LINGUISTIQUE DE CORPUS

3.1. Approche en linguistique de corpus

> « … *la langue a l'air assez différente quand on examine un grand morceau d'un coup* »
>
> [une remarque de Sinclair citée dans l'ouverture de l'ouvrage de Habert et al. (1997)]

Il va sans dire que depuis longtemps, la plupart des linguistes, surtout ceux qui exercent la linguistique de terrain, s'appuient sur des « corpus » de diverses formes dans leurs travaux. Or le terme « corpus » est employé dans notre étude avec un sens spécifique défini dans le cadre de la linguistique de corpus dont nous nous inspirons. Comme pour toute terminologie lorsqu'il s'agit de domaines assez « nouveaux » tels que la linguistique de corpus qui ne s'est développée qu'à partir des années quatre-vingts (Williams, 2005)[i], les avis divergent concernant la définition du terme « corpus », selon l'approche que l'on adopte ou le but/genre des analyses. Il nous semble donc utile et important de considérer au premier abord ce qu'implique la linguistique de corpus en général, et ensuite pour les analyses des fonctions discursives du birman parlé.

Plusieurs linguistes reconnaissent que le terme même « linguistique de corpus » qui est en fait la traduction littérale de *corpus linguistics* en anglais, issue d'une tradition anglo-saxonne, reste une notion plutôt floue en français [cf. Williams, 2006 ; UFR EILA, http://wall.jussieu.fr/recherche.shtml, etc.]. Williams (2006), dans son article intitulé « La linguistique et le corpus : une affaire prépositionnelle » explique son

[i] Selon Teubert (2009), la linguistique de corpus est entrée en scène dans les années soixante du siècle dernier. Le propos de Williams correspond à la remarque de Kerbrat-Orecchioni (1998) qui, au sujet des travaux sur la communication interpersonnelle, affirme qu'afin d'appréhender l'objet-langue et ses réalisations en milieu naturel, il est nécessaire d'analyser de très près, sur la base d'enregistrements de données 'authentiques', le fonctionnement d'échanges langagiers effectivement attestés. Pourtant il a fallu attendre en France les années 80 pour voir certains linguistes recourir systématiquement à cette pratique descriptive, qui reste encore du reste minoritaire (p. 51-52).

point de vue comme suit : dans le terme original en anglais *corpus linguistics*, le mot « linguistique » représente la discipline et le mot « corpus » décrit son objet, laissant le reste sujet à l'interprétation. Il soutient ainsi que « la puissance de l'anglais est dans l'ambiguïté » mais qu'en français la situation est plus complexe, étant donné que l'on ne peut pas juxtaposer les deux mots sans préposition. Or le choix de préposition entre 'de', 'des', ou 'sur' n'est pas chose simple, car chaque choix implique une interprétation différente : *de* signifiera la présence d'une discipline unique ; *des* suggère que plusieurs disciplines (et non plusieurs approches de la même discipline) sont en jeu ; *sur* implique que d'autres domaines de la linguistique peuvent utiliser les corpus sans faire de la linguistique de corpus *per se,* ce qui soulève la question de la nature des corpus (p.151)[i].

Mellet (2002 : 2) partage l'avis de Williams en soulignant que la notion de corpus « s'est complexifiée au cours des dernières années en fonction de la diversité des pratiques et des objectifs assignés à la constitution et à l'exploitation des corpus ». Rastier (2005 :31), pour sa part reconnaît que « la linguistique de corpus ne constitue aucunement un domaine de recherche unifié ». En effet, le titre même d'un des ouvrages clés sur le sujet pour les francophones,[ii] « *Les Linguistiques de corpus* » de Habert et al. (1997) souligne l'hétérogénéité du domaine de recherche. Ces derniers utilisent également le terme « la linguistique à base de corpus », dont la condition décisive est l'accès à de vastes ensembles des données linguistiques sous forme électronique. Ce débat pourrait continuer (sans doute à l'infini) si l'objet de notre étude portait sur cette nouvelle discipline, mais ce n'est pas le cas ici[iii]. En ce qui concerne nos analyses, nous proposons donc de souligner simplement quelques notions

[i] Pour en savoir davantage, nous vous invitons à consulter Williams G. (2006), « La linguistique et le corpus: Une affaire prépositionnelle », *Texto, revue de linguistique en ligne*. http://www.revue-texto.net/Parutions/Livres-E/Albi-2006/Williams.pdf

[ii] A en croire Cori et al. (2008 :5), Condamine (2005 :17), (f)ace à l'inflation de publications assez nombreuses dans le courant anglo-saxon *Corpus Linguistics*, on trouve peu d'ouvrages français équivalents.

[iii] Ici, nous nous permettons de citer Williams (2006) qui déconseille de se borner par les définitions : « Si la linguistique de corpus existe comme discipline autonome, où se trouvent les frontières avec d'autres disciplines ? Là, je retourne la question : avons-nous vraiment besoin de frontières quand toutes nos propres études sur le langage prouvent que les frontières n'existent pas ? La linguistique de corpus, comme d'autres disciplines de la linguistique, rentre parfaitement dans la notion de prototype, avec un nœud central et une périphérie qui glissera subtilement vers d'autres disciplines dans un continuum. Les catégories n'existent pas en soi, nous les créons pour mieux saisir la complexité. Parler des linguistiques de corpus c'est noyer le poisson, si tout le monde le fait, personne ne le fait, et tout le monde est perdant. La linguistique de corpus existe, elle est récente et sa méthodologie et son épistémologie se forgent. Pour la forger, il faut simplement la reconnaître » (p.157).

pertinentes de la linguistique de corpus et en particulier la façon dont nous les appliquons à notre enquête sur les particules énonciatives en birman.

Nous présentons les quatre approches adoptées dans notre analyse de corpus, avant d'arriver à la définition de corpus appliquée, qui est un peu différente de celle employée d'une façon générale parmi les linguistes.

3.1.1. Approche descriptive des faits réels

D'une manière générale, la linguistique de corpus s'intéresse à la langue en contexte sous la forme de grands ensembles de textes – les corpus – afin de « révéler les choix linguistiques opérés par des locuteurs dans des contextes réels » et à « comprendre les mécanismes de la communication » (Williams, 2005 :13). L'approche de la linguistique de corpus qui est par ailleurs l'apanage d'une linguistique descriptive, représente ainsi une « alternative à des démarches fondées sur l'introspection du linguiste ou sur l'élicitation de jugements des locuteurs » (Mondada, 2005 :75). C'est tout à fait un avis partagé par d'autres linguistes : par exemple, le motif essentiel de la linguistique de corpus est, selon Jacques (2005), l'intérêt grandissant de la linguistique pour ces aspects impossibles à traiter par l'introspection et l'intuition. Cela correspond/convient exactement à l'objectif principal de notre enquête, *i.e.* de chercher à comprendre le rôle et les fonctions discursives des particules en birman dont les fonctions grammaticales seules ne donnent pas une explication satisfaisante [cf. 1.3.2.2 : Morphèmes dépendants]. Cela n'est guère étonnant si l'on considère l'observation de Jacques (2005 :25)[i] que la capacité des phénomènes discursifs échappe souvent aux capacités d'invention hors situation réelle.

[i] A notre avis, l'article de Marie-Paule Jacques (2005) intitulé 'Pourquoi une linguistique de corpus ?' a bien esquissé des vertus ainsi que des critiques du domaine. Nous trouvons particulièrement utile la section sur les points forts de la linguistique de corpus qui sont (en résumé) de (1) mettre en lumière des fonctionnements qui échappent à l'intuition ; (2) corriger les intuitions sur le fonctionnement de la

En nous appuyons sur le corpus du birman parlé (de divers genres et locuteurs) enregistré que nous avons constitué, notre description des particules énonciatives se fonde sur des faits observables *i.e*. des traces (transcriptions) des paroles produites pour des raisons de communication entre êtres humains, et non sur des productions (parfois « artificielles ») produites par ou pour l'introspection des linguistes. Nous estimons ainsi qu'une telle approche nous permettra de dépasser la grammaire normative, et par conséquent c'est une approche qui nous semble idéale et séduisante, surtout pour mettre en lumière des fonctions discursives des particules birmanes, dont le sens est souvent à interpréter dans le contexte.

3.1.2. Approche contextualiste

> « *You shall know a word from the company it keeps* ».
> [Vous connaissez un mot par la compagnie qu'il tient]
> [Firth (1935), traduction citée dans William (2006)]

Le propos de Firth ci-dessus décrit précisément la situation des particules énonciatives en birman, dont la valeur sémantique est étroitement liée au contexte, comme nous l'avons indiqué précédemment. De ce fait, nous jugeons ce choix d'approche idéal, car la linguistique de corpus est, selon Williams (2006 :153), de par sa nature, contextualiste. Dans cette approche, il s'agit d'observer de grands ensembles de textes soigneusement choisis pour les besoins de la recherche linguistique. Par la suite, ces textes représentent une partie de la langue en action. Dans ce sens, l'environnement de la langue, avec tous les aspects sociolinguistiques, doit être pris en compte, c'est à dire, le contexte culturel et le contexte situationnel. De nos jours, personne ne niera plus que le sens (dans la compréhension de la langue) ne puisse pas être évalué en dehors du contexte situationnel. Reconnaissant que le contexte est primordial dans nos interprétations des particules énonciatives, nous consacrons également une place considérable au contexte dans la présentation des données afin de bien accompagner nos lecteurs.

3.1.3. Approche inductive (*corpus-drive approach*)

Parmi les deux approches fondamentales de la linguistique de corpus, notre travail s'inscrit plutôt dans l'approche inductive motivée par le corpus, *corpus-driven* en

---

langue ; Elle permet (3) d'avoir des indications en terme de fréquence et établir des relations statistiques entre ensembles de faits ; (4) d'atteindre et rendre compte de la variation. (p.25-26)

anglais[i] (cf. Sinclair 1987 ; Tognini-Bonelli 2001 ; Williams 2005 ; Hnin Tun 2006, ente autres) que de l'approche déductive appliquée au corpus, *corpus-based* en anglais (cf. Leech 1987, Biber 1998, Williams 2005, etc.). L'approche déductive[ii] utilise les donnés dans le corpus pour confirmer ou infirmer la validité des hypothèses. Les adeptes de cette approche cherchent à explorer des aspects et des phénomènes connus de la langue, dans les situations concrètes des corpus.

L'approche de notre choix, i.e. **l'approche inductive[iii] motivée par le corpus**, cherche à examiner en revanche sans *a priori* ni charpente théorique préétablie, des unités de corpus telles que des éléments récurrents (*recurring patterns* en anglais), des constructions lexicales ainsi que syntaxiques, afin d'aboutir « à la description des régularités linguistiques et à la formulation des conceptions théoriques » (Martelli, 2003 :15). Par exemple pour chaque particule dans notre étude, nous examinons ses éléments récurrents qui servent à identifier les contextes (syntaxiques ou situationnelles) dans lesquels les particules sont employées. Pendant les analyses nous posons des questions telles que **Est-ce que la particule X apparaît régulièrement avec les pronoms ou les noms ? Dans une construction affirmative ou négative ? Au début ou à la fin du tour de parole ?** et ainsi de suite. Nous estimons que les réponses à ces questions mettent en lumière les fonctions discursives des particules.

Toutefois il faut souligner que les approches inductive et déductive ne sont pas complètement séparées dans la pratique. Nous examinons les particules sans théories *a priori* en ce qui concerne leurs fonctions spécifiques : tout au long de notre enquête, nous ne savions (presque) jamais quels aspects discursifs ou syntaxiques se manifesteraient dans les concordances des données par exemple, mais il faut reconnaître aussi que nous entamons l'enquête avec une présomption, soutenue par nos précédentes recherches, que les particules birmanes ont des fonctions discursives. Néanmoins à notre connaissance, il n'existe pas encore de théories bien fondées ni de travaux

[i] (I)n a corpus-driven approach, the commitment of the linguist is to the integrity of the data as a whole, and descriptions aim to be compréhensive with respect to corpus evidence. The corpus, therefore, is seen as more than a repository of examples to back pre-existing théories or a probabilistic extension to an already well-defined system. The theoretical statements are fully consistent with, and reflect directly, the évidence provided by the corpus … The thoery has no independent existence from the évidence and the général methodological path is clear :observation leads to hypothesis leads to généralisation leads to unification in theoretical statement. [Tognini-Bonelli, 2001 : 84-85]

[ii] Connu également comme *top down* en anglais, qui signifie *démarche descendante*

[iii] *bottom-up approach* en anglais, qui siginifie *démarche ascendante*

suffisants sur les fonctions discursives des particules birmanes. En somme, on peut dire que notre approche est principalement inductive.

3.1.4. Approche probabiliste

En outre, comme l'a souligné Jacques (2005), les corpus (tels qu'ils sont conçus dans la discipline) peuvent révéler « des fonctionnements que l'intuition aurait pensé marginaux et qui sont en fait très répandus […] On considère facilement certains usages marginaux ou 'limites' alors qu'ils sont en fait des usages courants et vice-versa » (p.25). Or les corpus n'étant jamais un produit exhaustif ni fini, il est évident que les conclusions tirées à partir d'observations basées sur les corpus n'aboutissent pas à la description de la vérité « absolue » d'un système linguistique. Néanmoins un phénomène ou une construction langagière qui se répète et se produit avec une fréquence considérable ne peut pas être anodin(e) ni ne représente un cas singulier ou idiosyncratique. En tenant compte de la question d'usage, et par la suite des phénomènes observés et décrits dans un contexte, l'approche de la linguistique de corpus nous amène à concevoir la grammaire non comme un modèle absolu de la langue mais comme un modèle probabiliste[i]. Par « probabiliste », nous voulons dire tout simplement que **selon nos observations dans le corpus de cette étude, la particule examinée est susceptible de se produire dans une telle construction syntaxique dans un tel contexte en birman parlé contemporain.**

Cette approche probabiliste semble séduire de Robillard (2001) qui a eu l'audace de proposer d'appliquer la théorie du chaos à la linguistique. Selon lui, l'approche probabiliste représente la voie la plus appropriée pour saisir la variation et la coexistence, dans la langue, de faits 'réguliers' et 'irréguliers'. Il va même plus loin pour soutenir que les faits irréguliers constituent « un résidu de 'faits linguistiques' attestés mais néanmoins parasites parce qu'on ne comprend pas à quoi ils peuvent bien servir » (p. 179). D'une certaine manière n'est-ce pas le cas des particules énonciatives en birman, dont nous avons du mal à cerner les fonctions ? : elles n'affectent pas la grammaticalité de l'énoncé dans la construction syntaxique, donc on ne savait pas à quoi elles peuvent servir. Or leur présence est clairement attestée dans le discours naturel. Et c'est en nous servant de l'approche probabiliste de la linguistique de corpus

[i] Ne pas confondre avec 'probabilistic grammar' de Suppes (1972) qui porte sur une étude plus théorique

que nous envisageons de pouvoir faire apparaître leurs fonctions discursives dans la langue naturelle.

Par ailleurs, nous savons tous que lorsque l'on tâche d'examiner la langue réelle en cours d'utilisation, nous risquons de nous trouver face à « la résistance à la systématisation que présentent parfois les faits langagiers » (Candamine 2005 :16). Les linguistes qui travaillent sur la sémantique à base des corpus ont constaté qu'une telle approche s'oppose à « la vision d'un locuteur 'idéal' qui permettrait de décrire un modèle stable, contrôlé et prédictif » (ibid). Elle soutient ainsi que la prise en compte des corpus dans les analyses linguistiques (en particulier sémantique) suppose en effet « une rencontre avec un principe de réalité ». Si l'on tâche de décrire cette réalité langagière ou linguistique telle que le fonctionnement des particules énonciatives en birman, dont les règles semblent peu évidentes ou cohérentes « à première vue », il nous semble effectivement judicieux de commencer par l'approche probabiliste.

3.1.5. Définition du corpus de notre étude

Comme nous l'avons signalé précédemment, il existe des variantes lorsqu'il s'agit de la définition du terme « corpus ». Tout d'abord, notons d'emblée que **dans la linguistique de corpus, le terme *corpus* est utilisé comme abréviation de corpus informatisés**. La définition officielle de Sinclair, le doyen de la linguistique de corpus, est comme suit :

> Une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage
> [Sinclair, 1994 :2, traduction de Habert et al. 1997][i]

McEnry & Wilson (1996 :177) proposent en revanche les définitions à trois niveaux : (1) (d'une manière générale) un ensemble quelconque de textes ; (2) (définition plus courante) un ensemble de textes lisibles à la machine ; (3) (plus

[i] A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language (Sinclair, 1994 :2). N.B. c'est par ailleurs le point de vue adopté par EAGLES (European Advisory Group on Language Engineering Stardards).

précisément) une collection finie de textes lisibles à la machine, qui est censée être représentatifs d'une langue ou d'une variété de la langue[i].

Avec des textes réels et des données attestées, le corpus s'oppose aux exemples *ad hoc* forgés pour les besoins d'une théorie ou d'une étude. Il faut noter que le corpus « brut » n'obéit pas au jeu de règles érigées *a priori,* mais qu'il nous permet de reconstituer *a posteriori* des régularités et nous amène à des analyses partielles (Bommier-Pincemin, 1999[ii]).

En ce qui nous concerne, le corpus est tout simplement un ensemble de textes stockés sous forme électronique, lisibles à la machine et traitables informatiquement (pour saisir des concordances, par exemple) (William, 2005, Jacques 2005,). Il s'agit d'une collection de textes (transcriptions de la langue orale) de taille importante, constitués du moins des échantillons de texte qui dépassent le stade de la phrase. Il est peut-être utile de préciser aussi que le corpus n'est pas une simple collection de « sacs de mots », ni un archivage de textes transcrits. En effet dans la constitution de notre corpus, il s'agit d'une collection de textes généralement récoltés un peu « au hasard », certes, mais nous prenons bien soin de mettre l'accent sur :

- L'authenticité et l'aspect naturel de la langue (i.e. la langue enregistrée n'est à aucun moment produite pour les analyses linguistiques)
- Variété (de locuteurs et de genres) afin d'assurer la représentativité.

L'outil de base en analyses de corpus est le concordancier, un logiciel qui nous permet d'observer la fréquence des mots et d'identifier des collocations[iii], ce qui nous permet d'observer des propriétés distributionnelles des mots (les particules dans notre

---

[i] (1) (loosely) any body of text ; (most commonly) a body of machine readable text ; (3) (more strictly) a finite collection of machine-readable text, sampled to be maximally representative of a language, or variety (1996 :177, dans Pearson 42-43)

[ii] Cette dernière propose trois conditions à satisfaire dans la constitution des corpus : conditions de signifiance, d'acceptabilité et d'exploitabilité (p.416).
Conditions de signifiance : un corpus est constitué en vue d'une étude déterminée (*pertinence*), portant sur un objet particulier, une réalité telle qu'elle est perçue sous un certain angle de vue (et non sur plusieurs thèmes ou facettes indépendants, simultanément) (*cohérence*).
Conditions d'acceptabilité : le corpus doit apporter une représentation fidèle (*représentativité*), sans être parasité par des contraintes externes (*régularité*). Il doit avoir une ampleur et un niveau de détail adaptés au degré de finesse et à la richesse attendue en résultat de l'analyse (*complétude*).
Conditions d'exploitabilité : les textes qui forment le corpus doivent être commensurables (*homogénéité*). Le corpus doit apporter suffisamment d'éléments pour pouvoir repérer des comportements significatifs (au sens statistique du terme) (*volume*).

[iii] Une collocation (ou un colloquât) est un mot co-occurent, dont la fréquence est significative dans le contexte immédiat du nœud.

cas). Le Figure 1 illustre les résultats générés par le concordancier *Wordsmith* en ce qui concerne la particule KA (/ka[1]/ en transcription phonétique), surlignant les co-occurrences récurrentes : ici nous pouvons voir que KA se manifeste très fréquemment avec LEEH (/lɛ[3]/ en transcription phonétique).

Figure 1 : Exemple des collocations du lexème KA, générées par le concordancier

Il est évident que de tels résultats seuls, issus des analyses quantitatives ne nous révèlent pas grand-chose : ils requièrent d'être interprétés. C'est à partir de ces observations que nous envisageons d'établir des régularités dans l'emploi des particules afin de construire des conceptions théoriques.

Pour récapituler, la linguistique de corpus nous permet d'examiner la langue en contexte, sous la forme de grands ensembles de textes, et de repérer, d'une manière rapide et avec précision, les environnements dans lesquels les particules apparaissent. Par la suite la récurrence des environnements (con)textuels met en lumière les tendances dans l'emploi des particules.

3.2. Particularités birmanes : un défi en linguistique de corpus

Si, grâce à l'avancement informatique de nos jours, nous n'avons aucune difficulté pour saisir des textes en caractères birmans à l'ordinateur, il s'avère toutefois que la réalisation d'analyses de corpus en birman suscite des difficultés quelque peu inattendues. Nous les détaillerons avec des exemples dans 3.2.1, et terminerons le chapitre par l'explication de notre système de transcription peu conventionnel dans 3.2.2.

3.2.1. Langue syllabique, langue à tons

Pour commencer, rien que de mesurer ou décrire la taille du corpus devient « problématique » car il n'existe pas encore un programme de traitement de texte qui puisse faire un calcul automatique du nombre de mots en birman, comme cela se fait pour les langues occidentales. C'est à dire que la notion du « mot » (qui représente l'unité d'analyse de base en linguistique de corpus) est conçue sur la base des langues qui utilisent une écriture alphabétique, telles que l'anglais et le français. Or en ce qui concerne le birman, le système d'écriture est basée sur les syllabes. Par conséquent il est parfois impossible pour la machine de distinguer automatiquement entre un mot bi- ou polysyllabique et deux mots monosyllabiques, sans qu'un être humain le précise *a priori* pour la machine. Pour un locuteur natif, la lecture des énoncés est suffisante pour discerner les lexèmes qui sont particules ou non, mais on ne pourra le décider que par l'examen du contexte.

Pour illustrer ce phénomène, prenons par exemple deux morphèmes က /ka^{1}/ ('danser' ou marque de sujet ou de point de départ) ; et စား/sa^{3}/ ('manger'), qui semblent fonctionner (jusqu'ici) comme n'importe quel mot en anglais ou en français. Or ces deux morphèmes ensemble en birman peuvent former également un autre « mot » tel que ကစား /kə + sa^{3}/ qui signifie 'jouer'. A cela s'ajoute le fait que le birman

n'utilise pas non plus l'espace entre deux mots[i] comme c'est l'usage en français [cf. 1.2 : Représentation graphique du birman] : l'espace en birman marque d'habitude la fin d'un syntagme (*clause* en anglais) ou d'une phrase. De ce fait, il est par exemple impossible (en l'état actuel) au logiciel de reconnaître automatiquement combien de mots (morphèmes) à compter dans les deux énoncés (3.1) et (3.2) qui, uniquement d'après l'écriture en birman, semblent identiques. Or les exemples avec gloses en dessous indiquent qu'il y a 4 morphèmes dan (3.1) et 3 morphèmes en (3.2).

(3.1) သူကစားတယ်။	**(3.2)** သူကစားတယ်။

(3.1) သူ က စား တယ်။
θu^{2} ka^{1} sa^{3} tɛ^{2}
3SG MSN manger MFV
Il mange.

(3.2) သူ ကစား တယ်။
θu^{2} kə.sa^{3} tɛ^{2}
3SG jouer MFV
Il joue.

En outre, le logiciel concordancier que nous avons choisi pour cette étude, *Wordsmith Tools*[ii] (Mike Scott, 1998), n'est pas compatible avec les caractères birmans. Le créateur du logiciel ne cesse de le mettre à jour afin de le rendre compatible avec autant de polices de caractères que possible, mais pour ce faire il faut du moins que les textes soient saisis en caractère Unicode[iii]. Néanmoins les dernières polices birmanes ne sont que quasi Unicode[iv]. Aussi, comme solution pratique[i], avons-nous résolu à utiliser

[i] « Le birman s'écrit de gauche à droite sans séparer le plus souvent les mots. Deux signes de ponctuation, une barre ou deux barres, correspondent à nos virgules et à nos points ». [cf. Wikipédia : http://fr.wikipedia.org/wiki/Birman_%28langue%29]

[ii] Après avoir exploré plusieurs logiciels concordancier depuis que nous avons commencé les recherches à base de corpus informatisés, il s'est avéré que *Wordsmith Tools* nous semblait le plus efficace pour le birman et le plus facile à manipuler, sans avoir à apprendre d'abord les techniques et outils informatiques complexes. Afin de pouvoir nous servir du corpus déjà constitué (qui comprend un travail de transcription d'abord en birman et ensuite en caractère latin), nous avons choisi d'utiliser le même logiciel pour cette étude.

[iii] **Unicode** est une norme informatique, développée par le *Consortium Unicode,* qui vise à permettre le codage de texte écrit en donnant à tout caractère de n'importe quel système d'écriture un nom et un identifiant numérique, et ce de manière unifiée, quelle que soit la plate-forme informatique ou le logiciel. La dernière version, **Unicode 6.1.0**, est publiée depuis le 31 janvier 2012. [cf. Wikipédia, 13 avril 2012 : http://fr.wikipedia.org/wiki/Unicode]
Pour en savoir plus sur Unicode, voir : http://www.tuteurs.ens.fr/unix/editeurs/unicode.html

[iv] Conversion de textes en Unicode birman toujours en version béta et en phase de test [http://www.lexilogos.com/clavier/birman.htm].
Il est important de souligner que jusqu'aujourd'hui, les deux polices Unicode birmanes - *Padauk*, utilisé chez BBC (cf. http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&id=Padauk) et *Zawgyi*, utilisé chez

la transcription du corpus birman en caractères latins. Toutefois, cette solution conduit à un nouvel obstacle car le birman est une langue à tons [cf. 1.3.1. Aspects phonologiques], mais les caractères latins ne peuvent pas accommoder les marques de ton en birman, et les symboles diacritiques ne fonctionnent pas bien avec *Wordsmith*. En somme, nous avons conclu que les outils informatiques (ainsi que les méthodes d'analyse de corpus) tels qu'ils sont conçus initialement, ne permettent que des possibilités limitées pour faire des analyses de corpus birman, et par conséquent nous imposent certains travaux d'improvisation et remaniements. Nous proposons ainsi d'utiliser :

1) **les syllabes comme mesure du corpus** (le logiciel calcule le nombre de syllabes en birman à la place de mots pour d'autres langues) ; et

2) de mettre au point notre système de transcription propre que nous allons expliquer dans 3.2.2.

Néanmoins, après que les outils ont généré les statistiques (la liste des fréquences et des collocations), il nous reste encore une autre complication à résoudre avant de passer au travail d'interprétation. La plupart des particules qui nous intéressent étant homonymes ou polysémiques, il est nécessaire de faire quelques travaux manuels tels que trier et isoler des particules à examiner. Prenons par exemple deux lexèmes suivants : ပါ/pa²/ peut être particule de politesse ou signifier aussi 'être avec ; avoir avec soi' ; တယ်/tɛ²/ peut être particule de fin de phrase ou équivalent de 'très', ou une interjection hostile, équivalent de 'Dites donc !' En outre /tɛ²/ peut même être une partie du mot bi-syllabique ဟိုတယ်/ho².tɛ²/ 'hôtel' et ainsi de suite. Aussi pour identifier les fonctions discursives de la particule /tɛ²/, faut-il d'abord éliminer « à la main », parmi les 5 627 occurrences de /tɛ²/, celles qui ne correspondent pas à l'objet de notre enquête (ex. /tɛ²/ dans le mot /ho². tɛ²/ ou, avec le sens équivalent de 'très' sont à éliminer). Il s'agit en somme d'un travail de main d'œuvre assez intensif afin de

---

VOA (http://www.zawgyi.net/) - par exemple, ne sont pas tout à fait compatibles. Pourtant l'idée de l'unicode est que l'on puisse lire tout texte écrit en Unicode, quelle que soit la plateforme informatique ou le logiciel.

[i] Et intermédiaire en attendant les logiciels concordanciers qui puissent traiter les corpus en caractère birman.

cerner les fonctions discursives des particules en birman à base des corpus de grande taille. A notre avis cela vaut la peine, car les résultats que nous envisageons de découvrir seront sûrement de grande valeur.

En somme, là où un logiciel calculerait des mots en d'autres langues, en ce qui concerne le corpus birman en revanche, notre logiciel va calculer des syllabes. Rappelons aussi que nous ne tenons pas compte de l'aspect prosodique dans la présente étude, non sans reconnaître son importance dans le discours parlé, mais nous sommes d'avis qu'il est possible d'atteindre des résultats significatifs en ce qui concerne les fonctions discursives des particules birmanes à partir d'un examen des aspects morphosyntaxiques, comme le montreront nos résultats. Dans les langues occidentales, il y a de l'intonation. En birman aussi, certes, mais l'intonation n'a pas la même signification dans les deux langues : ce qui s'exprime par l'intonation en français en grande partie est souvent véhiculé par les particules en birman, où l'intonation semble être plutôt secondaire. Nous employons donc une économie provisoire, mais comme vous allez voir dans les résultats, ce que nous découvrons vaut et équilibre la prosodie française et quelque chose d'autre.

3.2.2. Système de transcription

Notre système peu conventionnel de transcription du corpus en caractère latin s'explique comme suit :a) Les tons sont indiqués par le nombre de voyelles :

- **Une voyelle (/a/ /o/ /u/ …)** **représente Ton 1 (court/haut)** ;
- **Deux voyelles (/aa/ /oo/ /uu/…)** **représentent Ton 2 (neutre/bas)** ;
- **Trois voyelles (/aaa/ /ooo/ /uuu/…)** **représentent Ton 3 (long/haut descendant)**

Le tableau 51 illustre quelques exemples avec notre système de transcription comportant les trois tons [voir la description des ton dans 1.3.1.2 : Tableau 7 et 8]. Dans les exemples, les tons sont marqué par les chiffres 1, 2, 3 en exposant.

Ton 1	Ton 2	Ton 3
စ /sa/ 'commencer'	စာ /saa/ 'lettre'	စား /saaa/ manger
ပို့ /po/ envoyer	ပို /poo/ être en plus, être en excès, dépasser	ပူး /pooo/ insecte, bactérie, transporter sur son dos, se montrer amoureux

ကု **/ku/** soigner, prescrire des médicaments	ကူ **/kuu/** aider, roucouler	ကူး **/kuuu/** traverser, copier
လန့် **/lan/** sursauter de peur	လံ **/laan/** classificateur :brasse, quatre coudées	လမ်း **/laaan/** chemin, rue

Tableau 51 : Illustration de système de transcription avec trois tons

b) Les consonnes aspirées sont doublées, comme l'illustre le tableau 52 [**N.B**. Le birman fait la distinction phonologique entre la consonne aspirée et non-aspirée. Cf. 1.3.1.1. Les consonnes]. Dans la transcription des exemples, la consonne aspirée est notée par un h en exposant [ex. /k^h t^h p^h s^h/ , etc.]

Consonne non-aspirée	Consonne aspirée
ကိုး **/kooo/** 9	ခိုး **/kkooo/** voler
တောင် **/tauun/** montaigne	ထောင် **/ttauun/** prison
ပန်း **/paaan/** fleur	ဖမ်း **/ppaaan/** attraper
စိတ် **/seiq/** esprit	ဆိတ် **/sseiq/** chèvre

Tableau 52 : Illustration de système de transcription avec consonnes non-aspirée et aspirée

c) En outre, A (en majuscule) remplace le schwa traditionnel /ə/, tel qu'il est illustré dans le tableau 53.

A	Consonne aspirée
က **/ka/** danser	ကလေး **/kA** leee/ enfant
စ– **/sa/** commencer	စကား **/sA** kaaa/ parole
လ **/la/** lune	အလကား **/A lA** kaaa/ Dépourvu de raison, gratuitement

Tableau 53 : Illustration de système de transcription avec le schwa /ə/

N.B. Pour la transcription des exemples dans les discussions, voir : Code de transcription]

En somme, les enregistrements sont d'abord transcrits en écriture birmane et ensuite retranscrits en caractère latin, selon le système expliqué ci-dessus. Rappelons aussi qu'à cause du *sandhi*, les consonnes ont souvent des prononciations différentes

selon l'environnement phonétique [cf. 1.3.1.3]. Par exemple, la particule de politesse ပါ /pa²/ se prononce /pa²/ ou /ba²/ selon la syllabe qui la précède. L'écriture birmane neutralise cet effet, et ainsi afin d'assurer une cohérence et une efficacité optimales, les textes sont systématiquement transcrits d'après le système d'écriture birman qui est relativement plus stable (que le birman oral). Nous ne tenons donc pas compte de l'effet du *sandhi* dans notre transcription : ပါ /pa²/ est ainsi transcrit tout au long du corpus comme /pa²/, quelle que soit la prononciation dans les enregistrements.

Pour terminer, nous devons souligner que notre corpus utilise un système d'annotation relativement simple, mais à notre avis adéquat pour la présente étude[i]. Par exemple, sans préjuger de l'importance de l'élément prosodique du discours parlé : des aspects prosodiques et paralinguistiques sont exclus de la transcription, et nous nous concentrons sur les aspects morpho-syntaxiques dans la présente enquête. En revanche, nous avons marqué les changements de locuteur [ex. <$x> pour locuteur X, <$y> pour locuteur Y, etc.], car nous nous intéressons au fonctionnement des tours de parole, qui est une des fonctions discursives qui nous intéresse. Nous notons également dans notre transcription des rires, et des pauses que nous jugeons longues d'une manière significative pour notre étude.

Nous allons voir dans le chapitre suivant la description détaillée du corpus analysé, et les transcriptions en birman et en transcription utilisée sont présentées dans le Tome 2.

[i] Il s'agit ici d'une des premières études sur les particules énonciatives en birman sur la base d'analyses de corpus informatisés de grande taille.