homme'	*umwiígeme*	'fille'
urwaaruka	'les jeunes'	*ahantu*	'endroit'
muhíra	'chez soi'	*inzu*	'maison'
inyubákwa	'construction'	*umutuúmba*	'colline'
izoóne	'zone'	*ikómiíne*	'commune'
iprovéensi	'province'	*igihúgu*	'pays'
igisagára	'ville'	*iparuwaáse*	'paroisse'
umuruúndikazi	'Burundaise'	*umukényezi*	'femme'
umugóre	'femme'	*umupfáasóni*	'femme'[2]

À ces noms s'ajouteraient quelques vocables verbaux et grammaticaux à indice d'usage également supérieur ou égal à 3 qui entrent souvent en collocation avec les noms en (164). Il s'agit des verbes en (165 a) et des éléments grammaticaux en (165 b) :

(165) a.

kubá	'être'	*kubáana*	'vivre avec'
kwiínjira	'entrer'	*kugenda*	'aller'
kubwíira	'dire à'	*guhakana*	'nier'
guhámagara	'appeler'		

b.

ni	'c'est'	*na*	'et'	*mu*	'dans'
murí	'dans'	*i*	'y'	*jeewé*	'moi'
kure	'loin'	*ca*	'de'	*ya*	'de'

On pourrait penser à ajouter plus tard des vocables moins usuels que sont par exemple les unités en (166) dont l'indice d'usage est inférieur à 3 comme :

---

[1] Connotation affective.

[2] Connotation de « haut niveau social ».

(166) *akagáramarúgaánda* 'un entêté'
imváamahaánga 'un étranger au pays'
umuremeshakiyaago 'un animateur'
umuzimyamuriro 'un voisin'

Ainsi donc, à partir d'un thème donné, l'enseignant peut puiser dans la liste des vocables de base, selon les besoins des apprenants, les vocables à enseigner. La liste constitue une banque de vocables hiérarchisés selon l'usage et dans laquelle l'enseignant puise. Reste à identifier les approches méthodologiques appropriées pour enseigner ces vocables.

2.3.3. LES APPROCHES MÉTHODOLOGIQUES

L'enseignement du vocabulaire exploite traditionnellement plusieurs approches méthodologiques. Nous nous proposons d'illustrer l'utilisation de trois d'entre elles, à savoir : le champ sémantique, les relations lexico-sémantiques paradigmatiques (essentiellement la synonymie et l'antonymie, l'hyponymie et l'hypéronymie) et les relations lexico-sémantiques syntagmatiques (les collocations).

Nous suggérons quelques activités fondées sur le vocabulaire de base du kirundi écrit que nous avons sélectionné. En fait, il s'agit de montrer la valeur ajoutée de la liste du vocabulaire de base dans l'enseignement du vocabulaire. Les propositions que nous formulons sont transférables du niveau primaire au niveau secondaire et même à l'enseignement du kirundi comme langue seconde.

2.3.3.1. Le champ sémantique

On appelle champ sémantique l'ensemble des lexies qui partagent une composante sémantique identificatrice de ce champ (Mel'cuk *et al.* 1995 : 173).

Voyons par exemple, à partir de l'extrait en (167), comment on peut envisager l'utilisation du champ sémantique pour enrichir le vocabulaire des enfants.

(167) [...] *Umunsi umwe hari umuhungu yagiye kubaramutsa kuko yakorana n'umugore w'uwo mugabo . Ashitse arabaramutsa kuko yagize n'Imana asanga uwo mupfasoni bakorana ni ho ari. Haheze akanya wa musore ati : Madame ngenda uku nyene? Nta na Amstel? Na we nya mugore yari yamaze gutuma kera, kuko yari azi n'ico afata* [...]. [Ubumwe 1990, n° 653, p. 8]

Traduction :

[...] Un jour, un jeune homme alla leur rendre visite parce qu'il travaillait avec la femme de cet homme. Quand il arriva, il les salua; la chance lui sourit car il se trouva que la femme avec laquelle il travaillait était là. Après un moment, le jeune homme dit : « Madame, je vais partir comme ça ? Pas même une Amstel? Mais la femme avait depuis longtemps déjà envoyé quelqu'un acheter, parce qu'elle savait ce qu'il prenait.

À partir de l'extrait, l'enseignant amène ses élèves à identifier dans un premier temps les relations sémantiques entre les vocables du texte et dans un second temps, à aller en dehors du texte et à trouver d'autres vocables qui peuvent être reliés au contexte du texte. C'est dans ce second temps qu'il puise dans le vocabulaire de base et qu'il privilégie les vocables à indice d'usage élevé ($U \geq 3$).

Dans un premier temps, la classe aboutira à un schéma des relations lexicales entre les vocables de l'extrait. Soit les vocables tirés de l'extrait en (167) :

(168)	*umuhúungu*	'garçon'	*umusóre*	'jeune homme'
	kugenda	'partir'	*gushika*	'arriver'
	kuramutsa	'visiter'	*umugabo*	'homme'
	umugóre	'femme'	*umupfáasóni*	'femme '[3]
	madaáme	'Madame'	*gufáta*	'prendre'
	amstel	'marque de bière'	*umusóre*	'jeune homme'

[3] Cf. note précédente.

La mise en relation de ces vocables fournirait un schéma similaire à celui-ci :

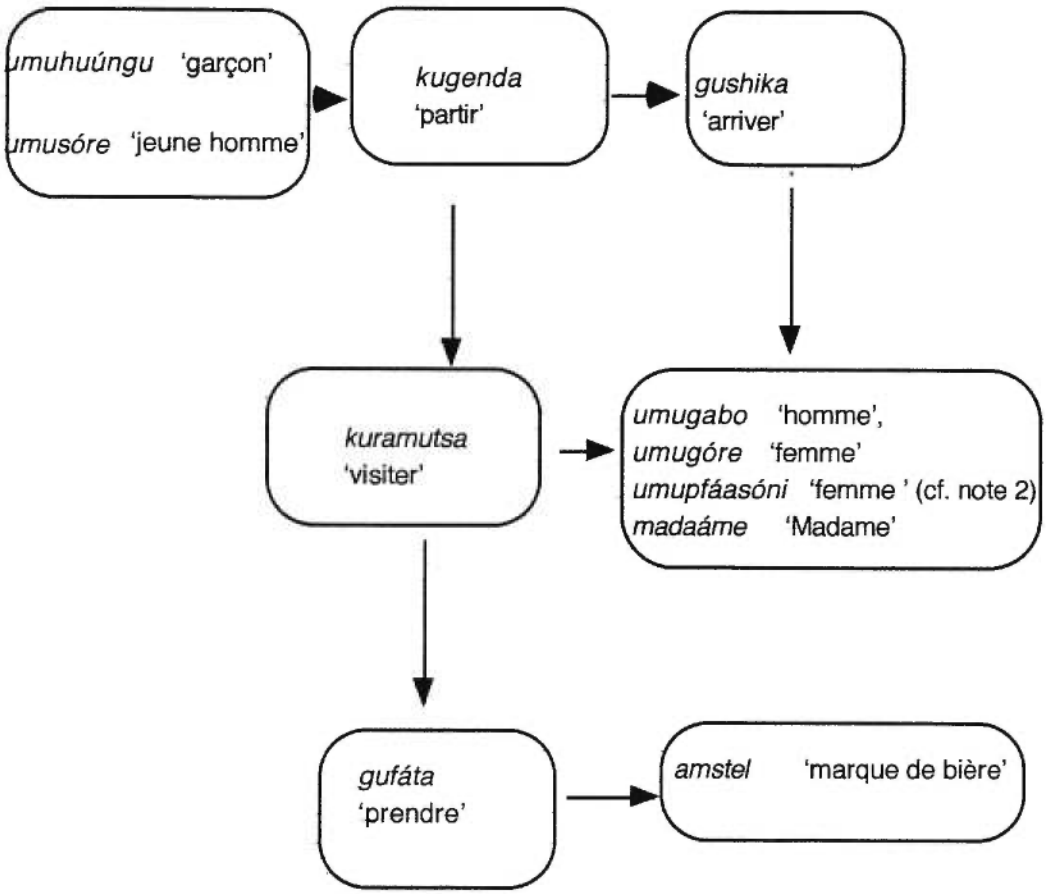


Schéma 2 - *Relations entre vocables d'un paragraphe*

À partir de ces relations entre éléments du texte, l'enseignant peut puiser dans le vocabulaire de base pour enrichir le vocabulaire de sa classe. Il pourrait par exemple utiliser les substantifs suivants dont $U \geq 3$:

(169)	*inzóga*	'boisson alcoolisée'	*urwáarwá*	'bière de banane'
	impéke	'bière de sorgho'	*ivyókuryá*	'nourriture'
	umugenzi	'ami'	*umuvyéeyi*	'parents'
	incúti	'parenté'	*ibiyéeri*	'bière Primus'

L'enseignant pourrait également ajouter les verbes suivants dont l'indice d'usage est élevé (U ≥ 3) :

(170)	*kwaakiira*	'recevoir'	*kunywá*	'boire'
	kuryá	'manger'	*gutéembeera*	'se promener'
	kuzimaana	'donner à boire aux visiteurs' / 'donner à manger aux visiteurs'		

Le schéma 2, enrichi de nouveaux vocables tirés du vocabulaire de base pourrait ressembler à celui-ci :

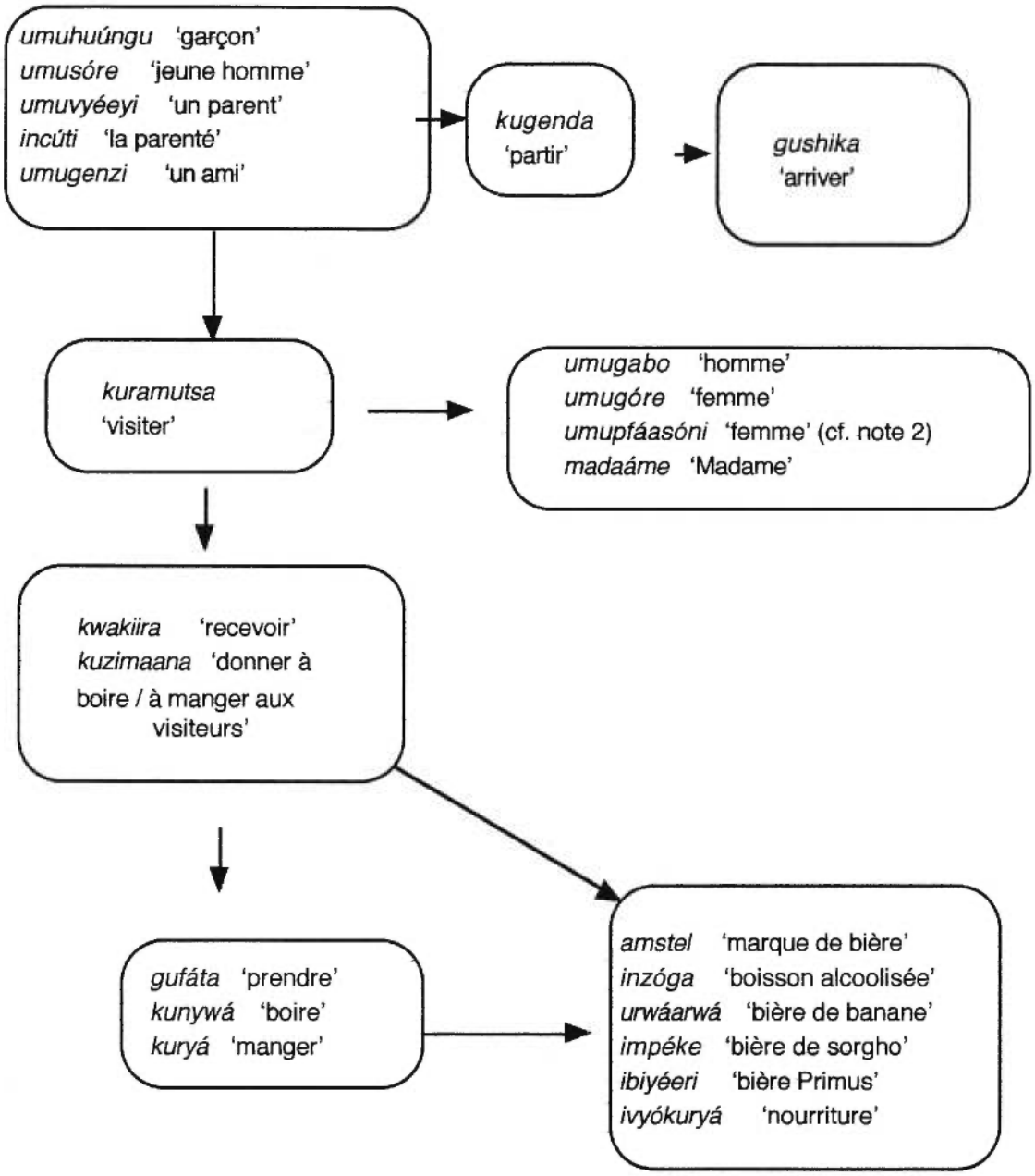


Schéma 3 - *Relations entre vocables tirés d'un paragraphe et de la liste du vocabulaire de base*

Ce type d'exercice permet l'identification de relations sémantiques entre les éléments du vocabulaire et ultimement l'élaboration de champs sémantiques comme celui de la parenté ou de la nourriture.

2.3.3.3. LES RELATIONS LEXICO-SÉMANTIQUES PARADIGMATIQUES

Tel que signalé au chapitre 2 § 3.2., les relations lexico-sémantiques cruciales en didactique des langues sont la synonymie, l'antonymie, l'hyponymie et l'hypéronymie. Ces relations sont fondamentales en classe de langue notamment en ce qui à trait à la définition (explication) des vocables.

2.3.3.3.1. La synonymie et l'antonymie

À partir de l'extrait en (167), l'enseignant puisera dans le vocabulaire de base pour préparer des exercices sur la synonymie et l'antonymie. Il amènera par exemple les élèves à trouver des quasi-synonymes au vocable *umupfáasóni* 'femme' (cf. note 2) à savoir *umugóre* 'femme' et *umukényezi* 'femme' ou des antonymes aux verbes en priorisant les vocables à haut indice d'usage qui fournissent les mots-formes du texte comme :

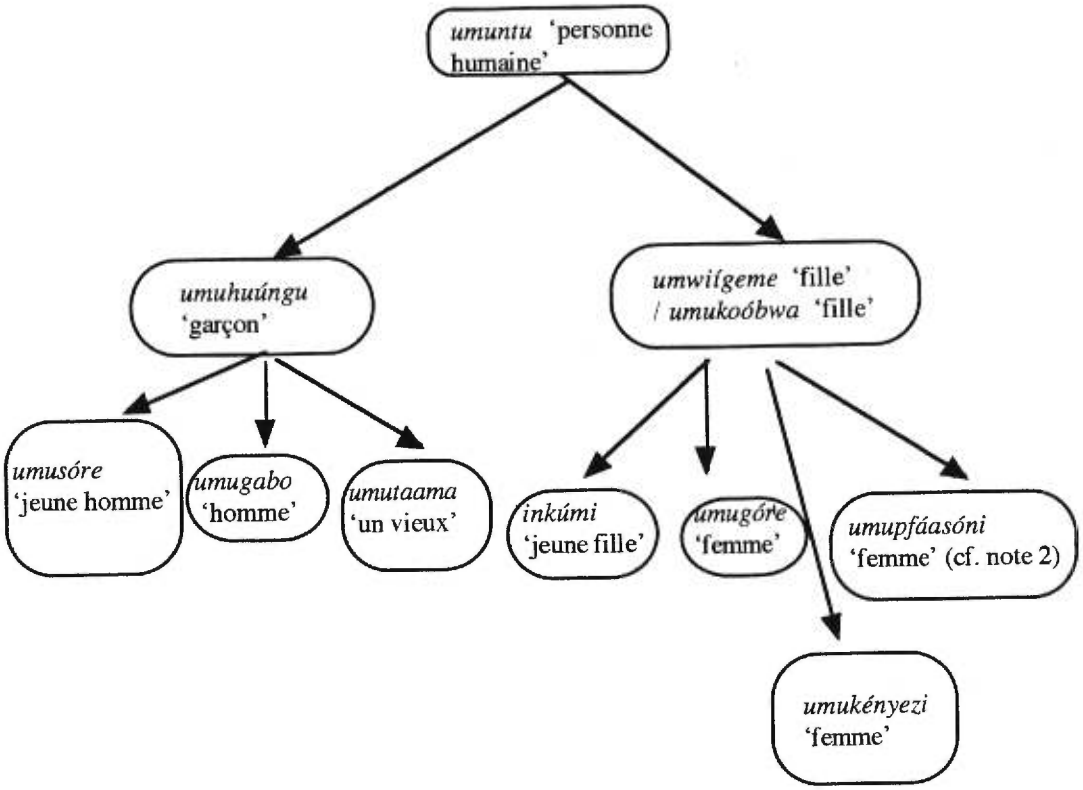
(171)	*yagiiye*	'il est allé'	versus	*yavuuye*	'il a quitté'
	ashitse	'il arrive'	versus	*agenda*	'il part'
	ataaha	'il rentre'	versus	*ashika*	'il arrive'

La prise en compte de l'indice d'usage fera que l'enseignant réservera pour plus tard l'enseignement du vocable *umucáancé* 'première femme' (U = -0,05)

2.3.3.3.2. L'hyponymie et l'hypéronymie

L'on se sert des relations d'hyponymie et d'hypéronymie pour amener les apprenants à construire des relations hiérarchiques simples entre les vocables d'un texte.

À partir de l'extrait en (167) et des vocables utilisés pour le sous-thème « personnes et lieux », on peut construire la hiérarchie suivante :



Il appartient à l'enseignant de complexifier cette hiérarchie en y ajoutant d'autres vocables tirés du vocabulaire de base comme par exemple les composés : *umushíingantaáhe* 'homme respectable', *séerugo* 'chef de ménage', etc.

2.3.3.4. LES COLLOCATIONS

On parle de collocation pour exprimer le fait que certains mots sont cooccurrents à d'autres avec une fréquence qui ne peut être due au hasard (Lewis 1997 : 215). Des travaux d'inventaire ont notamment été menés pour le français par M. Gross (1968) dans le cadre de la grammaire distributionnelle transformationnelle et ont abouti à des lexiques-grammaires avec notamment comme point d'encrage les verbes supports (ou opérateurs) comme *être*, *avoir* ou *prendre*. Ces derniers ont généralement des indices d'usage très élevés. Ainsi par exemple en français québecois les verbes *avoir*, *être* et *prendre* ont

respectivement des indices d'usage de 42 979; 40 957,10 et 2 206,30 (Beauchemin *et al.* 1992 : 747).

L'avantage des listes de fréquences est de permettre l'identification des verbes supports qui seront plus tard à la base des inventaires de collocations.

Soit par exemple, le vocable *gufáta* 'prendre'. Son indice d'usage est élevé (U = 219,26). Les élèves en connaissent le sens, mais ce vocable entre aussi dans la structure de locutions sémantiquement opaques ou plus ou moins compositionelles. Notre corpus contient par exemple les collocations suivantes :

(172)	a. *gufátana mu nda*	'être solidaire'	(w4)
	b. *gufáta mu mugongo* x	'aider *x* '	(w8,w12,w14)
	c. *gufáta ijambo*	'prendre la parole'	(w10)
	d. *gufáta urugendo*	'voyager'	(w10)
	e. *gufáta ingingo*	'décider'	(w10)
	f. *gufáta akayira*	's'en aller'	(w16)
	g. *gufáta amaráso*	'prélever du sang'	(w16)
	h. *gufáta nk'ámatáy'ábashitsi*	'prendre beaucoup soin de'	(w16)

Il appartient à l'enseignant de décider des collocations à enseigner en fonction du thème abordé. Les exemples en (d, f) pourraient bien cadrer avec le thème des « relations sociales » auquel renvoie l'extrait en (167). Les autres locutions pourraient être présentées plus tard.

La prise en compte de l'indice d'usage pourra donc permettre à l'enseignant :

- d'écarter dans les premiers enseignements des vocables non usuels,
- de hiérarchiser les vocables à enseigner,
- de disposer de vocables usuels pour les substituer aux non usuels dans l'élaboration des relations lexico-sémantiques.

L'utilisation de ces différentes méthodes permet de systématiser l'enseignement du vocabulaire à tous les niveaux académiques. Mais les besoins des apprenants contraignent à des choix spécifiques. C'est pourquoi nous distinguons dès à présent deux directions dans l'apprentissage / enseignement du vocabulaire, selon qu'il s'agit de la didactique de la langue maternelle ou de la langue seconde. Dans le premier cas, nous distinguerons le niveau primaire, le niveau secondaire et l'alphabétisation des adultes. Dans le second cas, il s'agira de voir quel vocabulaire enseigner aux apprenants selon leur niveau en kirundi L2.

3. VOCABULAIRE DE BASE ET DIDACTIQUE DU KIRUNDI L1

3.1. À L'ÉCOLE PRIMAIRE

À l'école primaire, le kirundi est une langue d'enseignement jusqu'en 3e année; tous les cours sont donnés en kirundi. Dès la quatrième année, il devient une matière d'enseignement; le français devenant la langue d'enseignement.

Le vocabulaire de base du kirundi écrit peut être utilisé dans les phrases-types qui servent de base à l'apprentissage de la lecture / écriture, et dans l'élaboration de textes simplifiés.

3.1.1. UTILISATION DU VOCABULAIRE DE BASE DANS LES PHRASES-TYPES

L'enseignement de la lecture et de l'écriture dans les écoles primaires du Burundi privilégie la méthode globale[4]. Elle part de la phrase vers les lettres. Les phrases-types sont choisies en fonction des correspondances phono-graphiques à enseigner. De plus, ces phrases de départ doivent refléter le milieu socio-culturel de l'enfant. Un même manuel, élaboré par le B.E.R.[5], est utilisé à l'échelle du pays. Une de ces phrases-types est par exemple la suivante :

---

[4] En alphabétisation, on parle de méthode analytique (Halaoui 1999 : 5).

[5] Bureau d'Éducation Rurale.

(173) *Kadege atugemuriye akayóga*
'Kadege' 'nous amène' 'de la bière'

Les trois mots-formes qui composent cette phrase relèvent des vocables suivants que nous présentons avec leurs indices d'usage :

Kadege	Kadege	'Kadege' (nom propre)	
kugemura	[-gemur-]	'amener des cadeaux'	(U = 2,65)
akayóga	inzóga	'boisson alcoolisée'	(U = 15,56)

Sans remettre en question la méthode, nous nous interrogeons sur le choix des vocables utilisés dans la phrase. Ne serait-il pas plus judicieux de former les phrases-types avec les vocables les plus fréquents, tout en gardant des phrases qui reflètent le milieu de l'enfant?

Ainsi pour enseigner la correspondance phono-graphique entre le phonème /e/ et le graphème [e], on pourrait envisager d'utiliser les vocables renfermant ce phonème et qui ont un indice d'usage plus élevé que ceux utilisés en (173). Ainsi à la place de la phrase-type en (173), on aurait celle en (174) où le vocable [-komer-] 'être en bonne santé' a un indice d'usage de U = 94,46.

(174) *Peetéro arakomeye*
'Pierre' 'est en bonne santé'

Nous croyons que le choix des vocables qui composent les phrases-types ne doit pas se fonder seulement sur des critères phonologiques et culturels mais aussi sur des critères quantitatifs.

Une fois l'alphabet maîtrisé, l'élève doit s'exercer à lire. Mais que lit-il ? Nous suggérons que le vocabulaire de base serve à l'élaboration de textes simplifiés qui serviraient de matériel de lecture pour les lecteurs débutants.

3.1.2. VOCABULAIRE DE BASE ET TEXTES SIMPLIFIÉS

Les ouvrages de l'école primaire proposent un matériel de lecture allant d'une variété de phrases à des textes. Le vocabulaire de base peut aider l'enseignant à utiliser dans les phrases à lire des vocables à haut indice d'usage et à proposer des textes de lecture simplifiés fondés dans un premier temps sur des vocables à indice d'usage élevé, puis englobant plus tard des vocables dont l'indice d'usage est faible. De tels textes, enrichis d'autres procédés (images, taille des lettres, etc.) débouchent sur la littérature pour enfants (Beccaria 1976); une littérature encore à écrire pour une société à tradition orale comme l'est la société du Burundi.

L'élaboration de textes simplifiés opère par la simplification du vocabulaire utilisé. L'on procède soit en remplaçant les mots moins fréquents par ceux qui sont plus fréquents, soit en paraphrasant le mot difficile. L'on recourt aussi à :

- la régularisation : faire ressembler un texte à un texte standard;
- l'explicitation : ajouter des mots au texte pour le rendre facile à lire;
- la suppression de mots moins fréquents;
- la répétition : l'on répète un vocable au lieu d'utiliser un para-synonyme plus difficile.

Pour plus d'informations sur la simplification des textes et l'élaboration du matériel de lecture, l'on consultera Nation (1992 : 177-189), Beccaria (1976), et McCullough & Chacko (1976).

Voici à titre d'exemple, des phrases fondées sur des vocables à haut indice d'usage. Nous fournissons la traduction en vis-à-vis.

(175)	*Peetéro ni umunyéeshuúre.*	'Pierre est un élève.'
	Yiiga néeza.	'Il étudie bien.'
	Arakúunda abavyéeyi bíiwé.	'Il aime ses parents.'
	Arabáfasha gukóra.	'Il les aide à travailler.'
	Baramútuma kuraaba umwáana.	'Ils l'envoient surveiller l'enfant.'

Voici aussi un extrait du journal *Ndongozi* (w1 p2) dont nous proposons une version simplifiée en utilisant les vocables à indice d'usage supérieur à 3.

(176)	*Haábaaye umugabo*	'Il était une fois un homme'
	n'úmukényezi wiíwé	'et sa femme'
	baákundana	'qui s'aimaient
	urw'úmugabo n'úmugóre.	'vraiment.'
	Umuúnsi umwé,	'Un jour'
	umugóre abwiira umugabo wiíwé ati	'la femme confia à son mari : '
	mugábo wanje ndagúkunda	'chéri, je t'aime beaucoup,'
	si nzóopfá ngúhemukiye	'je ne te tromperai jamais'
	kaándi nzokugumiriza ibanga.	'et je te serai toujours fidèle'.
	Búkeeye	'Peu de temps après',
	wáa mugabo acudika	'l'homme se lia d'amitié'
	n'úmusirikare. [...]	'avec un militaire [...]'

Si l'on remplace certains vocables et expressions à faible indice d'usage par des vocables à indice d'usage élevé et si l'on modifie la longueur des phrases, l'on obtient le texte suivant :

(177)	*Haábaaye umugabo*	'Il était une fois un homme
	n'úmugóre.	et une femme'.
	Baárakúundana caane.	'Ils s'aimaient beaucoup'.
	Umuúnsi umwé,	'Un jour'
	umugóre abwiira umugabo ati :	'la femme dit à son mari':
	ndagukunda	'Je t'aime beaucoup'
	si nzóopfá ngíiye ku wuúndi mugabo.	'je te serai toujours fidèle'.
	Uwo mugabo yarí afíse umugenzi	'Le mari avait un ami'
	w'úmusoda. [...]	'qui était soldat [...]'

De tels textes constituent, pour les enfants, des lectures simples dont le vocabulaire n'est pas difficile. Mais plus l'apprenant progresse, plus son vocabulaire s'étend. Les textes peuvent alors contenir des vocables plus rares dont le sens peut être inféré à partir du contexte. L'enseignant a la latitude de puiser dans la liste des vocables moins fréquents pour enrichir les textes soumis à la lecture.

3.2. À l'ÉCOLE SECONDAIRE

À l'école secondaire, le kirundi est une matière d'enseignement et non une langue d'enseignement. Tous les cours se donnent en français à l'exception du cours de kirundi.

Dans le secondaire, les élèves ont un niveau avancé en kirundi. L'enseignement du vocabulaire en classe de kirundi L1 s'oriente alors essentiellement vers les vocables à faible indice d'usage, les vocables métalinguistiques et les vocables reliés à un domaine d'activité (McCarthy 1990). Il porte aussi sur la structure morpho-sémantique des vocables (les dérivés et les composés).

3.2.1. L'UTILISATION DES VOCABLES À FAIBLE INDICE D'USAGE

Le vocabulaire à faible indice d'usage permet à l'enseignant d'enrichir le vocabulaire des étudiants d'unités lexicales rares et morphologiquement plus complexes. En partant, par exemple, du schéma 2, l'enseignant pourra l'enrichir avec les vocables nominaux suivants dont U est inférieur à 3 :

(178)	*baamwáana*	'beaux-parents'
	inábukwe	'belle-mère'
	umukámakare	'vieille femme'
	bwaanashaámba	'agronome'
	incáabwéenge	'intellectuel'
	imódoka	'automobile'
	igaáriyamoóshi	'train'
	ihúutihúuti	'empressement'
	ikáazé	'bienvenue'
	imasuwá	'bateau'
	umuzimyamuriro	'voisin'
	amamírwangóhe	'tard dans la soirée'
	ibitóongati	'plantes décoratives'
	intíicantiíkize	'quantité insignifiante'

agashíinguuracúmu	'la dernière bière servie'
ikibagabaga	'un adjoint'
ivyúunyunyú	'sels minéraux
ubuvúukagíhugu	'citoyenneté'
imváamahaánga	'un étranger'
ikiréengazúuba	'coucher du soleil'
umwuúzukuruza	'arrière-petit-enfant'

3.2.2. LES VOCABLES MÉTADISCURSIFS

Le vocabulaire métadiscursif est utilisé notamment pour formuler des consignes relatives aux tâches à accomplir en classe de kirundi ou pour désigner les parties du texte. Certains vocables métadiscursifs sont présents dans notre vocabulaire fondamental (dans les tranches dont U < 3). L'enseignant pourra donc les sélectionner dès qu'il aborde une tâche qui les requiert. Ce sont :

(179)

impfunyapfunyo	'résumé'	*incáamaké*	'synthèse'
intáangamáraara	'introduction'	*intáango*	'début'
ivyeérekeye	'ce qui est relatif à'	*icíiyumviiro*	'idée'
ibiráaba	'ce qui concerne'	*gushígikira*	'soutenir'
gusíguura	'expliquer'	*gutaahuura*	'comprendre'
gutoomoora	'expliciter'	*guhéraheza*	'compléter'
gusóbaanura	'expliquer en détail'		
gutáandukanya	'différencier'		
gutoohooza	'mener une recherche'		

3.2.3. LES VOCABLES RELIÉS À UN DOMAINE D'ACTIVITÉ

Il s'agit principalement de vocables reliés à l'agriculture, à l'élevage, à la santé et au développement rural. L'enseignant consultera le vocabulaire de base pour repérer les vocables reliés au thème qu'il aborde. Pour l'agriculture par exemple, il y aura :

- les noms de saisons agricoles : *impeéshi* 'petite saison sèche', *umutaasuro* 'début des pluies', *icí* 'saison sèche', etc.
- les types de sols : *urubuye* 'sol pierreux'
- les nutriments : *ibitabizo* 'engrais', *amasé* 'bouse', etc.
- l'encadreur agricole : *bwaanashaámba* 'agronome'
- les étapes de la maturation des cultures : *urushúurwé* 'floraison', *urubimba* 'gousses', etc.
- les problèmes et les solutions : *inkúkuura* 'érosion', *imiserége* 'canal'.

3.2.4. LA STRUCTURE MORPHO-SÉMANTIQUE DES VOCABLES

À des niveaux avancés, comme au collège, les étudiants pourraient aborder la structure morpho-sémantique des vocables. L'enseignant pourrait par exemple former à partir du vocabulaire de base des mots morphologiquement complexes et amener les étudiants à en découvrir le sens sans consulter le dictionnaire. On pourrait aussi partir d'un vocable du vocabulaire de base et constituer une liste de tous les dérivés et composés dans lesquels il se retrouve.

Il faut faire remarquer ici que l'étude de la structure morpho-sémantique des vocables peut et doit même commencer à l'école primaire. S'il est vrai que l'écolier possède une connaissance non explicite des règles morpho-sémantiques, cela n'empêche pas l'enseignant de lui proposer divers exercices adaptés à son niveau cognitif. L'enseignant pourrait par exemple demander à l'apprenant d'encercler dans une liste de substantifs à base verbo-nominale ceux qui désignent des personnes. Cet exercice mettrait à contribution les connaissances de l'enfant sur le suffixe agentif [-yi] en kirundi (cf. § 3.2.2.1).

3.3. EN ALPHABÉTISATION DES ADULTES

Le programme du Ministère des Affaires Sociales prévoit pour les adultes qui n'ont pas eu la chance de fréquenter l'école un programme d'alphabétisation. Ce programme intègre des thèmes sur le développement tel que la consommation de l'eau potable, la limitation des naissances, etc.

Selon une typologie proposée par Halaoui (1999), il s'agit d'une alphabétisation spécialisée (dite parfois fonctionnelle). Le but visé est l'enseignement de la lecture, de l'écriture et du calcul, enseignement sur lequel viennent se greffer des connaissances sur des thèmes choisis.

Généralement, l'alphabétisation se déroule en trois temps : la préalphabétisation, l'alphabétisation proprement dite et la postalphabétisation (Halaoui 1999). Dans le premier temps, on vise à préparer l'analphabète à l'alphabétisation proprement dite où se fait l'apprentissage de la lecture, de l'écriture et du calcul. Quant à la postalphabétisation, elle vise la consolidation des connaissances acquises et parfois à l'acquisition de nouvelles connaissances spécialisées.

Nous nous proposons d'illustrer l'utilisation du vocabulaire de base sélectionné en alphabétisation proprement dite et en postalphabétisation. L'on comprendra que notre propos est centré sur la lecture et l'écriture.

Rappelons que l'adulte maîtrise la langue orale et qu'il a essentiellement besoin d'apprendre à lire et à écrire. La question est donc double : que doit-il lire ? et que doit-il écrire ? Et en quoi le vocabulaire de base aidera-t-il ?

Il faut d'abord souligner que le vocabulaire de base présente l'avantage de fournir un contenu lexical aux cours d'alphabétisation; cela est d'autant plus important que dans le programme actuel du Ministère des Affaires Sociales[6], il est difficile de savoir, pour le niveau lexical, ce que l'alphabétisé connaît ou ne connaît pas. Le vocabulaire de base pourrait aussi aider dans l'élaboration d'outils de postalphabétisation et d'évaluation.

---

[6] Nous l'avons expérimenté en tant que formateur des alphabétiseurs de 1990 à 1991.

3.3.1. QUOI LIRE?

L'on peut distinguer deux moments : le début de l'alphabétisation et la période où l'alphabétisé doit asseoir ses connaissances afin de ne pas retomber dans l'illétrisme.

En début d'alphabétisation, le vocabulaire de base aide dans l'élaboration des phrases-types. Celles-ci sont formées de vocables les plus fréquents de la langue, afin que l'apprenant puisse les retrouver dans presque tous les textes écrits.

Des études sur la lisibilité (McCullough & Chacko 1976 : 174-175, Boyer 1987) ont montré en effet que la répétition des formes lexicales connues augmente la rapidité de l'acquisition de la lecture. Il s'agit en fait de répétition de mêmes stimuli visuels.

Lorsque l'adulte connaît son alphabet, il a besoin de textes simplifiés pour soutenir sa motivation à lire. Ici aussi le vocabulaire de base aide à mettre au point ces textes.

Enfin, le vocabulaire de base permet de renforcer l'aspect fonctionnel de l'alphabétisation en fournissant des vocables moins fréquents qui renvoient souvent à des réalités politiques, économiques, médicales, judiciaires, etc. dont l'alphabétisé a besoin. Pour le domaine judiciaire, par exemple, on a :

(180)	*ingingo*	'article'
	umucaáamanza	'juge'
	umushíkirizamaánza	'procureur'
	séentare ntahinyúzwa	'Cour suprême'
	ibwíirizwa nshingiro	'constitution'
	itégeko-bwiírizwa	'décret-loi'
	itégéko mpánavyáaha	'code pénal'
	itégeko ngendérwakó	'loi-cadre'
	séentáre rúbaámba	'chambre criminelle'
	gasáambuuramaánza	'Cour de cassation'

Le matériel de lecture devrait aussi comprendre des textes que l'alphabétisé «fréquentera» vraisemblablement dans sa vie. On peut citer : la lettre d'ami, les contrats, la convocation, les communiqués-radio, les contes, etc. De même, il faudra pour la postalphabétisation :

- élaborer des textes simplifiés pour soutenir la lecture des adultes récemment alphabétisés qui ne maîtrisent pas encore la lecture et l'écriture du kirundi;
- rendre disponibles des écrits de vulgarisation scientifique dans des domaines d'importance comme la santé, la justice, le développement rural (cf. liste de vocables pour les confectionner);
- ultimement, rendre disponibles les journaux en kirundi.

3.3.2. QUOI ÉCRIRE ?

Dès que l'alphabétisé connaît son alphabet se pose la question des activités motivantes qui puissent l'amener à continuer d'écrire. Outre la dictée traditionnelle, nous suggérons des sujets d'écriture qui cadrent avec les besoins de l'alphabétisé et qui peuvent faire intervenir certaines unités du vocabulaire de base. On peut citer comme pour la lecture : les lettres d'amis, les baux et les contrats, la convocation, les communiqués-radio, etc. L'on pourrait aussi initier des journaux ruraux où les alphabétisés seraient les « journalistes ».

4. VOCABULAIRE DE BASE ET DIDACTIQUE DU KIRUNDI L2

S'il est un domaine où les vocabulaires de base ont servi le plus, c'est dans l'enseignement des langues secondes. Rappelons pour mémoire Gougenheim *et al.* (1964) pour le français et Thorndike & Lorge (1944) pour l'anglais.

L'on peut distinguer, pour simplifier, deux types de vocabulaires : un vocabulaire général (composé de vocables de haute fréquence et de basse fréquence) et un vocabulaire spécialisé.

Le choix du vocabulaire à privilégier dépend du profil du débutant en L2 et de ses besoins. Un coopérant qui travaille au Ministère de l'Agriculture du Burundi aura par exemple besoin, en plus du vocabulaire général, de vocables relatifs aux saisons culturales, aux types de sols, etc. alors qu'un touriste pourrait bien se contenter d'un vocabulaire général.

Il est donc de première importance de définir le profil du débutant en kirundi L2. Nous renvoyons à Bérard (1991 : 33-43) pour un inventaire des types de profils.

Nous partirons ici du cas assez commun où un débutant en kirundi L2 n'a pas besoin de vocabulaire spécialisé. Il a par contre besoin d'un vocabulaire général pour s'exprimer à l'oral et à l'écrit mais aussi pour lire le journal et échanger avec ses interlocuteurs. Bref, il a besoin d'un vocabulaire à haut indice d'usage, qui permet de développer sa compétence à écouter, parler, lire et écrire.

La taille de ce vocabulaire général varie selon les recherches. Pour l'anglais par exemple, il est estimé à environ 3 000 vocables (Nation 1992 : 5). Pour le français, Gougenheim *et al.* (1964) fournissent une liste de 1 475 vocables tandis que Matoré (1963) en retient 5 000. Il est, pour le kirundi, constitué de 4 025 vocables.

À l'intérieur des 4 025 vocables de base, nous établissons des niveaux afin d'obtenir une certaine gradation du vocabulaire. Cette gradation correspond en didactique des L2 aux « niveaux », ceux-ci allant du niveau des débutants à celui des avancés.

Les 4 025 vocables du kirundi écrit se répartissent donc comme suit :

Niveau	*V*	*% de V*	*N*	*% de N*	*Usage*
Débutants	1 288	32,00%	90 334	87,22%	≥ 3
Intermédiaires	1 905	47,3%	4 649	4,48%	$3 > U \geq 0$
Avancés	832	20,6%	1 124	1,08%	$U < 0$

Tableau 39 - *Répartition des vocables selon le niveau des apprenants*

Aux §§ 4.1 et 4.2, nous nous proposons de voir comment ce vocabulaire peut être utilisé pour développer la compétence à écouter, parler, lire et écrire.

4.1. ÉCOUTER ET PARLER

Généralement, le développement de l'écoute requiert l'utilisation de textes divers : chansons, informations, débats, etc. Mais pour des débutants en kirundi L2, l'on pourrait suggérer des exercices d'écoute de textes simplifiés et des exercices à trou.

De plus, le kirundi étant une langue à ton et à quantité vocalique, des exercices sur les paires minimales permettraient de vérifier si l'apprenant distingue les vocables qu'il entend[7]. L'on pourrait par exemple lui lire des vocables et il ferait correspondre les vocables à des images. L'enseignant trouvera dans la liste du vocabulaire de base les unités lexicales à utiliser.

Afin de développer les capacités d'expression orale, l'on peut proposer à l'apprenant :

- des exercices de répétion des vocables tirés de la liste;
- des exercices de substitution où entrent en jeu des vocables sélectionnés selon leur usage : ex. *inyuma* 'derrière', *imbere* 'devant', *heejuru* 'au-dessus', *munsí* 'en dessous';
- de décrire et de faire deviner aux pairs des objets ou des personnes dénotés par des vocables de base;
- de confectionner une banque d'images correspondant aux vocables de base et de faire retrouver une image à partir d'une description;
- de paraphraser certains vocables ou de remplacer des vocables par des para-synonymes contenus dans le vocabulaire de base;
- des exercices sur les tons et la quantité vocalique axés sur la répétition et le jeu. L'on pourrait par exemple demander à la classe de montrer l'image de l'objet ou de l'action dénotée par le vocable prononcé par l'apprenant.

---

[7] L'on a par exemple les paires de vocables verbaux suivants :

[-sib-] 'proférer des insultes' (U = 3,77) *versus* [-siib-] 's'absenter' (U = 13,63)
[-yag-] 'fondre' (U = -0,05) *versus* [-yaag-] 'converser' (U = 22,51)

Notons ici que tous ces exercices peuvent prendre des formes différentes selon l' approche didactique adoptée. Dans une approche structurale par exemple, les exercices ont un rôle premier; ils sont souvent mécaniques et laissent peu de place à la créativité de l'apprenant (Bérard 1991:13).

Par contre, dans une approche communicative, les exercices ont un rôle d'appoint, de régulateur. Ils favorisent les productions langagières des apprenants et les aident à surmonter leurs blocages (Bérard 1991 : 44). Dans ce contexte, on peut recourir à certaines techniques comme le jeu de rôle où les apprenants utilisent le vocabulaire connu à des fins diverses : donner un ordre, demander quelque chose, exprimer une émotion, etc. (Germain 1993 : 201-215, Bérard 1991 : 56, 95-99).

4.2. LIRE ET ÉCRIRE

La compétence à lire du débutant en kirundi L2 se développe notamment par la lecture de textes simplifiés. Le but poursuivi est de fournir à l'apprenant un matériel de lecture correspondant à son niveau de développement linguistique.

L'on estime pour l'anglais qu'un vocabulaire d'environ 3 000 vocables suffit pour couvrir les besoins de lecture d'un apprenant de niveau 6 (avancé) en anglais L2 (Nation 1992 : 116). Nous ne disposons pas de données exactes pour le français. On peut cependant constater que le dictionnaire fondamental de Gougenheim (1961) comporte 3000 vocables, celui de Matoré (1963) 5 000 et celui de Juilland *et al.* (1970) 5 082.

Sur les divers aspects de l'élaboration de textes simplifiés, l'on consultera notamment Nation (1992 : 177-189). Pour ce qui est spécifiquement du plan lexical, l'on se reportera à l'exemple proposé en (169) pour s'en faire une idée.

L'existence d'un vocabulaire de base permet d'opérer une simplification objective de textes.

Outre la lecture de textes simplifiés, les activités de lecture exploitant le vocabulaire de base peuvent être des textes à trous que l'apprenant complète avec des listes de vocables fournis tirés du vocabulaire de base, des exercices de substitution de vocables, etc.

Il va sans dire que le matériel de lecture tiendra compte des caractéristiques phonologiques de la langue : les tons et la quantité vocalique. L'on se reportera aux

activités suggérées pour l'écoute et l'expression orale, pour développer la discrimination tonale et celle de la longueur vocalique.

Il nous faut rappeler ici que l'orthographe actuelle du kirundi ne note pas les tons et la quantité vocalique. Faut-il alors introduire ces deux aspects de la langue dans les cours de kirundi L2 (ou même L1) alors que les apprenants ne les retrouveront pas dans des documents de lecture authentiques ?

Selon Forges & Mayugi (1988 : 126-127), la réponse est affirmative car « *l'absence de notation de la tonalité et de la quantité amène fréquemment des quiproquos. [...] Une décision officielle (devait être) prise* [...] ».

Halaoui (1999 : 4) évoque la même difficulté en ces termes :

> *« Les tons n'étant pas notés, on ne saurait demander à l'apprenant qui rencontre l'un de ces mots dans la lecture d'une phrase, d'avoir immédiatement à l'esprit tous les autres mots et de réaliser celui qui lui semble s'accorder du point de vue du sens avec les autres mots de la phrase ».*

Pour le kirundi, l'absence de notation des tons et de la quantité vocalique dans les textes nous semble liée à deux raisons :

- les recherches en phonologie du kirundi sont récentes et peu diffusées;
- de nombreux lecteurs (dont des enseignants et autres fonctionnaires) n'ont pas été formés à la transcription et à la lecture de la tonalité et de la quantité vocalique.

Le développement de la compétence en écriture du kirundi L2 passe aussi par la maîtrise des graphèmes et par l'emploi correct des vocables dans des phrases. À ce niveau, l'orthographe du kirundi pose des problèmes qui sont spécifiques dont notamment :

- des graphèmes particuliers à la langue : ex. *ts, pf*, etc.
- des suites graphiques particulières : ex. *mbw, ndw, nshw*, etc.

L'on contournera ces difficultés en présentant d'abord les vocables à structure phono-graphique simple, les autres devant être introduits graduellement et de manière limitée.

Quant à la syntaxe, elle est marquée en kirundi par l'accord de classe (cf. tableau 12, p.110). Cet aspect est à exploiter à l'écrit. Les vocables seront utilisés dans des phrases simples en privilégiant les plus fréquents à l'intérieur des classes d'accord.

Ici, les données morphologiques fournies au chapitre 4 § 3 pourraient être d'une grande utilité pour sélectionner les éléments morphologiques à utiliser dans les exemples. La liste des vocables de base permettrait ainsi de sélectionner les éléments lexicaux, les données fréquentielles sur la morphologie permettant quant à elles de sélectionner les affixes à retenir en priorité. Un tel couplage aboutirait à des exemples qui utiliseraient les vocables et les morphèmes fréquents de la langue. Par exemple, il est plus pertinent d'utiliser le radical [-gabo] 'homme' avec l'adjectif [-níni] 'grand' et le préfixe de classe [-mu-]$_{cl.1}$ qu'avec le préfixe [-ru-] moins fréquent.

(181) *umugabo* *muníni ararima* 'un homme de grande taille cultive'
abagabo *baníni bararima* 'des hommes de grande taille cultivent'
urugabo *runíni rurarima* 'un homme de grande taille' [+péjoratif]

De nombreuses activités d'écriture sont possibles à partir des vocables les plus usuels : identification et décodage des relations lexicales impliquant seulement des vocables de base à partir d'un thème donné, reproduction des vocables vus pendant un court laps de temps, des substitutions de mots dans un texte de manière à garder les accords intacts et sans trop altérer le sens de la phrase, etc.

5. CONCLUSION

Le chapitre 5 nous a permis d'esquisser quelques pistes d'exploitation du vocabulaire de base du kirundi écrit en didactique du kirundi L1 ou L2.

Comme on a pu le constater, la liste est une donnée brute que l'enseignant peut utiliser pour enrichir le vocabulaire des apprenants. Cet enrichissement doit tenir compte des besoins des élèves et utiliser un ensemble de techniques variées comme le champ sémantique, les relations lexico-sémantiques et les collocations.

Dans le primaire, en alphabétisation des adultes et en L2, les principales utilisations sont liées à l'élaboration des phrases-types, à la simplification des textes et à l'élaboration du matériel de lecture et d'écriture.

À l'école secondaire, le vocabulaire de base peut être utilisé dans l'enrichissement lexical, activité fondée sur des vocables à faible indice d'usage, des vocables métalinguistiques, des vocables techniques et une analyse de la structure morpho-sémantique des vocables.

Une telle exploitation de la liste des vocables de base du kirundi écrit est idéalement à lier à l'exploitation d'une base de données textuelles.

Signalons également que l'élaboration de la liste du vocabulaire de base constitue une assise solide pour l'élaboration d'une banque d'images qui servirait dans les classes de niveau primaire ou dans les cours pour débutants en kirundi L2.

CONCLUSION GÉNÉRALE

La présente recherche visait à élaborer un vocabulaire de base du kirundi à partir d'un corpus écrit d'environ 100 000 mots-formes. Du même coup, elle devait permettre de prendre la mesure des difficultés que posent la lexicométrie du kirundi et le traitement automatique de la morphologie de la langue.

L'étude opère donc à deux niveaux d'analyse : le niveau lexical pour les aspects lexicométriques et le niveau morphologique pour le traitement automatique de la morphologie du kirundi.

Au niveau lexical, notre objectif était de dégager, à partir des 103 561 mots-formes du corpus, une liste de vocables de base du kirundi écrit. Pour y parvenir, il nous fallait régler préalablement un certain nombre de problèmes qui se posent en statistique lexicale du kirundi, puis à opérer la sélection des vocables.

Une des principales difficultés en lexicométrie du kirundi réside dans les nombreuses interventions manuelles à mener sur le corpus. Elles visent notamment à restituer les tons et la quantité vocalique aux mots-formes du corpus car les journaux dépouillés ne notent pas ces deux aspects de la phonologie de la langue. À cette difficulté s'ajoute la désambiguïsation manuelle de nombreux homographes lexicaux, syntaxiques et morphologiques qui jalonnent le corpus.

Il a fallu également procéder à la séparation du morphèmes dicto-modal (*nti* 'ne...pas') et des morphèmes locatifs (*ho* 'y', *ko* 'sur' *mwo* 'dans', *yo* 'y') avec le verbe et restituer les voyelles élidées aux mots-formes.

Une autre difficulté réside dans le choix des lemmes. Dans l'ensemble, nous avons retenu les options de notre dictionnaire de référence (Rodegem 1970). Nous avons ramené les mots-formes verbaux et adjectivaux à leurs radicaux. Pour les substantifs, nous avons retenu comme lemme la forme au singulier et à la classe neutre, tandis que les mots-formes grammaticaux ont été lemmatisés sur une base formelle; les fréquences des homographes étant établies à partir des index fournies par *WordCruncher*.

Grâce aux index des mots-formes fournis par *WordCruncher* et à ceux des radicaux fournis par l'analyseur morphologique du kirundi, nous avons pu dégager pour tout le corpus un vocabulaire de 4 025 vocables répartis en trois tranches selon l'indice d'usage (U). La première tranche regroupe les vocables dont $U \geq 3$, la deuxième ceux dont $3 > U \geq 0$ et la troisième ceux dont $U < 0$. La première tranche constitue le noyau du vocabulaire de base.

En regard de nos hypothèses de départ, nous avons pu constater que, comme pour les autres langues, les mots grammaticaux connaissent les plus hautes fréquences. Sur les 50 vocables les plus fréquents, les vocables grammaticaux sont au nombre de 30.

Si l'on considère la fréquence des catégories grammaticales, les mots grammaticaux constituent - tout comme pour le français et l'anglais - la catégorie la plus fréquente suivie de celle des substantifs, puis de celle des verbes et enfin des adjectifs.

Quant à la stabilité des fréquences des catégories grammaticales en kirundi, stabilité exprimée par un coefficient de variation, nous avons constaté que les mots grammaticaux constituent la catégorie la plus stable. Viennent ensuite les substantifs et les verbes et enfin les adjectifs. En français oral (Gougenheim *et al.* 1964), ce sont les verbes qui sont les plus stables suivis des mots grammaticaux, des adjectifs et enfin des substantifs.

Nous avions également repris l'hypothèse généralement admise que les formes simples sont plus fréquentes que les formes complexes. Nous avons constaté à cet égard que les substantifs à base nominale, moins complexes du point de vue morphologique, sont plus fréquents que les substantifs à base verbo-nominale dont la formation résulte de la concaténation de plusieurs morphèmes.

Rappelons aussi que les mots-formes verbaux non suffixés sont plus nombreux que les suffixés (70%). Il en est de même pour les substantifs à base verbo-nominale non suffixés qui constituent un peu plus de 60% (4 320 / 6 892) des substantifsà base verbo-nominale.

Nous pouvons dire donc que dans l'ensemble, nos résultats confirment les hypothèses émises au niveau lexical. Qu'en est-il du niveau morphologique ?

L'analyse morphologique automatique du corpus avait pour but d'en dégager les morphèmes les plus fréquents de manière à compléter les index de vocables par des index de morphèmes. À ce niveau, l'analyseur morphologique du kirundi a permis d'effectuer un travail manuellement irréalisable.

La complexité morphologique de la langue constitue la principale difficulté qui entrave l'analyse morphologique automatique. Malgré une bonne performance de l'analyseur, nous avons dû mener de nombreuses interventions manuelles notamment pour régler les cas de réduplication du radical, de variantes lexicales et de paires de radicaux dont l'un ressemble formellement à un dérivé de l'autre.

L'analyse morphologique automatique a porté sur un ensemble de morphèmes du verbe et du substantif.

Pour le verbe, l'analyse a porté sur les marqueurs aspectuels et les suffixes de dérivation.

Des trois marqueurs aspectuels du kirundi, le morphème de l'inaccompli [-a] est le plus fréquent. Viennent ensuite le morphème de l'accompli [-ye] et celui de l'inaccompli inchoatif [-e].

Quant aux suffixes de dérivation, les plus fréquents sont dans l'ordre l'applicatif [-ir-], le passif [-u-], le causatif [-i-] et l'associatif [-an-].

Pour les substantifs, l'analyse a porté sur les préfixes de classes, les suffixes de dérivation et les dérivatifs thématiques nominaux.

Si l'on considère les préfixes de classes, les plus fréquents sont dans l'ordre [- ri-] cl.5, [- mu-] cl.1 & cl.3, [- ba-] cl.2, [-n-] cl.9 & cl.10 et [-ki-] cl.7.

Quant aux suffixes de dérivation, les plus fréquents sont dans l'ordre le passif [-u-], l'applicatif [-ir-], l'associatif [-an-] et le causatif [-i-].

Enfin, les dérivatifs thématiques nominaux les plus fréquentes sont [-a] et [o]; ils dominent sensiblement les autres.

Ces résultats sont susceptibles de nombreuses applications dans divers domaines, particulièrement en didactique du kirundi langue maternelle et langue seconde.

En didactique du kirundi langue maternelle, le vocabulaire de base peut être utilisé pour l'élaboration des phrases-types, point de départ de l'enseignement de la lecture et de l'écriture ainsi que pour la mise au point de textes simplifiés, outils importants pour l'apprentissage de la lecture.

Le vocabulaire de base peut également être mis à contribution dans l'enseignement systématique du vocabulaire au secondaire et dans la vulgarisation scientifique.

En didactique du kirundi L2, le vocabulaire sélectionné pourrait servir de base à de nombreux exercices visant développer les habiletés à l'écoute, à l'expression orale, à la lecture et à l'écriture.

Si cette recherche a permis de jeter quelques lumières sur certains aspects lexicaux et morphologiques du kirundi, bien des aspects n'ont pas pu être abordés.

Nous n'avons par exemple pas pu procéder à la comparaison des vocabulaires des deux journaux *Ndongozi* et *Ubumwe* dont est tiré notre corpus. Une étude sur la richesse lexicale des deux journaux permettrait d'en avoir une idée différenciée.

Également, nous n'avons que très sommairement abordé les relations lexico-sémantiques à l'intérieur de notre corpus. Sur la base des vocables sélectionnés

et de bien d'autres de la langue, on pourrait prolonger l'investigation sur la synonymie, l'antonymie et les relations hiérarchiques.

Du point de vue didactique, l'utilisation du vocabulaire de base serait idéalement à lier à l'exploitation d'une base de données textuelles. Bien plus, la liste constitue une assise solide pour l'élaboration d'une banque d'images qui servirait dans les classes de niveau primaire ou dans des cours pour débutants en kirundi L2.

RÉFÉRENCES

Alegria *et al.* 1996. « Automatic morphological analysis of Basque », *Literary and linguistic computing*. Volume 11 (4). 193-205.

Anglin, J.M. 1993. *Vocabulary development : a morphological analysis*. Chicago : University of Chicago Press.

Anis, J. *et al.* 1988. *L'écriture, théories et descriptions*. Bruxelles : De Boeck-Wesmael.

Aristizabal, G.C. 1938. *Détermination expérimentale du vocabulaire écrit pour servir à l'enregistrement de l'orthographe à l'école primaire*. Thèse de doctorat. Louvain : Université de Louvain.

Aronoff, M. 1976. *Word formation in generative grammar*. Cambridge : M. I. T. Press.

Attar, R. *et al.* 1978. « Linguistic tools for retrieval systems », *Journal of the association for computing machinery*. 25 : 52-66.

Auchlin, A. J. *et al.* 1981. « Réflexions sur les marqueurs de structuration de la conversation », *Études de linguistique appliquée*. 44 : 88-103.

Baayen, R.H. 1989. *A corpus-based approach to morphological productivity. Statistical analysis and psycholinguistic interpretation*. Amsterdam : Centrum voor Wiskunde en Informatica.

Bagnal, N. 1993. *Newspaper language*. Oxford : Butterworth-Heinemann.

Barth, G. 1961. *Recherches sur la fréquence et la valeur des parties du discours en français, en anglais et en espagnol*. Paris.

Baudot, J. 1992. *Fréquences d'utilisation des mots en français écrit contemporain*. Montréal : Presses de l'Université de Montréal.

Bauer, L. 1994. « Semantic categories », *dans* Asher, R.E. (ed.). 1994. *The encyclopedia of language and linguistics*. Volume 7. Oxford / New York / Seoul / Tokyo : Pergamon Press. 3799-3800.

Baylon C. & P. Fabre. 1978. *Grammaire systématique de la langue française*. Paris : Nathan.

Baylon, C. & P. Fabre 1978. *La sémantique*. Paris : Fernand Nathan.

Beauchemin, N. & M. Théoret. 1984. « Micro-Solivo : un lemmatiseur semi-automatique pour le québécois parlé », *Revue de l'association québécoise de linguistique*. 3 (3). 19-38.

Beauchemin, N. & P. Martel. 1979. *Vocabulaire fondamental du québecois parlé*. Sherbrooke : Université de Sherbrooke.

Beauchemin N. *et al.* 1983. *Vocabulaire du québecois parlé en Estrie. Fréquence, dispersion, usage*. Sherbrooke : Université de Sherbrooke.

Beauchemin, N. *et al.* 1992. *Dictionnaire de fréquence des mots du français parlé au Québec. Fréquence, dispersion, écart réduit*. New York / San Francisco / Bern / Baltimore / Frankfurt / Am Main / Berlin / Wiem / Paris : Peter Lang.

Beccaria, M.J. 1976. « La lisibilité dans un journal d'enfant : « Pomme d'Api », dans Bentolila, 1976. *Recherches actuelles sur l'enseignement de la lecture*. Paris : Communication et langage / Éditions Reitz. 192-202.

Bélanger, G. 1992. *Étude des relations cohésives grammaticales : perspective traductologique et typologique*. Thèse de doctorat. Sherbrooke : Université de Sherbrooke.

Bérard, E. 1991. *L'approche communicative. Théorie et pratiques*. Paris : CLE international.

Bloomfield, L. 1933. *Language*. New York : Holt, Rinehart and Winston.

Booij, G. 1977. *Dutch morphology : a study of word formation in generative grammar*. Dordrecht : Foris.

Boyer, J.-Y. 1987. *Le rôle de la redondance au niveau de la structure textuelle en lecture*. Thèse de doctorat. Ottawa : Université d'Ottawa.

Bradley, D.C. 1983. *Computational distinctions of vocabulary type*. Indiana : Bloomington.

Brown, P.F. *et al.* 1990. « A statistical approach to machine translation », *Computational linguistics*. 16. 79-86.

Brunet, É. 1981. *Le vocabulaire français de 1789 à nos jours : d'après les données du Trésor de la langue française*. Genève : Slatkine.

Brunet, É. 1983. *Le vocabulaire de Proust*. Genève : Slatkine-Champion.

Brunet, É. 1985. *Le vocabulaire de Zola*. Genève / Paris : Slatkine-Champion.

Brunet, É. 1988. *Le vocabulaire de Victor Hugo*. Genève : Champion-Slatkine.

Bybee, J.L. 1994. « Morphological universals and change », *dans* Asher, R.E. (ed.). 1994. *The encyclopedia of language and linguistics.* Volume 5. Oxford / New York / Seoul / Tokyo : Pergamon Press. 2557-2562.

C.N.R.S. 1971. *Dictionnaire des fréquences. Vocabulaire littéraire des XIX^e^ et X X^e^ siècle.* Paris : Didier. 7 volumes.

Carrière, L. 1952. *Le vocabulaire français fondamental.* Thèse de doctorat. Montréal : Institut Saint-Georges.

Carroll, J.B. *et al.* 1971. *The American heritage word frequency.* New York : American heritage publishing Co.

Carstairs-McCarthy, A. 1994. « Morphological universals », *dans* Asher, R.E. (ed.). 1994. *The encyclopedia of language and linguistics.* Volume 5. Oxford / New York / Seoul / Tokyo : Pergamon Press. 2553-2557.

Carter, R. & M. Mc Carthy. 1988. *Vocabulary and language teaching.* London / New York : Longmans.

Cheydleur, F. 1929. *French idiom list based on a running count of 1 183 000 words.* New York : MacMillan.

Clark, V.E. 1993. *The lexicon in acquisition.* Cambridge : Cambridge studies in linguistics.

Conquet, A. & F. Richaudeau. 1976. « Cinq méthodes de mesure de la lisibilité », *dans* Bentolila, A. 1976. *Recherches actuelles sur l'enseignement de la lecture.* Paris : Communication et langage / Éditions Reitz. 203-211.

Corbin, D. 1987. *Morphologie dérivationnelle et structuration du lexique.* Tübingen : Niemeyer.

Cordier, F. 1994. *Représentation cognitive et langage : une conquête progressive.* Paris : Armand Colin.

Cossette, A. 1994. *La richesse lexicale et sa mesure.* Paris : Honoré Champion éditeur.

Coste, D. *et al.* 1976. *Un niveau-seuil.* Neuchatel : Hatier.

Daniel, D. & D. Delas-Demon. 1979. *Dictionnaire des idées par les mots.* Paris : Robert.

De Chantal, F. 1997. *La transgression linguistique dans* Normance *et dans* Bagatelles pour un massacre *de Louis-Fernand Céline. Étude stylométrique.* Thèse de doctorat. Montréal : Université de Montréal.

De Colombel, V. 1986. *Phonologie quantitative et synthématique : propositions méthodologiques et théoriques avec application à l'ouldémé.* Paris : SELAF.

De Saussure, F. 1916. *Cours de linguistique générale.* [1972]. Lausanne : Payot.

Debièvre, M. 1977. « Essai d'application des méthodes de la statistique linguistique au problème posé par l'attribution du texte de la version française du roman de Tristan », *dans* David, J. & R. Martin. 1977. *Études de statistique linguistique.* Paris : Klincksieck.

Deweye, G. 1923. *Relative frequency of English speech sounds.* Cambridge.

Deweze, A. 1989. *Informatique documentaire.* Paris / Milan / Barcelone / Mexico : Masson.

Dhal, H. 1979. *Word frequencies of spoken American English.* Essex : Verbatim.

Diop, A. *et. al.* 1975. « La morphologie du Wolof fondamental. Évaluation d'une méthode statistique », *dans* CNRS. 1975. *Les langues sans tradition écrite. Méthode d'enquête et de description.* Paris : SELAF. 471-486.

Dottrens, R. & D. Massarenti. 1963. *Vocabulaire fondamental du français.* [3e édition]. Neuchâtel : Delachaux / Niestlé.

Dubois, J. 1962 a. *Étude sur la dérivation suffixale en français moderne contemporain.* Paris : Larousse.

Dubois, J. 1962 b. *Le vocabulaire politique et social en France de 1869 à 1872.* Paris : Larousse.

Dubois, J. *et al.* 1994. *Dictionnaire de linguistique.* Paris : Larousse.

Dubrocard, M. 1985. « Problèmes d'attribution : choix des éléments significatifs », *dans* CNRS. 1985. *Hommage à Pierre Guiraud.* Nice : Faculté de Nice. 187-191.

Dugast, A. 1980. *La statistique lexicale.* Genève : Slatkine.

Dyen, I. 1975. *Linguistic subgrouping and lexicostatistics.* The Hague : Mouton.

Ellegard , A. 1962. *A statistical method for determinng authorship, the Junius letters,* 1769-1772. Göteborg.

Évrard, É. 1966. « Étude de statistique sur les affinités de cinquante-huit dialectes bantous », *dans* Faculté des lettres et sciences humaines de Strasbourg. 1966. *Statistique et analyse linguistique.* Paris : PUF. 85-94.

Faik-Nzuji, C.P. 1992. *Éléments de phonologie et de morphophonologie des langues bantoues.* Louvain-La-Neuve : Peeters.

Faitelson-Weiser, S. & R. Gingras. 1992. « La disponibilité suffixale », *Langue et linguistique.* 18 : 39-66.

Forges, G. & N. Mayugi 1988. « Approche structuro-globale de la lecture et de l'écriture des éléments supra-segmentaux à portée phonologique », *Revue de phonétique appliquée*. (82-84) : 125-133.

Fortier, G. 1993. *Le vocabulaire des adolescents et des adolescentes du Québec. Fréquence, répartition, disponibilité.* Montréal : Les Éditions Logiques.

François, F. 1978. *Éléments de linguistique appliquée à l'étude du langage de l'enfant*. Paris : Éditions J.-B.Baillière.

Frandin, B. 1994. « L'approche à deux niveaux en morphologie computationnelle et les développements récents de la morphologie », *T.A.L.* Volume 35 (2). 9-49.

Galisson, R. 1971. *Inventaire thématique et syntagmatique du français fondamental*. Paris : Hachette - Larousse.

Geeraerts, D., S. Grondelaes & P. Bakema. 1994. *The structure of lexical variation. Meaning, naming and context*. Berlin / New York : Mouton de Gruyter.

Genouvrier, É. & J. Peytard. 1970. *Linguistique et enseignement du français*. Paris : Larousse.

Germain, C. 1993. *Évolution de l'enseignement des langues : 5 000 ans d'histoire*. Paris : Hurtbise HMH, Ltée - CLE international.

Gill, S.J. & A.M. Renker 1992. « From phoneme to text. An overview of Makah », *dans* Buchholzer, G.P. (ed.). 1992. *Amerindian Languages & Informatics*. Paris : Amerindia. 185-205.

Goldsmith, J. & F. Sabimana. 1989. « The kirundi verb », *dans* Jouannet, F. 1989. 19-63.

Goody, J. 1994. *Entre l'oralité et l'écriture*. Paris : PUF.

Gordon, B. 1983. « Lexical access and lexical decision : mecanism of frequency sensitivity », *Journal of verbal learning and verbal behavior*. 22. 24-44.

Gougenheim, G. *et al.* 1964. *L'élaboration du français fondamental (1er degré). Étude sur l'établissement d'un vocabulaire et d'une grammaire de base.* Nouvelle édition refondue et augmentée. Paris : Didier.

Gougenheim, G. 1961. *Dictionnaire fondamental de la langue française*. Paris : Didier.

Gratton, M. & P. Barbaud. 1981. « Vocabulaire, milieux sociaux et méthodes d'enseignement », *dans* Gagné, G. & M. Pagé. 1981. *Études sur la langue parlée des enfants québecois.* Montréal : Les presses de l'université de Montréal. 265-283.

Greenberg, J.H. (ed). 1966. *Universals of language.* Cambridge : MIT Press.

Greidanus, T. 1990. *Les constructions verbales en français parlé. Étude quantitative et descriptive de la syntaxe de 250 verbes les plus fréquents.* Tübingen : Max Niemeyer Verlang.

Greimas, A.J. & J. Courtés. 1993. *Sémiotique. Dictionnaire raisonné de la théorie du langage.* Paris : Hachette.

Gross, M. 1968. *Grammaire transformationnelle du français : syntaxe du verbe.* Paris : Larousse.

Guillaume, H. *et al.* 1978. *Démographie linguistique : approche quantitative.* Paris : SELAF.

Guiraud, P. 1955. *Index du vocabulaire du théâtre classique.* Paris : Klincksieck.

Guiraud, P. 1960. *Problèmes et méthodes de la statistique linguistique.* Paris : P.U.F.

Hachten, W.A. 1993. *The growth of media in the third word. African failures, asian successes.* Mes : Iowa State University Press.

Halaoui, N. 1999. « L'éducation de base en Afrique noire ». Montréal : Département de linguistique et de traduction. Université de Montréal. Manuscrit.

Halle, M. 1973. « Prolegomena to a theory of word formation », *Linguistic inquiry.* 4 : 3-16.

Hatzfeld, A. *et al.* 1964. *Dictionnaire général de la langue française, du commencement du XVII^e siècle jusqu'à nos jours.* 7^e édition. Paris : Delagrave.

Haton, J.-P. & M.-C. Haton. 1989. *L'intelligence artificielle.* Paris : P.U.F.

Haygood, J.D. 1936. *Le vocabulaire fondamental du français.* Paris : E. Droz.

Hazael-Massieux, M-C. 1993. *Écrire en créole. Oralité et écriture aux Antilles.* Paris : L'Harmattan.

Henmon, V.A.C. 1924. *French word book on a count of 400 000 running words.* Madison : University of Wisconsin Madison.

Hofland, K. 1992. *Word frequencies in British and American English.* Bergen / Harlow : Longman.

Horn, E. 1926. *A basic English vocabulary : 10 000 words most commonly used in writing*. Iowa.

Houis, M. 1977. «Plan de description systématique des langues négro-africaines», *Afrique et Langage*. 7.

Howes, D.H. 1966. « A word count of spoken English », *Journal of verbal learning and verbal behavior*. 5. 572-606.

Hug, M. (1979). *La distribution des phonèmes en français / Die phonem-verteilung im deutschen*. Genève : Slatkine-Champion.

Humphreys, L.R. 1992. « The simplified English », *dans* Tommola, H., K. Varantola, T. Salmi-Tolonen & J. Schopp (eds). 1992. *Euralex'92. Proceedings*. Part II. Tampere : Tempereen Yyliopisto. 353-363.

Imbs, P. 1960. *L'emploi des temps verbaux en français moderne. Essai de grammaire descriptive*. Paris : Klincksieck.

Johansson, S. & K. Hofland. 1989. *Frequency analysis of English vocabulary and grammar*. Volume1. Oxford : Clarendon Press.

Johnson D.D. *et al.* 1983. *The Gin word book for teachers. A basic lexicon. A reference tool for classroom teachers*. Ginn and Company.

Jolivet, R. 1982. *Descriptions quantifiées en syntaxe du français*. Genève / Paris : Slatkine.

Jouannet, F. 1989. *Modèles en tonologie (kirundi et kinyarwanda)*. Paris : C.N.R.S.

Juilland, A.& E. Chang-Rodriguez. 1964. *Frequency dictionary of Spanish*. London / Paris : Mouton & Co / The Hague.

Juilland, A. *et al.* 1970. *Frequency dictionnary of French words*. La Haye : Mouton.

Juilland, M. 1985. « Étude quantitative des champs sémantiques et morphosémantiques dans une oeuvre littéraire », dans Charpentier, C. & J. David. 1985. *La recherche française par ordinateur en langue et littérature*. Genève / Paris : Slatkine / Champion.

Kaalep, H.-J. 1997. « An estonian morphological analyser and the impact of a corpus on its development », *Computer and the humanities*. Volume 31 (2). 115-133.

Karlsson, F. 1992. « Swetwol : a comprehensive morphological analyser for Swedish », *Nordic Journal of linguistics*. Volume 15 (1). 1-45.

Karlsson, F. 1994. « Computational morphology », dans Asher, R.E. 1994. *Encyclopedia of language and linguistics*. Volume 5. Oxford / New York / Séoul / Tokyo : Pergamon Press. 2570 - 2573.

King, M. 1995. « Traitement automatique des langues : état des lieux », *dans* Huot, H. & H. Portine. 1995. *La linguistique appliquée aujourd'hui : problèmes et méthodes*. Amsterdam : De Werelt. 7-18.

Kinkade, M.D. 1992. « Dictionary appendices : somme Upper Chehalis Solutions », *dans* Buchholzer, G.P. (ed.). 1992. *Amerindian Languages & Informatics*. Paris : Amerindia. 31-45.

Kiparsky, P. 1982. *Explanation in phonology*. Dordrecht : Foris.

Koskenniemi, K. 1983. *Two level morphology : a general computational model for word-form recognition and production*. Helsinki : University of Helsimki.

Kucera, H. & W.N. Francis. 1967. *Computational analysis of present-day American English*. Providence : Brown university.

Labbé, D. 1990. *Le vocabulaire de François Mitterand*. Paris : Presses de la fondation nationale des sciences politiques.

Labov, W. 1970. « The logic of non standard English », *dans* Frederick, W. (ed.). 1970. *Language and poverty : perspective on a theme*. Chicago : Markham. 153-189.

Labov, W. 1972. *Sociolinguistics patterns*. Philadelphia : University of Pennsylvania Press.

Lafon, P. 1984. *Dépouillement et statistique en lexicométrie*. Genève-Paris : Slatkine-Champion.

Lafon *et al.* 1985. *Le machinal. Principes d'enregistrement informatique des textes*. Paris : Klincksieck

Lapierre, A. 1972. *Fréquence et distribution des temps simples et des personnes verbales en français moderne*. Thèse de doctorat. Strasbourg : Centre de philologie romane.

Lebart, L. & A. Salem. 1988. *Analyse statistique de données textuelles*. Paris : Bordas.

Levin, B. 1993. *English verb classes and alternations*. London / Chicago : The University of Chicago Press.

Lewis, M. 1997. *Implementing the lexical approach. Putting theory into practice*. Hove : Language teaching publications.

Ligier, F. & L. Varagnolo. 1986. « Quel vocabulaire enseigner et comment l'enseigner. Entrevue avec Jacques Rebuffot », *dans* Ligier, F. & L. Savoie. 1986. *Didactique en questions. Le point de vue de 22 spécialistes en français langue seconde*. Beloeil : Les Éditions La Lignée. 144-154.

Lipka, L. 1990. *An outline of English lexicology. Lexical structure, word semantics and word-formation*. Tübingen : Niemeyer.

Lyne, Anthony A. 1985. *The vocabulary of french business correspondence. Word frequences, collocations, and problems of lexicométric method*. Genève / Paris : Slatkine / Champion.

Lyons, J. 1995. *Linguistic semantics. An introduction*. Cambridge : Cambridge University Press.

Mackey, W. *et al.* 1970. *Le vocabulaire disponible du français*. Paris : Didier.

MacNamara, J. 1982. *Names for things. A study of human learning*. Cambridge : M.I.T. Press.

Manczak, W. 1966. « Fréquence et évolution », *dans* Faculté des lettres et sciences humaines de Strasbourg. 1966. *Statistique et analyse linguistique*. Paris : PUF. 99-103.

Martel, P. 1984. « Concordances et divergences entre français fondamental et québécois fondamental », *Revue de l'association québécoise de linguistique* ». 3 (3). 39-61.

Martel, P. 1986. « Richesse lexicale et variable sociologique », *dans* CNRS. 1985. *En hommage à Muller. Méthodes quantitatives et informatiques dans l'étude des textes*. Genève : Slatkine / Champion. 559-608.

Martin, E. 1997. « Le traitement des corpus textuels à l'Institut National de la Langue Française », *Literary and linguistic computing*. Volume 12 (1). 37-45.

Martinet, A. 1979. *Grammaire fonctionnelle du français*. Paris : Didier.

Matoré, G. 1963. *Dictionnaire du vocabulaire essentiel (les 5 000 mots fondamentaux)*. Paris : Larousse.

Mazareno, J.R. & S.J. Mazareno 1988. *A cluster approach to elementary vocabulary instruction*. Newark : Delaware.

McCarthy, M. 1990. *Vocabulary*. Oxford : Oxford University Press.

McCullough, C. & C. Chacko. 1976. « La production de matériel pour l'enseignement de la lecture », dans Girolami-Boulinier & C. Mémin. 1976. *L'enseignement de la lecture. Problèmes et réflexions.* Paris / Neuchâtel : Delachaux et Niestlé, Éditeurs / Les presses de l'Unesco. 163-185.

Meeussen, A.E. 1959. *Essai de grammaire rundi.* Tervuren : Annales du Musée royal de l'Afrique centrale.

Meillet, A. & M. Cohen (éds). 1924. *Les langues du monde.* Paris : Champion.

Mel'čuk, I. 1993. *Cours de morphologie générale. Volume 1. Introduction et première partie : le mot.* Montréal : Presses de l'Université de Montréal.

Mel'čuk, I. 1994. *Cours de morphologie générale. Volume 2. Deuxième partie : significations morphologiques.* Montréal : Presses de l'Université de Montréal

Mel'čuk, I. 1996. *Cours de morphologie générale. Volume 3 : Troisième partie : moyens morphologiques . Quatrième partie : syntactiques morphologiques.* Montréal : Presses de l'Université de Montréal.

Mel'čuk, I. *et al.* 1984. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques I.* Montréal : Presses de l'université de Montréal.

Mel'čuk, I. *et al.* 1995. *Introduction à la lexicologie explicative et combinatoire.* Louvain : Éditions Duculot.

Mel'čuk, I. & É. Bakiza. 1987. « Les classes nominales en kirundi », *Bulletin de la Société linguistique de Paris.* LXXXII-1 : 283-341.

Ménard, N. 1978. « Richesse lexicale et mots rares », *Le français moderne.* XLVI. (1). 33-43.

Ménard, N. & L. Santerre. 1979. « La richesse lexicale individuelle comme marqueur sociolinguistique », *Cahiers de linguistique.* 1 : 165-188.

Ménard, N. 1983. *Mesure de la richesse lexicale. Théories et vérifications expérimentales. Études stylométriques et sociolinguistiques.* Genève / Paris : Slatkine / Champion.

Ménard, N. 1988. « Calcul de la cohésion lexico-sémantique des textes : aspects méthodologiques et recherches d'indices statistiques », dans Roper, J.P.G. 1988. *Computers in literary and linguistic research / L'ordinateur et les recherches littéraires et linguistiques.* Paris-Genève : Champion-Slatkine.

Ménard, N. 1989. « Mesure des relations lexico-sémantiques dans des textes scientifiques : problèmes méthodologiques », *Méta.* XXXIV (3). 468-478.

Merrilees, B. *et al.* 1992. « Editing and concording the Dictionarius of Firmin Le ver (1440) », *dans* Russon Wooldridge, T. 1992. *Historical dictionary databases*. Toronto: Centre for Computing in the Humanities. 8-19.

Montler, T. 1992. « Some computer applications for pacific northwest amerindian linguistics », *dans* Buchholzer, G.P. (ed.). 1992. *Amerindian Languages & Informatics*. Paris : Amerindia. 83-109.

Moreau, M.L. & M. Richelle. 1981. *L'acquisition du langage*. Bruxelles : Pierre Mardaga.

Muller, Ch. 1968. *Étude de statistique lexicale. Le vocabulaire du théâtre de Pierre Corneille*. Paris : Larousse.

Muller, Ch. 1977. *Principes et méthodes de la statistique lexicale*. Paris : Hachette.

Muller, Ch. 1979 a. « Code écrit et code parlé - Les personnes verbales », *dans* Muller, Ch. 1979. *Langue française et linguistique quantitative. Recueil d'articles*. 57-64. Genève : Slatkine.

Muller, Ch. 1979 b. « Passé simple et passé composé dans le vers classique », *dans* Muller, Ch. 1979. *Langue française et linguistique quantitative. Recueil d'articles*. 253-256. Genève : Slatkine.

Muller, Ch. 1979 c. *Étude de statistique lexicale. Le vocabulaire du théâtre de Pierre Corneille*. Paris : Larousse.

Muller, Ch. 1985. *Langue française, linguistique quantitative, informatique*. Genève / Paris : Slatkine / Champion.

Nation, I.S.P. 1992. *Teaching and learning vocabulary*. New York : Newbury House Publishers.

Nation, I.S.P. 1994. « Morphology in language learning and teaching », *dans* Asher, R.E. (ed.). 1994.*The encyclopedia of language and linguistics*. Volume 5. Oxford / New York / Seoul / Tokyo : Pergamon Press. 2584-2585.

Nattinger, J. R. & J.S. DeCarrico. 1992. *Lexical phrases and language teaching*. Oxford : Oxford University Press.

Ndayishinguje, P. 1978. *Contribution à la phonétique du kirundi*. Thèse de doctorat. Paris : Université de Paris III.

Nemni , M. 1986. « Les maux des mots », *dans* Ligier, F. & L. Savoie. 1986. *Didactique en questions. Le point de vue de 22 spécialistes en français langue seconde*. Beloeil : Les Éditions La Lignée. 155-180.

Nkanira, P. 1971. *Vues de Gustave Guillaume sur les langues à classes nominales et essai d'application à la catégorie du nom en burundais.* Mémoire de maîtrise. Québec : Université Laval.

Nkanira, P. 1984. *La représentation et l'expression du temps grammatical en kirundi. (Essai de description psychomécanique).* Thèse de doctorat. Québec : Université Laval.

Ntirampeba, P. 1988. *Le vocabulaire des contes kirundi. Essai de statistique lexicale.* Mémoire de licence. Bujumbura : Université du Burundi.

Ntirampeba, P. 1993. *Description des langues bantoues et lemmatisation. Le cas du kirundi.* Mémoire de maîtrise. Montréal : Université de Montréal.

OFIL. 1994. *Guide des produits et services d'ingénierie linguistique accessibles en France.* Paris : Observatoire français et international des industries de la langue.

Oflazer, K. 1994. « Two-level description of turkish morphology », *Literary and linguistic computing.* Volume 9 (2). 137-149.

Paquot, A. 1988. *Les Québecois et leurs mots. Étude sémiologique et sociolinguistique des régionalismes lexicaux au Québec.* Québec : Presses de l'Université Laval.

Parisi, D. & C. Castelfranchi. 1988. « Disambiguisation in a lexicaly based sentence understanding », *dans* Steven, I. Small *et al.* (ed). 1988. *Lexical ambiguity resolution. Prespectives from psycholinguistics, neurobiology and artificial intelligence.* San Mateo : Morgan Kaufmann Publisher Inc. 129-150.

Picoche, J. 1977. *Précis de lexicologie française. L'étude et l'enseignement du vocabulaire.* Paris : Nathan.

Piot, M. 1988. « Coordination-subordination. Une définition générale », *Langue française.* 77. 5-18.

Powell, F.C. 1982. *Statistical Tables for the social, biological and physical sciences.* Cambridge / New-York : Cambridge University Press.

Préfontaine, G. & R. Préfontaine. 1968. *Vocabulaire oral des enfants de 5 à 8 ans au Canada français.* Montréal : Beauchemin.

Primeau, G. & G. Labelle. 1981. « État et évolution du vocabulaire d'enfants québecois de neuf à douze ans », *dans* Gagné, G. & M. Pagé. 1981. *Études sur la langue parlée des enfants québecois.* Montréal : Les presses de l'université de Montréal. 133-145.

Quémada, B. 1962. « Actes du colloque international sur la mécanisation des recherches lexicologiques », *Cahiers de lexicologie*. 3.

Rey, A. *et al.* 1992. *Dictionnaire historique de la langue française*. Paris : Dictionnaires Le Robert.

Richards, J. *et al.* 1985. *Longman dictionary of applied linguistics*. Essex : Longman.

Richards, J.C. 1974. *Error analysis : perspectives on second language acquisition*. London : Longman.

Risland, H.D. 1945. *A basic vocabulary of elementary school children*. New York : The Macmillan Company.

Rivière, R. 1979. *Vocabulaire de base de la langue écrite. Répartition par thèmes, index alphabétique*. Riennes : Wesmael-Charlier.

Robert, A.H. (éditeur). 1965. *A statistical linguistic analysis of American English*. The Hague.

Robert, M.1960. *Le système des temps du français dans la langue moderne des journaux*. Strasbourg : Centre de philologie romane.

Robert, P. 1976. *Le Petit Robert*. Paris : Le Robert.

Rodegem, Firmin. M. 1967. *Précis de grammaire rundi*. Bruxelles / Gand : E.Story-scientia s.p.r.l.

Rodegem, Firmin M. 1970. *Dictionnaire Rundi-Français*. Tervuren : Annales du musée royal de l'Afrique centrale.

Rousseau, R. 1985. *Vocabulaire écrit des enfants et milieux d'appartenance. Une échelle multidimensionnelle et graduée du vocabulaire écrit des enfants de 7 à 12 ans*. Rimouski : Université du Québec à Rimouski.

Roy, G.R. 1976. *Contribution à l'analyse du syntagme verbal : étude morpho-syntaxique et statistique des coverbes*. Québec / Paris : Presses Universitaires de Laval / Klincksieck.

Sankoff, G. & H. Cedergren. 1971. « Some results of a sociolinguistic study of Montreal French », *dans* Regna, D. 1971. (ed.). *Linguistic diversity in canadian society*. Edmonton : Linguistic Research Inc. 61-87.

Sapir. E. 1921. *Language*. New York : Harcourt, Brace & World.

Scalise, S. 1984. *Generative morphology*. Dordrecht : Foris.

Sgarbas, K. *et al.* 1995. « A PC-KIMMO-Based morphological description of modern Greek », *Literary and linguistic computing*. Volume 10 (3). 189-203.

Spencer, A. 1991. *Morphological theory. An introduction to word structure in generative grammar.* London : Basil Blackwell.

Sproat, R. 1992. *Morphology and computation.* Cambridge / London : MIT Press.

Tashdjan, A. 1972. *Dictionnaire d'accès à l'information.* Abidjan : Institut de linguistique appliquée.

Ters, F. *et al.* 1964. *L'échelle Dubois-Buyse d'orthographe usuelle française.* Neuchatel : H. Messeiller.

Tesitelova, M. 1992. *Quantitative linguistics.* Amsterdam-Philadelphia : John Benjamins Publishing Company.

Tharp, J.B. *et al.* 1939. *A basic French vocabulary.* New York : Henry Holt.

Thompson, L.C., M.T.Thompson & R. Hsu. 1992. « A computerized dictionary of Thompson Salish », *dans* Buchholzer, G.P. (ed.). 1992, *Amerindian Languages & Informatics.* Paris : Amerindia. 3-30.

Thorndike, E.L. 1921. *The teacher's word book.* New York.

Thorndike, E.L. & I. Lorge 1944. *The teacher's word book of 30 000 words.* New York : Teachers college / Columbia University.

Touratier, C. 1983. « Définition du verbe (à propos de l'indonésien et du malgache) », *dans* Cercle linguistique d'Aix-en-Provence. 1983. *Les parties du discours.* Aix-en-Provence : Service des publications de l'Université de Provence.

Tournier, J. 1985. *Introduction descriptive à la lexicogénétique de l'anglais contemporain.* Paris-Genève : Champion-Slatkine.

Trudel, R. & R. Antonius. 1991. *Méthodes quantitatives appliquées aux sciences humaines.* Montréal : Centre Éducatif et Culturel Inc.

Tubach, J-P. 1995. « Reconnaissance de la parole : questions et recherches », *dans* Huot, H. & H. Portine. 1995. *La linguistique appliquée aujourd'hui : problèmes et méthodes.* Amsterdam : De Werelt. 7-18.

Tubach, J-P. 19989. *La parole et son traitement automatique.* Paris : Masson.

Tzoukermann, E. & M. Liberman. 1990. « A finite-state morphological processor for Spanish », *Computational linguistics.* 3. 277-281.

Vachek, J. 1989. « Thougts on some fifty years of research in written language », *dans* Vachek, J. 1989. *Written language revisited.* 197-214. Amsterdam / Philadelphia : John Benjamins Publishing Company. 197-214.

Vander Beke, G.E. 1929. *French word book.* New York : MacMillan.

Vaneste, A. 1988. « Pour une description statistique de la structure morphologique du lexique », *Travaux de linguistique*. 16 : 123-144.

Vikis-Freibergs, V. 1974. *Fréquence d'usage des mots au Québec*. Montréal : Les Presses de l'Université de Montréal.

Vinette, R. 1943. *Recherche sur le vocabulaire des enfants : essai de détermination expérimentale du vocabulaire employé par les enfants des écoles primaires à Montréal*. Thèse de doctorat. Montréal : Institut pédagogique Saint-Georges.

Wallace, J.M. 1982. *Teaching vocabulary*. London : Heinmann Educational books.

Ward, C.F. 1926. *Minimum French vocabulary test book*. New York.

West, M. 1953. *A general service list of English words with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. London / Harlow : Longmans-Green and co ltd.

Wijnands, P.H. 1989. « Systèmes-experts et terminologie », *Méta*. XXXIV. (3). 502-508.

Yule, G.U. 1944. *The statistical study of literary vocabulary*. Cambridge : Archon Books.

ANNEXES

ANNEXE 1 :
Liste des numéros, pages et colonnes saisis par sous-corpus

Sous-corpus 1 : w1

1. Nd-1975-10-p8-c 1 & 3[1].
2. Nd-1976-4-p7-c 1.
3. Nd-1976-6-p2-c 2 & 1.
4. Nd-1976-10-p6-c 1 & 3.
5. Nd-1976-18-p6-c 1 & 2.
6. Nd-1976-23-p6-c 3 & 4.
7. Nd-1977-15-p4-c 1 & 4.
8. Nd-1977-23-p5-c 1 & 2.
9. Nd-1978-11-p6-c 1.
10. Nd-1979-20-p4-c 4 & 2.

Sous-corpus 2 : w2

1. Nd-1990-11-p2-c 1.
2. Nd-1990-17-p7-c 4.
3. Nd-1991-33-p4-c 1 & 2 & 3 & 4.
4. Nd-1991-32-p1-c 2.
5. Nd-1991-52-p3-c 2 & 3.
6. Nd-1992-64-p2-c 3 & 2 & 1.
7. Nd-1992-70-p1-c 1 & 2 & 3.
8. Nd-1992-76-p6-c 2.
9. Nd-1993-97-p6-c 4.
10. Nd-1993-98-p6-c 3.

---

[1] Il faut lire comme ceci : « l'extrait est tiré de Ndongozi (Nd.) de 1975, n°10, à la page 8, aux colonnes 1 & 3 ». Les colonnes se suivent dans l'ordre de tirage.

11. Nd-1994-103-p2-c 1.
12. Nd-1994-104-p13-c 4 & 2 & 1.
13. Nd-1994-105-p7-c 1 & 2 & 3.
14. Nd-1994-114-p8-c 4 & 2 & 3.
15. Nd-1994-118-p5-c 1 & 2.
16. Nd-1991-31-p4-c 3.
17. Nd-1991-52-p16-c 2.
18. Nd-1991-50-p1-c 1 & 2 &3.
19. Nd-1992-72-p6-c 1 & 2 & 3 & 4.
20. Nd-1993-97a-p1-c 2.

Sous-corpus 3 : w3

1. Nd-1974-12-p1-c 1 & 2 & 3 & 4.
2. Nd-1974-16-p1-c 1 & 2 & 3, p2-c 4.
3. Nd-1974-17-p4-c 1 & 2.
4. Nd-1975-1-p3-c 1 & 2 & 3 & 4.
5. Nd-1976-1-p5-c 1 & 2.
6. Nd-1976-10-p2-c 1 & 2.
7. Nd-1976-n2-p1-c 1 & 2.
8. Nd-1976-21-p1-c 1 & 2 & 3.
9. Nd-1976-24-p5-c 1 & 2 & 3.
10. Nd-1976-7-p7-c 1 & 2.
11. Nd-1977-20-p1-c 1 & 2 & 3.
12. Nd-1977-23-p7-c 2 & 3.
13. Nd-1977-24-p2-c 1 & 2.
14. Nd-1977-3-p1-c 2 & 3 & 4.
15. Nd-1977-6-p1-c 2 & 3 & 4.
16. Nd-1978-13-p3-c 3 & 4.
17. Nd-1978-17-p6-c 1 & 2 & 3.
18. Nd-1978-3-p3-c 1 & 2 & 3.
19. Nd-1979-15-p3-c 2 & 3 & 4, p4-c 1.
20. Nd-1979-23-p7-c 1 & 2 & 3.

Sous-corpus 4 : w4

1. Nd-1990-11-p1-c 1 & 2 & 3.
2. Nd-1990-12-p1-c 1 & 2, p2-c 1.
3. Nd-1990-16-p1-c 2, p2-c 1.
4. Nd-1990-19-p7-c 1 & 2.
5. Nd-1990-22-p6-c 1 & 2.
6. Nd-1990-26-p1-c 1 & 2 & 3 & 4.
7. Nd-1990-29-p2-c.1 & 2.
8. Nd-1991-31-p5-c1 & 2
9. Nd-1990-13-p1-c 1.
10. Nd-1991-35-p1-c 1 & 2 & 3, p6-c 1.
11. Nd-1991-36-p6-c 1 & 2.
12. Nd-1991-38-p6-c 1 & 2 & 3 & 4.
13. Nd-1991-43-p5-c 1 & 2 & 3.
14. Nd-1991-44-p4-c 1 & 2.
15. Nd-1991-48-p4-c 1 & 2 & 3.
16. Nd-1991-53-p1-c 2 & 3, p3-c 1.
17. Nd-1992-54-p5-c 1 & 2.
18. Nd-1992-56-p5-c 1 & 2 & 3.
19. Nd-1992-59-p5-c 1 & 2 & 3 & 4.
20. Nd-1992-65-p5-c 1 & 2.
21. Nd-1992-67-p5-c 1 & 2 & 3.
22. Nd-1992-68-p5-c 1 & 2 & 3.
23. Nd-1992-70-p5-c 1 & 2.
24. Nd-1992-71-p5-c 1 & 2.
25. Nd-1992-72-p5-c 1 & 2.
26. Nd-1992-75-p5-c 1 & 2.
27. Nd-1992-76-p5-c 1 & 2 & 3.
28. Nd-1992-77-p2-c 1 & 2 & 3 & 4.
29. Nd-1993-78-p5-c 1 & 2.
30. Nd-1993-86-p8-c 1 & 2.

31. Nd-1993-88-p1-c 1 & 2 & 3-p2-c1.
32. Nd-1993-90-p10-c 1 & 2 & 3.
33. Nd-1993-93-p6-c 1 & 2 & 3.
34. Nd-1993-95-p6-c 1 & 2 & 3.
35. Nd-1993-98-p5-c 1 & 2 & 3.
36. Nd-1993-99-p5-c 1 & 2 & 3 & 4.
37. Nd-1994-100-p6-c 2 & 3 & 4.
38. Nd-1994-101-p6-c 1 & 2 & 3.
39. Nd-1994-107-p5-c 1 & 2 & 3.
40. Nd-1994-108-p6-c 1 & 2 & 3.
41. Nd-1994-113-p6-c 1 & 2 & 3.
42. Nd-1994-115-p5-c 1 & 2 & 3.
43. Nd-1994-118-p6-c 1 & 2 & 3 & 4.
44. Nd-1994-119-p15-c 1 & 2 & 3 & 4.
45. Nd-1994-120-p5-c 1 & 2 & 3.

Sous-corpus 5 : w5

1. Nd-1974-22-p4-c 1 & 2 & 3 & 4.
2. Nd-1976-10-p1-c 1 & 2 & 3.
3. Nd-1977-15-p6-c 1 & 2.
4. Nd-1977-20-p4-c 1 & 2 & 3.
5. Nd-1977-4-p4-c 2 & 3.
6. Nd-1978-10-p8-c 1 & 2.
7. Nd-1978-18-p8-c 1 & 2 & 3..
8. Nd-1979-18-p7-c 1 & 2.
9. Nd-1979-20-p5-c 3 & 4..
10. Nd-1979-22-p6-c 1 & 2.

Sous-corpus 6 : w6

1. Nd-1990-7-p5-c 1 & 2 & 3.
2. Nd-1990-14-c 6 & 1 & 2.
3. Nd-1990-19-c 4 & 1 & 2 & 3.
4. Nd-1990-25-p7-c 1 & 2.

5. Nd-1990-27-p6-c 1 & 2.
6. Nd-1990-8-p2-c 1 & 2 & 3.
7. Nd-1991-38-p1-c 1 & 2 & 3 & 4.
8. Nd-1991-41-p3-c 1 & 2 & 3.
9. Nd-1991-44-p6-c 3 & 4.
10. Nd-1991-49-p5c 1 & 2 & 3.
11. nd-1991-51-p1-c 2 & 3, p5-c 1 & 2.
12. Nd-1991-52-p8-c 2 & 3 & 4.
13. Nd-1991-53-p5-c 1 & 2 & 3.
14. Nd-1992-56-p7-c 1 & 2 & 3.
15. Nd-1992-65-p3-c 1 & 2 & 3.
16. Nd-1992-67-p8-c 1 & 2 & 3.
17. Nd-1992-70-p15-c 1 & 2 & 3.
18. Nd-1992-75-p13-c 1 & 2 & 3.
19. Nd-1993-82-p5-c 1 & 2 & 3.
20. Nd-1993-84-p8-c 1 & 2 & 3 & 4.
21. Nd-1993-85-p15-c 1 & 2 & 3.
22. Nd-1993-87-p8-c 1 & 2 & 3.
23. Nd-1993-91-p11-c 1 & 2 & 3.
24. Nd-1993-93-p8-c 1 & 2 & 3.
25. Nd-1993-95-p11-c 1 & 2 & 3.
26. Nd-1993-96-p14-c 1 & 2 & 3 & 4.
27. Nd-1993-99-p8-c 1 & 2 & 3 & 4.
28. Nd-1994-100-p7-c 1 & 2 & 3.
29. Nd-1994-108-p14-c 1 & 2 & 3.
30. Nd-1994-112-p7-c 1 & 2 & 3.

Sous-corpus 7 : w7

1. Nd-1975-13-p1-c 1 & 2 & 3 & 4.
2. Nd-1975-16-p3-c 1 & 2 & 3 & 4.
3. Nd-1975-4-p1-c 1 & 2 & 3 & 4.
4. Nd-1975-7-p1-c 1 & 2 & 3 & 4.
5. Nd-1976-1-p1-c 1 & 2.
6. Nd-1976-13-p3-c 2 & 3.

7. Nd-1976-20-p1-c 1 & 2.
8. Nd-1976-no21-p4-c 1 & 2.
9. Nd-1976-22-p2-c 3 & 4.
10. Nd-1977-no1-p5-c 1 & 2.
11. Nd-1977-16-p1-c 2 & 3 & 4.
12. Nd-1977-21-p2-c 2 & 3.
13. Nd-1977-5-p4-c 1 & 2.
14. Nd-1977-6-p3-c 1 & 2.
15. Nd-1978-12-p5-c 1 & 2 & 3.
16. Nd-1978-18-p5-c 1 & 2.
17. Nd-1978-21-p2-c 3 & 4.
18. Nd-1979-17-p1-c 3 & 4.
19. Nd-1979-20-p1-c 1 & 2 & 3.
20. Nd-1979-21-p1-c 1 & 2.

Sous-corpus 8 : w8

1. Nd-1990-6-p1-c 1 & 2 & 3 & 4 & 5.
2. Nd-1990-16-p8-c 1 & 2 & 3 & 4.
3. Nd-1990-26-p3-c 1 & 2 & 3 & 4.
4. Nd-1991-33-p1-c 2 & 3 & 4, p3-c1 & 2.
5. Nd-1991-34-p1-c 1 & 2 & 3 & 4.
6. Nd-1991-35-p3-c 1 & 2 & 3.
7. Nd-1991-36-p2-c 3 & 4-p3-c1.
8. Nd-1991-38-p1-c 1 & 2 & 3, p3-c2 & 3 & 4.
9. Nd-1991-39-p1-c 1 & 2 & 3 & 4.
10. Nd-1991-42-p1-c 1 & 2 & 3 & 4-p3-c3.
11. Nd-1991-47-p1-c 1 & 2 & 3, p2-c2 & 3.
12. Nd-1992-57-p3-c 1 & 2 & 3.
13. Nd-1992-59-p4-c 1 & 2 & 3.
14. Nd-1992-62-p3-c 1 & 2 & 3.
15. Nd-1992-63-p2-c 1 & 2 & 3.
16. Nd-1992-64-p3-c 1 & 2 & 3.
17. Nd-1992-66-p3-c 1 & 2.
18. Nd-1992-67-p2-c 1 & 2 & 3 & 4.

19. Nd-1992-68-p2-c 1 & 2 & 3 & 4.
20. Nd-1992-69-p4-c 1 & 2 & 3 & 4.
21. Nd-1992-72-p2-c 1 & 2 & 3.
22. Nd-1992-73-p3-c 1 & 2 & 3.
23. Nd-1992-74-p2-c 1 & 2 & 3.
24. Nd-1992-76-p2-c 1 & 2 & 3 & 4.
25. Nd-1993-81-p2-c 1 & 2 & 3 & 4.
26. Nd-1993-85-p2-c 1 & 2 & 3.
27. Nd-1993-86-p6-c 1 & 2 & 3.
28. Nd-1993-90-p2-c 1 & 2 & 3 & 4.
29. Nd-1993-91-p3-c 1 & 2 & 3 & 4.
30. Nd-1993-93-p2-c 1 & 2 & 3.
31. Nd-1993-96-p2-c 1 & 2 & 3 & 4.
32. Nd-1993-99-p2-c 1 & 2 & 3.
33. Nd-1993-100-p2-c 1 & 2 & 3 & 4.
34. Nd-1993-101-p2-c 1 & 2 & 3.
35. Nd-1994-103-p2-c 1 & 2 & 3.
36. Nd-1994-104-p2-c 1 & 2 & 3.
37. Nd-1994-106-p2-c 1 & 2 & 3.
38. Nd-1994-107-p2-c 1 & 2 & 3.
39. Nd-1994-109-p2-c 1 & 2 & 3.
40. Nd-1994-111-p2-c 1 & 2 & 3.
41. Nd-1994-113-p1-c 1.
42. Nd-1994-115-p2-c 1 & 2 & 3.
43. Nd-1994-116-p2-c 1 & 2 & 3 & 4.
44. Nd-1994-118-p2-c 1 & 2 & 3.
45. Nd-1994-120-p4-c 1 & 2 & 3 & 4.

Sous-corpus 9 : w9

1. U1976-104-p8-c 1.
2. U1975-105-p6-c 2.
3. U1978-149-p8-c 1 & 3.
4. U1978-150-p8-c 1 & 3.
5. U1978-156-p1-c 3 & 2.

6. U1978-165-p8-c 1, p6-c2 & 3.
7. U1980-217-p3-c 2 & 3.
8. U1980-242-p8-c 1, p3-c 2.
9. U1980-249-p6-c 1.
10. U1980-260-p8-c 1 & 2.

Sous-corpus 10 : w10

1. U1990-668-p7-c 2.
2. U1990-662-p2-c 3.
3. U1992-746-p8-c 2.
4. U1990-672-p1-c 3.
5. U1990-674-p6-c 2.
6. U1990-690-p1-c 1.
7. U1991-683-p1-c 1.
8. U1991-685-p1-c 1.
9. U1991-703-p3-c 1.
10. U1991-723-p8-c 1.
11. U1991-729-p1-c 1.
12. U1991-739-p7-c 1.
13. U1992-746-p8-c 2.
14. U1991-776-p5-c 2 & 3 & 1.
15. U1992-782-p4-c 1 & 2.
16. U1993-802-p2-c 4 & 1.
17. U1993-826-p2-c 3 & 1.
18. U1994-844-p10-c 1 & 4.
19. U1994-865-p4-c 4 & 2.
20. U1994-874-p7-c 1 & 2.

Sous-corpus 11 : w11

1. U1974-71-p1-c 1 & 2.
2. U1975-101-p1-c 2 & 3.
3. U1975-99-p1-c 1.
4. U1978-147-p1-c 1 & 2.

5. U1978-149-p1-c 1 & 2.
6. U1978-157-p1-c 1 & 2 & 3 & 4 & 5.
7. U1980-104-p1-c 1.
8. U1980-215-p2-c 1 & 2.
9. U1980-217-p16-c 1 & 2.
10. U1980-218-p1-c 1 & 2 & 3.
11. U1980-225-p1-c 1 & 2 & 3.
12. U1980-228-p1-c 1 & 2 & 3.
13. U1980-232-p1-c 1 & 2 & 3 & 4 & 5.
14. U1980-238-p1-c 1 & 2.
15. U1980-247-p1-c 1 & 2 & 3 & 4.
16. U1980-249-p1-c 1 & 2.
17. U1980-252-p1-c 1 & 2.
18. U1980-257-p1-c 1 & 2 & 3.
19. U1980-262-p2-c 1 & 2.
20. U1980-155-p1-c 1 & 2.

Sous-corpus 12 : w12

1. U1990-632-p1-c 1 & 2 & 3.
2. U1990-638-p1-c1 & 2 & 3 & 4, p2-c 1.
3. U1990-644-p1-c 2 & 3 & 4-et-p04-c 1.
4. U1990-646-p1-c 1 & 2 & 3.
5. U1990-647-p2-c 1 & 2.
6. U1990-649-p2-c 1 & 2.
7. U1990-654-p1-c 1 & 2 & 3.
8. U1990-658-p1-c 4, p3-c 1 & 2.
9. U1990-670-p1-c 1.
10. U1990-672-p2-c 1 & 2.
11. U1990-677-p1-c 1.
12. U1991-684-p1-c 1.
13. U1991-686-p1-c 2.
14. U1991-689-p1-c 1.
15. U1991-691-p1-c 1.
16. U1991-692-p1-c 1.

17. U1991-701-p1-c 2.
18. U1991-703-p1-c 1.
19. U1991-706-p1-c 1.
20. U1991-713-p1-c 1.
21. U1991-709-p1-c 1.
22. U1991-712-p1-c 2.
23. U1991-713-p1-c 1.
24. U1991-715-p1-c 1.
25. U1991-721-p1-c 1.
26. U1992-735-p1-c 1.
27. U1992-744-p1-c 1.
28. U1992-746-p3-c 1.
29. U1992-747-p1-c 1.
30. U1992-759-p1-c 1.
31. U1992-762-p1-c 1.
32. U1992-770-p1-c 1.
33. U1992-771-p1-c 1.
34. U1992-773-p1-c 2.
35. U1992-775-p1-c 1 & 2 & 3.
36. U1992-779-p1-c 1.
37. U1992-780-p1-c 1.
38. U1992-782-p1-c 1 & 2 & 3.
39. U1992-783-p1-c 1 & 2 & 3.
40. U1993-791-p2-c 1.
41. U1993-792-p2-c 1 & 2.
42. U1993-793-p1-c 1.
43. U1993-798-p1-c 3.
44. U1993-802-p1-c 1 & 2.
45. U1993-805-p1-c 2.
46. U1993-807-p1-c 2.
47. U1993-808-p1-c 2.
48. U1993-811-p1-c 1 & 2.
49. U1993-813-p6-c 1.
50. U1993-817-p1-c 1 & 2.
51. U1993-818-p1-c 2 & 3.

52. U1993-822-p1-c 1 & 2.
53. U1993-824-p1-c 1 & 2.
54. U1993-827-p1-c 4.
55. U1993-833-p1-c 3.
56. U1993-842-p1-c 2 & 3.
57. U1994-843-p1-c 2 & 3.
58. U1993-844-p1-c 2 & 3.
59. U1994-864-p2-c 2 & 3.
60. U1994-975-p1-c 2 & 3.

Sous-corpus 13 : w13

1. U1974-75-p1-c 1 & 2.
2. U1974-80-p1-c 1 & 2 & 3.
3. U1974-87-p3-c 1.
4. U1974-90-p3-c 1 & 2.
5. U1975-93-p4-c 1 & 2 & 3.
6. U1975-99-p3-c 1.
7. U1978-146-p5-c 1 & 2.
8. U1978-147-p6-c 2 & 3.
9. U1978-154-p4-c 2 & 3 & 4 & 5.
10. U1978-155-p2-c 1 & 2 & 3.
11. U1978-159-p4-c 1 & 2.
12. U1980-230-p7-c 1 & 2 & 3.
13. U1980-243-p4-c 1 & 2.
14. U1980-248-p6-c 1 & 2 & 3 & 4 & 5.
15. U1980-258-p5-c 1 & 2.

Sous-corpus 14 : w14

1. U1990-635-p4-c 3.
2. U1990-656-p6-c 1 & 2 & 3 & 4.
3. U1990-657-p1-c 1 & 2.
4. U1990-666-p12-c 1.
5. U1990-675-p5-c 1.

6. U1990-676-p1 & 2.
7. U1990-688-p7-c 1.
8. U1991-691-p1-c 1, p2-c 1.
9. U1991-692-p7-c 1.
10. U1991-700-p5-c 1.
11. U1991-702-p5-c 1.
12. U1991-703-p5-c 1.
13. U1991-710-p5-c 1.
14. U1991-723-p6-c 1.
15. U1992-760-p15-c 1.
16. U1992-778-p7-c 1.
17. U1992-781-p6-c 1.
18. U1992-792-p7-c 1.
19. U1993-805-p7-c 1.
20. U1993-807-p7-c 1.
21. U1993-815-p1 & 7.
22. U1993-817-p1 & 4.
23. U1993-820-p4-c 1.
24. U1993-828-p1-c 2 & 3.
25. U1993-831-p1-c 1, p2-c 1 & 2.
26. U1994-840-p1 & 2.
27. U1994-858-p12-c 1.
28. U1993-861-p1-c 2 & 3.
29. U1994-867-p9-c 1.
30. U1994-871-p7-c 1.

Sous-corpus 15 : w15

1. U1974-74-p6-c 1 & 2 & 3.
2. U1974-80-p4-c 1 & 2.
3. U1975-99-p5-c 1.
4. U1976-105-p3-c 1 & 2 & 3 & 4.
5. U1977-118-p4-c 1 & 2.
6. U1978-144-p6-c 3 & 4.
7. U1978-149-p6-c 1 & 2.

8. U1978-154-p6-c 1.
9. U1978-162-p3-c 1 & 2.
10. U1978-166-p9-c 1 & 2.
11. U1980-229-p5-c 1.
12. U1980-231-p3-c 1 & 2.
13. U1980-238-p4-c 1 & 2.
14. U1980-239-p4-c 1 & 2 & 3 & 4.
15. U1980-258-p12-c 1 & 2.

Sous-corpus 16 : w16

1. U1990-638-p8-c 1.
2. U1990-653-p8-c 1.
3. U1990-662-p8-c 1.
4. U1990-669-p8-c 1.
5. U1990-670-p8-c 1.
6. U1990-674-p8-c 1.
7. U1991-684-p8-c 1.
8. U1991-694-p8-c 1.
9. U1991-699-p8-c 1 & 2.
10. U1991-702-p8-c 1 & 2.
11. U1991-713-p8-c 1.
12. U1991-730-p8-c 1.
13. U1991-733-p8-c 1.
14. U1992-749-p8-c 1.
15. U1992-756-p8-c 1.
16. U1992-766-p8-c 1.
17. U1992-772-p8-c 1.
18. U1992-784-p8-c 1.
19. U1993-789-p8-c 1.
20. U1993-790-p8-c 1.
21. U1993-795-p8-c 1.
22. U1993-799-p8-c 2 & 3.
23. U1993-800-p8-c 1.
24. U1993-807-p8-c 1.

25. U1993-815-p8-c 1.
26. U1993-832-p8-c 1.
27. U1994-843-p8-c 1.
28. U1994-847-p8-c 1.
29. U1994-856-p12-c 1.
30. U1994-858-p12-c 1.

ANNEXE 2

Exemple de résultats fournis par l'analyseur morphologique du kirundi

Statistiques GlobalesW1=======

Nombre Total de Verbes:	792
Nombre Total de Substantifs:	731
Nombre Total de Substantifs (à base verbo-nominale) :	205
Nombre Total de Substantifs (à base nominale) :	526
Nombre Total d'Adjectifs:	48

Fréquence des Radicaux dans les Verbes

ak 1
akir 1
am 8
ambur 1
ba 23
babar 2
bandany 1
bandw 1
candag 1 (...)

Fréquence des Radicaux dans les Substantifs (à base verbo-nominale)
ank 1
aruk 3
ba 4
ban 1
baz 2
biik 2
gend 2 (...)

Fréquence des Radicaux dans les Substantifs (à base nominale)

ago 1
aka 6
akatsi 1
ana 10
anya 1
atsi 1
ayi 1
boko 2 (...)

Fréquence des Radicaux Adjectivaux

bi 1
inshi 21
iza 2
ke 6
kuru 10
nini 1
shasha 2
to 2

Fréquence des Aspects Verbaux

a 441 e 81 ye 177 ø 93

Fréquence des Préfixes de Classe

mu 101
ba 98
mi 35
ri 99
ma 58
ki 50
bi 67
ru 15
bu 39
n 58
aka 18
tu 3
ku 37
ha 4
autre 49

Fréquence des dérivatifs thématiques nominaux

a 99 e 15 i 0 o 46 u 0 yi 30 sans 15

Fréquence des Suffixes dans les Verbes

an 24
ir 93
u 45
ish 17
ur 1
i 45
ik 5
uk 2
agir 2
sans 558

Fréquence des Suffixes dans les Substantifs

an 10
ir 14
u 22
ish 3
i 9
sans 147

ANNEXE 3

Occurrences de quelques radicaux adjectivaux dans les 16 sous-corpus

Radical	*Glose*	*w1*	*w2*	*w3*	*w4*	*w5*	*w6*	*w7*	*w8*	*w9*	*w10*	*w11*	*w12*	*w13*	*w14*	*w15*	*w16*	F_o
[-BÍ]	'mauvais'	1	0	2	2	0	2	5	4	0	3	0	4	4	7	4	2	**30**
[-BÍSI]	'cru'	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	**1**
[-HIÍRE]	'respectable'	0	0	0	0	0	0	1	3	0	1	1	6	0	1	0	0	**13**
[-IÍNSHI]	'en grande quantité' / 'nombreux'	21	11	38	70	12	31	11	30	5	20	14	47	32	16	24	26	**408**
[-IIZÁ]	'beau'	2	5	7	10	3	6	5	15	3	3	4	6	8	7	5	8	**98**
[-KÉ]	'en petite quantité'	6	6	2	16	3	4	6	19	4	2	2	14	16	15	6	4	**124**
[-KÉEYA]	'en petite quantité'	0	0	0	2	0	0	0	0	2	0	0	0	0	0	0	1	**4**
[-KÉEYI]	'en petite quantité'	0	1	0	0	1	0	0	1	1	1		1	1	1	2	0	**11**
[-KÚRU]	'grand'	12	7	1	22	2	12	9	20	4	3	16	38	5	5	2	7	**165**
[-NÍNI]	'grand'	1	2	1	6	1	3	1	6	3	1	1	2	1	1	4	1	**35**
[-SA]	'seul'	0	1	0	4	2	0	1	3	0	1	0	1	2	2	2	6	**25**
[-SHÁ]	'nouveau'	0	0	0	1	0	0	0	2	0	2	1	5	0	0	0	1	**12**
[-SHAÁSHA]	'nouveau'	2	0	2	6	1	1	1	6	2	4	5	22	2	2	4	0	**60**
[-TAGATIFÚ]	'saint'	0	0	0	0	0	0	3	1	0	0	0	0	0	0	0	0	**4**
[-TÓ]	'petit'	2	1	2	3	1	5	4	1	0	1	2	4	1	0	2	4	**33**
[-TOÓTO]	'vert'	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	**3**
[-TÓOYI]	'petit'	0	0	0	2	2	4	0	2	00	0	0	4	2	0	2	4	**22**

ANNEXE 4

Typologie des thèmes

Thème	*L'individu*	*Les relations*	*La société*	*Le développement*	*La culture*
S	- personnes et lieux	- salutations	- agriculture et élevage	- école et règlements	- musique
o	- parties du corps	- famille	- commerce	- hygiène	- religion
u	- couleurs	- voisinage	- le marché	- transport	- mort
s	- habits	- école	- ville *versus* campagne	- unités de mesure	- arts
-	- santé	- travaux à la maison	- tradition *versus* modernité	et nombres	plastiques
t	- vie et mort	- relations parents-enfants	- illétrisme	- commerce	- sport
h	- nourriture et boissons	- relations garçons et filles	- gouvernement	- transport	- médias
è	- étapes de la vie	- mariage	- démocratie, élections	- technologie	- théâtre
m	- plans de carrière		- partis politiques		- jeux
e			- alcoolisme		
s			- violence et vol		
			- patriotisme		
			- immigration		

ANNEXE 5 Écart type, coefficient de variation et fréquences théoriques de quelques vocables à indice de dispersion négatif

Vocable	F_o	σ	v	*D*	*U*	*Glose*
IGAÁRIYAMOÓSHI	1	0,25	3,96	-0,02	-0,02	'train'
IMBÚRABÚGIRÍRE	2	0,50	3,97	-0,03	-0,05	'personne difficile à satisfaire'
UMUKÍRO	5	1,27	4,06	-0,05	-0,25	'membre du mouvement Chiro'
URWÁARA	4	1,08	4,30	-0,11	-0,44	'ongle'
UMUGÍRANÉEZÁ	2	0,65	5,23	-0,35	-0,70	'membre d'un organisme humanitaire'

L'écart type et le coefficient de variation sont calculés à partir des écarts entre les fréquences théoriques des vocables dans les 16 sous-corpus (de w1 à w16) et leurs fréquences réelles. Voici ces fréquences théoriques pour les vocables ci-dessus :

Vocable	*w1t*	*w2t*	*w3t*	*w4t*	*w5t*	*w6t*	*w7t*	*w8t*	*w9t*	*w10t*	*w11t*	*w12t*	*w13t*	*w14t*	*w15t*	*w16t*
IGAÁRIYAMOÓSHI	0,03	0,05	0,05	0,12	0,03	0,07	0,05	0,11	0,02	0,05	0,05	0,15	0,04	0,08	0,04	0,08
IMBÚRABÚGIRÍRE	0,06	0,10	0,10	0,23	0,05	0,15	0,10	0,23	0,05	0,10	0,10	0,30	0,07	0,15	0,08	0,15
UMUKÍRO	0,14	0,25	0,24	0,58	0,13	0,37	0,24	0,56	0,12	0,24	0,24	0,76	0,18	0,38	0,19	0,38
URWÁARA	0,11	0,20	0,19	0,46	0,10	0,29	0,19	0,45	0,10	0,19	0,19	0,60	0,15	0,30	0,15	0,30
UMUGÍRANÉEZÁ	0,06	0,10	0,10	0,23	0,05	0,15	0,10	0,23	0,05	0,10	0,10	0,30	0,07	0,15	0,08	0,15