Enjeux et méthodes pour la création de corpus en langues peu dotées. Application à la classification de textes pour l'apprentissage du birman.

Jennifer Lewis-Wong

INSTITUT NATIONAL DES LANGUES ET CIVILISATIONS ORIENTALES
École doctorale n°265 : *Langues, Littératures et Sociétés du monde*
ERTIM Equipe de Recherche Textes, Informatique, Multilinguisme (EA2520)
LACITO Langues et Civilisations à Tradition Orale (UMR 7107)

THÈSE

présentée par

Jennifer Lewis-Wong

soutenue le 27 janvier 2023
pour obtenir le grade de Docteur de l'INALCO
en Traitement automatique des langues

Enjeux et méthodes pour la création de corpus en langues peu dotées. Application à la classification de textes pour l'apprentissage du birman.

Thèse dirigée par :

M. Mathieu Valette	Professeur des universités, Inalco
M[me] San San Hnin Tun	Maître de conférences, Inalco

Rapporteurs :

M. Alexis Michaud	Directeur de recherche, CNRS
M. Justin Watkins	Professeur des universités, SOAS

Membres du jury :

M. Michel Antelme	Professeur des universités, Inalco
M. Thomas François	Maître de conférences, UCLouvain
M[me] San San Hnin Tun	Maître de conférences, Inalco
M. Alexis Michaud	Directeur de recherche, CNRS
M. Mathieu Valette	Professeur des universités, Inalco
M. Justin Watkins	Professeur des universités, SOAS

Remerciements

Je tiens à remercier tout d'abord mes deux directeurs de thèse, Mathieu Valette et Sayama San San Hnin Tun, qui ont non seulement accepté de diriger ce qui semblait parfois être un sujet de recherche plutôt alambiqué, mais ont continué à m'encourager malgré mes doutes quant à ma capacité à arriver à ce stade. Je leur suis extrêmement reconnaissante pour leur patience au fil des ans, non seulement en tant que directeurs de thèse, mais aussi en tant que professeurs de TAL et de langue birmane respectivement.

J'ai également une dette de gratitude envers François Stuck, dont les encouragements doux, mais fermes m'ont convaincu de terminer. Ses yeux de lynx qui ont parcouru mon manuscrit m'ont impressionné au plus haut point et ont grandement amélioré la qualité de la rédaction. Les erreurs qui subsistent sont cependant bien les miennes.

Si c'est grâce aux personnages inoubliables du département Asie du Sud-est et Pacifique de l'Inalco que j'ai pris goût aux langues et cultures de l'Asie du Sud-est, c'est grâce aux enseignants de linguistique et de TAL, la plupart membres de l'ERTIM, que j'ai découvert le TAL et tout son nouveau monde de possibilités et de frustrations que je ne peux que qualifier de partiellement addictif. L'environnement de travail à l'ERTIM a toujours été rempli de gentillesse, de respect, d'enthousiasme, de bonne compagnie et de bonne humeur, grâce à son personnel permanent, à ses doctorants passés et présents, à ses stagiaires et aux camarades de classe de ma promotion de Master, un groupe d'étudiants particulièrement sympathique et soudé. On ne peut pas demander mieux.

Sans l'aide de Vincent Berment, ce projet n'aurait jamais vu le jour, car c'est son adaptation de son outil de segmentation laotien pour le birman qui m'a permis d'envisager ce projet. Lorsque j'ai commencé, c'était le seul outil de segmentation disponible pour le birman et son système m'a permis d'expérimenter la segmentation d'une manière qui n'aurait pas été possible avec d'autres méthodes. Il a toujours été encourageant et prêt à apporter les petites modifications techniques dont j'avais besoin. Merci Vincent !

Pour leur aide à la création du corpus birman, je tiens à remercier non seulement Sayama San San Hnin Tun, mais aussi May Hmoo et Khwankhaw Sangkhaphathanon pour la correction de corpus et les discussions intéressantes sur le birman, et l'écrivain-traducteur Moe Thet Han pour les textes du corpus littéraire.

Au fil des ans, de nombreux amis et collègues m'ont aidée à résoudre des problèmes techniques, car je suis parfois un peu lente à la détente, ayant appris la programmation à un âge, disons plus avancé que d'habitude. Je remercie particulièrement François Stuck,

Satenik Mkhitaryan, Damien Nouvel, Li Yun Yan et Xiaoting Luo pour leur aide et leurs encouragements.

Ce manuscrit a été rédigé en LaTeX grâce à la formation dispensée par Thomas Pellard (2018). L'inclusion des polices birmanes mutuellement incompatibles n'aurait pas été possible avec un traitement de texte classique. Je lui suis donc particulièrement reconnaissante pour ses conseils en la matière.

Je tiens également à remercier les nombreux professeurs de langues étrangères que j'ai connus au fil des années en tant qu'étudiant sur trois continents, bien trop nombreux pour être cités par leurs noms, mais à qui j'exprime ma gratitude pour leurs connaissances, leur générosité et leur patience. Sans leur influence, mon amour et ma fascination pour les langues étrangères n'auraient pas pu s'épanouir et me soutenir pendant si longtemps. Chaque professeur a une façon unique d'enseigner, et une vision unique de sa langue, que j'ai toujours trouvées enrichissantes. La valeur de l'expérience d'un professeur est difficile à quantifier, mais pour cette étudiante au moins, elle est appréciée avec gratitude.

Enfin, je n'ai pas de mots assez forts pour remercier Richard Wong, la perle rare qui m'a non seulement encouragée, mais aussi supportée pendant toutes ces années d'études, en cuisinant pour mes amis et en me gardant le moral. Ce n'est pas sans un certain sentiment de culpabilité que je peux enfin t'annoncer que tu as le temps de prendre réellement ta retraite. 感激不盡！

Je dédie cette thèse à la mémoire de mon arrière-grand-père, Eddie Smith, postier et linguiste amateur passionné, dont la lecture des mémoires témoignant de ses années de travail comme sapeur-télégraphiste pendant la Grande Guerre m'a donné confiance en ma propre curiosité intellectuelle ces dernières années.

Introduction

> Without grammar very little can be conveyed,
> without vocabulary nothing can be conveyed.
> (Wilkins 1972)

1 Contexte et problématique

1.1 *Le rôle du vocabulaire dans l'apprentissage d'une langue étrangère*

La corrélation entre connaissances lexicales et d'autres compétences linguistiques en langue étrangère (de lecture, de compréhension de l'oral, de l'écrit) constatée par la recherche en linguistique appliquée (Laufer 1992 ; Alderson 2005 ; Milton 2010) donne une place importante à l'apprentissage du vocabulaire. Mais, si l'importance du vocabulaire dans l'apprentissage d'une langue étrangère paraît une évidence, il n'en va pas de même quant aux moyens de l'acquérir, car l'acquisition de la grande quantité de vocabulaire nécessaire à la communication constitue un défi de taille pour l'apprenant. Selon Meara (1980), les apprenants de langues étrangères eux-mêmes reconnaissent que l'acquisition du vocabulaire leur pose des difficultés considérables, surtout après qu'ils ont dépassé le niveau débutant. Si l'apprenant aspire à une compétence proche du locuteur natif, la tâche peut paraître insurmontable, car l'exposition lexicale d'un apprenant de langue étrangère (notamment celui d'une langue peu enseignée) est généralement beaucoup moins riche que celle d'un apprenant de sa langue maternelle. Des estimations de la taille du vocabulaire d'un locuteur natif varient (Nation 1993), mais on estime par exemple qu'un locuteur natif anglophone éduqué possède un vocabulaire de l'ordre de 16 000 à 20 000 mots ou plutôt familles de mots (Schmitt 2010) sans tenir compte des mots composés et locutions. Il se peut que cela soit moins pour d'autres langues ayant assimilé moins d'éléments d'autres langues et dont la morphologie est plus riche (Ringbom 1983), mais la perspective demeure intimidante pour un débutant. Au lieu de se concentrer sur un objectif aussi

ambitieux, Nation et Meara (2010) préconisent une approche axée sur les besoins de l'apprenant et sur l'utilité du vocabulaire, déterminée par la fréquence et la répartition des éléments dans un corpus de textes représentatif. Ces informations sont utilisées pour créer une liste de vocabulaire de base ordonnée par fréquences décroissantes. Jusqu'ici, l'élaboration de ces listes a surtout concernée le monde anglophone, mais il existe maintenant, pour de nombreuses langues, des ouvrages répertoriant de tels vocabulaires fréquentiels de base, tirés d'un traitement statistique de corpus de grande envergure. Ces listes de vocabulaire de haute fréquence composées du vocabulaire de base le plus utile sont complétées par d'autres listes de fréquence issues de corpus spécialisés centrés sur les besoins et les intérêts des apprenants (le langage académique ou des affaires, par exemple). Quant à l'apprentissage et au renforcement du vocabulaire au-delà d'une liste de référence de base, Nation (2001) recommande surtout la lecture extensive de textes appropriés au niveau de l'apprenant : de textes simples (des textes destinés à des enfants natifs), simplifiés (modifiés afin de limiter le vocabulaire du texte et diminuer le vocabulaire de basse fréquence) ou authentiques. Pour faciliter l'apprentissage du vocabulaire donc, il faut créer des corpus de textes pour établir quel vocabulaire est de haute fréquence, et ensuite avoir un moyen d'estimer le niveau de difficulté de textes ou *lisibilité*.

1.2 *L'importance du contexte dans le traitement automatique*

L'apprentissage de toute langue étrangère se déroule dans un contexte culturel hétérogène : le bagage culturel de l'apprenant d'une part, et le contexte culturel de la langue cible que l'apprenant essaie d'apprendre d'autre part. Pour l'enseignant de langue, le premier est parfois très difficile à cerner, car non seulement les apprenants proviennent de cultures différentes, mais peuvent aussi avoir eu des expériences linguistiques variées. Pour l'apprenant, le contexte culturel de la langue cible qu'il rencontre au cours de son apprentissage est parfois inattendu, car il comprend des aspects très divers, tels que l'histoire et les us et coutumes des locuteurs de la langue, ainsi que le contexte pédagogique, c'est-à-dire les ressources pédagogiques disponibles, la tradition d'enseignement de la langue, voire les croyances des locuteurs natifs sur leur propre langue. Ce contexte pédagogique peut prendre plus d'importance dans le contexte d'apprentissage de langues peu enseignées, s'il n'existe qu'un dictionnaire bilingue par exemple, l'apprentissage

sera forcément influencé par la façon dont le dictionnaire présente la langue. Gardant cette notion de contexte pédagogique à l'esprit, nous commencerons, dans une première partie, par exposer le contexte de la langue birmane et son apprentissage en tant que langue étrangère, avant de passer à un précis de la langue elle-même, à son système d'écriture et aux particularités de son informatisation.

Cette volonté de prendre en compte le contexte linguistique du birman se poursuivra dans le chapitre suivant sur la création de corpus. Par sa nature même, on peut considérer que, le travail sur corpus peut être considéré comme une recherche de contexte linguistique. Nous nous efforcerons de situer notre travail dans son contexte linguistique certes, mais aussi dans un contexte pédagogique. Nous tenons compte des travaux scientifiques antérieurs en linguistiques et en lexicologie dans la création de nos corpus authentiques (qui représentent des aspects de la langue en général), mais dans la préparation de ces corpus authentiques nous nous référons aussi à un corpus composé de ressources didactiques, pour tenir compte du point de vue des enseignants de la langue birmane.

Si notre démarche et son expérimentation est volontairement axée sur corpus, il nous semble qu'il est difficile de la caractériser uniquement comme *corpus-based approach* (approche basée sur corpus) ou *corpus-driven approach* (approche motivée par le corpus) selon la distinction de Tognini-Bonelli (2001). Selon cette définition, la première utilise les données du corpus pour valider ou raffiner une hypothèse, le deuxième infère des hypothèses à partir des données du corpus. La création de listes de fréquence détaillée dans la deuxième partie est clairement induite des corpus, et un certain nombre de nos questions de recherche surviennent du corpus uniquement. N'ayant pas l'intuition du locuteur natif, mais de l'apprenant de langue, de premier abord, cette approche motivée par le corpus paraît la plus sûre. Toutefois, nous utilisons aussi comme corpus du matériel didactique, des manuels et des sites web d'apprentissage du birman, qui reflètent les hypothèses sur la langue de leurs auteurs. Ce sont ces informations, combinées avec nos recherches sur la nature du vocabulaire de l'apprenant qui informent notre approche du traitement automatique de la segmentation du birman détaillée dans le dernier chapitre de la partie concernant la langue birmane.

1.3 *La lisibilité pour les langues peu enseignées et peu dotées*

Historiquement, les chercheurs en lisibilité ont eu recours à deux types de stratégies pour estimer la difficulté des textes. La première est l'élaboration de formules de lisibilité qui s'appuient sur des mesures de caractéristiques superficielles des textes, telles que les proportions de mots et de syllabes dans la phrase (comme l'indice FKGL[1]), la fréquence des mots polysyllabiques (comme l'indice SMOG[2]) ou la longueur de la phrase et le pourcentage de mots polysyllabiques (comme l'indice GFI[3]). La seconde, permise par les avancées en traitement automatique de langues et en apprentissage automatique, s'appuie sur des modèles statistiques de lisibilité plus complexes, basés sur des caractéristiques textuelles plus profondes et des corpus de textes de référence déjà classifiés par niveau de difficulté.

Ces deux approches posent plusieurs problèmes pour la lisibilité des langues peu enseignées comme langue étrangère. D'abord, la plupart des travaux sur la lisibilité mesurent la difficulté des textes pour des locuteurs natifs, et il ne va pas de soi que la lisibilité d'un texte pour un lecteur natif soit la même que pour un lecteur qui lit une langue étrangère (François 2011). Il faut donc développer de nouvelles formules, ou créer des modèles basés sur des corpus de textes destinés aux lecteurs apprenants de langue étrangère déjà classifiés par niveau de difficulté. Trouver des textes classés par niveau de difficulté n'est pas toujours possible et se heurte au problème du lecteur idéalisé, sans prendre en compte les compétences individuelles. Le deuxième problème est que ces méthodes, développées pour une langue spécifique (jusqu'alors, la recherche sur la lisibilité en langue étrangère s'est concentrée majoritairement sur l'anglais), ne sont pas nécessairement adaptées ou facilement adaptables à d'autres langues, bien que parfois une formule de lisibilité puisse donner des résultats satisfaisants pour des langues complètement différentes[4].

L'idée de mettre à profit les informations sur la fréquence lexicale est prometteuse, car la *méthode* ne serait pas spécifique à une langue, et ne demanderait pas un corpus classifié. L'autre avantage est qu'elle fait référence à la langue

1. FKGL = Flesch-Kincaid Grade Level (Kincaid et al. 1975)

2. SMOG = Simple Measure of Gobbledygook (McLaughlin 1969). Cette formule calcule la lisibilité pour un morceau de texte de trente phrases.

3. GFI = Gunning-Fog Index (Gunning 1952)

4. Das et Roychoudhury (2004) et Das et Roychoudhury (2006) (cités par Islam et al. (2012)) notent que l'indice FKGL donne des résultats satisfaisants pour le bengali, par exemple.

elle-même et non pas au lecteur ou aux jugements subjectifs sur la difficulté des textes qui sont utilisés pour créer les corpus déjà classifiés par niveaux de difficulté. Bien entendu le vocabulaire n'est pas le seul facteur qui contribue à la difficulté de lecture d'un texte. D'autres facteurs linguistiques comme la complexité syntaxique jouent un rôle important dans la compréhension d'un texte, mais il faudrait aussi considérer des facteurs extralinguistiques, tels que les connaissances préalables du lecteur, sa motivation, la densité d'information, la structure logique du texte y figurent également (Carrell 1987). Toutefois, l'usage d'un seul facteur, la fréquence lexicale, se justifie. Non seulement par l'importance de l'acquisition lexicale dans l'apprentissage d'une langue étrangère que nous avons évoquée, mais aussi par sa contribution importante à la difficulté. Selon Chall (1958), le niveau de difficulté du vocabulaire contribuerait à au moins 80% de la variabilité des indices de lisibilité. On peut dire que ce type de mesure de lisibilité est plus modeste, et ne cherche qu'à juger le vocabulaire d'un texte et rien d'autre, une lisibilité du texte en termes du vocabulaire. Il est bien possible qu'un même texte peut y avoir des lisibilités de différents niveaux selon que l'on juge la syntaxe, le vocabulaire, la cohésion textuelle et ainsi de suite. Il est particulièrement utile d'examiner le vocabulaire séparément des autres facteurs pour des lecteurs du birman langue étrangère plus avancés pour éventuellement pouvoir évaluer le vocabulaire de textes authentiques et recommander les textes qui contiennent le vocabulaire le plus utile à apprendre. L'utilisation de la fréquence du vocabulaire pour classifier les textes peut se considérer comme une mesure d'utilité de textes pour le lecteur et non seulement de difficulté. Comme l'a remarqué Leech (2011), le principe selon lequel le vocabulaire plus fréquent est plus important à apprendre ne peut guère être nié en tant que principe général, mais les apprenants ont tendance à surutiliser le vocabulaire à haute fréquence. Identifier le vocabulaire de fréquence légèrement moins fréquent, mais utile et trouver des textes pour apprendre en contexte peut aider à l'enrichissement du vocabulaire de l'apprenant.

1.4 *Objectifs*

Trois tâches donc constituent la base de notre travail : création de corpus représentatifs, élaboration de listes de vocabulaire de haute fréquence et l'implémentation d'une méthode de lisibilité basée uniquement sur la fréquence

lexicale, toutes les trois dans un contexte spécifique, l'apprentissage d'une langue étrangère peu enseignée, le birman. Les langues moins enseignées sont souvent des langues peu dotées en ressources pédagogiques et peu dotées en outils et ressources numériques (comme les corpus) pour un traitement informatique. Nous procéderons donc en soulevant les particularités rencontrées dans le traitement du birman, qui pourraient, nous l'espérons, s'avérer utiles pour le traitement d'autres langues peu enseignées et peu dotées informatiquement.

2 Plan de thèse

Cette présentation de notre travail est divisée en deux parties.

2.1 *La première partie : la création de corpus birman*

La première partie concerne les spécificités de la création de corpus en birman.

Le premier chapitre introduit la langue birmane elle-même, sa place dans la société birmane et le contexte de son apprentissage comme langue étrangère. Nous résumons ensuite les caractéristiques générales de la langue pertinentes pour notre travail, en particulier son système d'écriture. Suit une discussion sur l'informatisation du birman et les problèmes qu'elle a posés historiquement dans la création de polices de caractères et la mise en œuvre de la norme Unicode.

Le deuxième chapitre présente nos corpus, à la fois des corpus de textes authentiques et des corpus de textes didactiques utilisés pour prendre en compte le contexte éducatif dans la préparation de nos corpus pour le traitement automatique. La fin de ce chapitre détaille les prétraitements plus simples nécessaires pour éviter d'introduire des erreurs dans le corpus.

Le troisième chapitre concerne la segmentation de textes birmans en tokens. Nous commençons par un état de l'art de la segmentation automatique du birman et une introduction à notre outil de segmentation suivie d'une évaluation des performances de l'outil et l'utilisation de l'outil pour la correction de corpus. La deuxième partie de ce chapitre concerne l'épineuse question de la définition des tokens pour le traitement automatique, toujours appelés *words* par les informaticiens birmans. Nous exposons les aspects de segmentation qui peuvent s'appliquer à tout type de traitement statistique et automatique du birman, comme la définition des frontières des entités nommées et le rôle de l'espace typographique, mais nous regardons aussi comment les chercheurs en

linguistique, en lexicographie et les auteurs de manuels didactiques ont abordé ce problème. Enfin, nous détaillons les stratégies utilisant les fréquences des n-grammes de syllabes pour identifier les « mots ».

2.2 *La deuxième partie : la lisibilité par la fréquence lexicale du birman langue étrangère*

La deuxième partie porte sur la méthode utilisée pour élaborer les listes de fréquence lexicale et l'usage d'une liste générale pour classer les textes birmans par ordre de difficulté, utilisant une méthode de lisibilité.

Le quatrième chapitre concerne la fréquence lexicale. Après un bref état de l'art, nous détaillons les méthodes que nous avons explorées pour prendre en compte la taille relativement faible de nos corpus authentiques, prenant en compte la fréquence et la dispersion des types dans les corpus. Ces démarches nous ont permis de créer une liste de fréquence lexicale globale à utiliser pour classer des textes par la suite. La fin du chapitre donne un aperçu de la capacité de la liste lexicale à couvrir le vocabulaire d'un petit corpus et une comparaison entre notre liste globale et le vocabulaire des manuels de notre corpus didactique.

Le cinquième et dernier chapitre concerne la classification de textes birmans utilisant les données de fréquence lexicale. La première partie explique la difficulté rencontrée dans l'utilisation de la plupart des méthodes existantes de lisibilité pour les langues peu dotées et peu enseignées. Ensuite, nous expliquons la méthode que nous avons utilisée et la petite modification que nous avons effectuée pour tenir compte de la petite quantité de données dont nous disposons : le tri par ordre de difficulté utilisant l'algorithme de tri par insertion binaire et l'entraînement d'une machine à vecteurs de support sur deux corpus représentatifs de textes faciles et difficiles pour réaliser le tri. Nous démontrons l'efficacité de cette méthode sur un petit corpus de textes didactiques.

3 Conventions

Bien que le terme *Myanmar* soit la dénomination officielle, pour désigner le pays, le terme *Birmanie* reste tant dans l'usage de tous les jours que dans la recherche scientifique francophone. Nous employons aussi *birman* pour la langue et comme racine de dérivation pour les adjectifs (*la langue birmane*) et les substantifs désignant les personnes (*un Birman*, une personne de nationalité

birmane). Les chercheurs birmans qui communiquent en anglais utilisent souvent *Myanmar* pour le pays, mais aussi et *myanmar* pour la langue (calqué sur မြန်မာဘာသာ /mjàɴmàbàθà/ *myanmar + langue*, *la langue birmane*), alors que les chercheurs qui ne sont pas des Birmans utilisent souvent *Burma* et *burmese*. Il faut donc considérer ces termes comme des synonymes.

Le code à trois lettres établi par l'Organisation internationale de normalisation pour représenter le nom de la langue birmane est `mya` pour le code ISO 639-3, et `mya` ou `bur` pour le code ISO 639-2, `mya` étant le code terminologique et `bur` le code bibliographique[5]. Le code ISO 639-1 à deux lettres `my` est utilisé dans les adresses et noms de sites web.

Les termes en gris clair renvoient au glossaire.

3.1 *Translittération orthographique et transcription phonémique*

Quand il s'agit d'exemples en birman, nous utiliserons les systèmes de translittération et transcription disponibles en ligne sur le site Burmese Character Picker[6]. La translittération orthographique qui illustrent le système d'écriture seront données entre chevrons ⟨ ⟩, alors que la transcription phonémique en API[7] sera indiquée par des barres obliques //[8]. Par exemple : မြန်မာဘာသာ, qui veut dire *la langue birmane*, s'écrit avec des lettres équivalentes ⟨mǰnˣmaba̰θa⟩ et se prononce approximativement /mjàɴmàbàθà/. Le tableau A.1 qui détaille le système de translittération est donné en annexe. Les noms propres birmans sont souvent écrits en lettres latines selon l'usage en anglais pour faciliter la lecture. La prononciation thaïlandaise est transcrite de manière similaire à celle du birman. La prononciation chinoise est donnée en hànyǔ pīnyīn et celle du japonais utilise le système Hepburn.

Sauf indication contraire, la langue d'origine des exemples est le birman. Les équivalents français donnés dans les exemples ne constituent pas une traduction exhaustive.

5. http://www.loc.gov/standards/iso639-2/php/code_list.php consulté le 13 juin 2019.
6. https://r12a.github.io/pickers/mymr-my/
7. API = Alphabet phonétique international
8. La transcription est donnée à titre indicatif et ne reflète pas nécessairement la vraie prononciation, car elle ne prend pas en compte ni les exceptions ni les modifications dans la prononciation des syllabes liées, le sandhi, (Bernot, Cardinaud et al. 2001) qui ne sont pas indiquées dans la forme écrite de la langue.

Première partie

La création de corpus birman

Chapitre 1

La langue birmane

Nous introduisons dans ce chapitre le contexte linguistique du birman d'abord, portant sur son enseignement et ses caractéristiques de base, suivi d'une explication plus détaillée de son système d'écriture et de son informatisation, car le traitement automatique concerne exclusivement l'analyse de la représentation écrite de la langue, même s'il s'agit de la langue orale transcrite. Une compréhension de ces aspects est indispensable lors de la création de corpus birman, pour éviter des erreurs lors de l'analyse statistique.

1.1 La langue birmane dans la société birmane

La langue birmane မြန်မာဘာသာ /mjàɴmàbàθà/, appelée officiellement le *Myanmar* en anglais, est la langue officielle de la République de L'Union de Myanmar (figure 1.1), officiellement la seule langue de l'administration, des médias et de l'éducation (Lo Bianco 2016)[1].

Le développement et la préservation du birman sont assurés par la *Myanmar Language Commission* ou MLC, qui fait partie du Département des langues ethniques du Myanmar[2] sous le ministère de l'éducation. La MLC publie des dictionnaires, un guide d'orthographe et une grammaire de la langue. La commission est aussi à l'origine du système de translittération MLCTS[3].

Selon *Ethnologue* (Eberhard et al. 2019), en 2000 la langue comptait quelque 32 millions de locuteurs natifs en Birmanie, soit plus des trois quarts de la population. A l'aide des chiffres fournis par les Nations Unies sur la population du

1. Depuis 2014, l'importance des 135 langues minoritaires surtout dans le domaine éducatif a été reconsidérée à travers le développement d'une politique nationale linguistique (Lo Bianco 2016).
2. http://www.dmnl.gov.mm/?lang=my, en birman
3. Myanmar Language Commission Transcription System

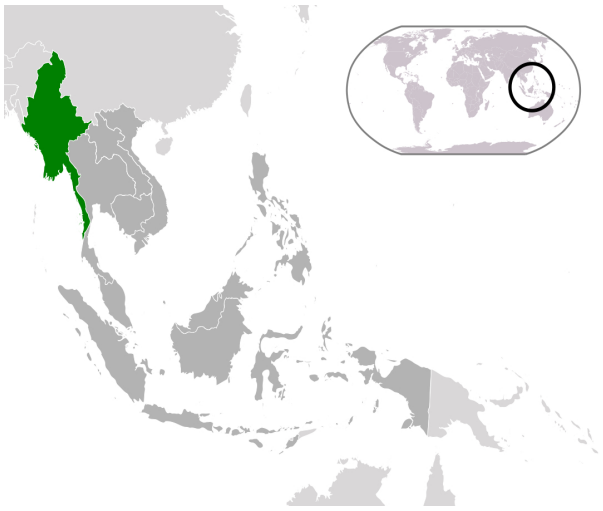


FIG. 1.1 : La Birmanie (Myanmar) en Asie. Source : Wikimedia Commons

pays[4], le nombre de locuteurs natifs actuels peut s'estimer à plus de 37 millions. Il faudrait ajouter à ce chiffre le nombre de Birmans d'outre-mer (environ 2 millions au total, dont 1,4 millions en Thaïlande selon le recensement officiel de 2014 (Department of Population 2015)), mais le nombre de locuteurs de birman est difficile à estimer, car ces réfugiés de nationalité birmane sont souvent membres des minorités ethniques du Myanmar et ne sont pas forcément de langue maternelle birmane[5]. Langue officielle et langue de l'éducation nationale, le birman est pratiqué comme deuxième langue par la plupart de membres des minorités ethniques dans le pays. *Ethnologue* (Eberhard et al. 2019) estime le nombre de locuteurs de birman comme deuxième langue à environ dix millions de personnes.

Nous faisons une distinction entre l'apprentissage du birman comme *deuxième langue* (L2) et l'apprentissage du birman comme *langue étrangère* (BLE) (Gass et Selinker 2000). Dans les deux cas, il s'agit d'une langue apprise après la langue maternelle (L1), mais dans le cas du premier, l'apprentissage se fait dans un contexte birmanophone, pas nécessairement dans un environnement d'appren-

4. Selon les estimations démographiques de la division Population des Nations Unies (United Nations, 2017), la population totale du pays était de 46 095 462 de personnes en 2000, et de 53 855 735 en 2018.

5. Des estimations chiffrées concernant l'étendu de la diaspora du Myanmar en Asie se trouvent dans Egreteau (2012).

tissage formel, alors que dans le cas d'apprentissage d'une langue étrangère, il s'agit d'un processus d'apprentissage du birman à l'âge adulte et dans un contexte plus formel, souvent dans une salle de classe avec un formateur de langue. Comme le terme implique, l'apprentissage d'une langue étrangère se passe souvent, mais pas forcément exclusivement, dans l'environnement de la langue maternelle de l'apprenant (un étudiant français qui apprend le birman en France, par exemple), avec un accès restreint aux locuteurs natifs. C'est l'apprentissage du birman comme langue étrangère qui nous intéresse, tout en étant conscient des croisements possibles entre les types d'apprentissages.

1.2 Le birman comme langue étrangère

Les premiers étrangers à s'intéresser à la langue birmane et sa description furent surtout des missionnaires, à commencer par la grammaire du missionnaire anglais Felix Carey (1814). Les premiers étudiants étrangers de renom furent les missionnaires protestants américains Adoniram Judson (1788–1850) et Ann Hasseltine Judson (1789–1826). En plus de leurs traductions de textes chrétiens (du catéchisme et des livres de la Bible), Adoniram Judson a compilé un dictionnaire birman-anglais (Judson 1826 ; Judson 1852) et un précis de grammaire de la langue (Judson 1842), inclus dans les éditions ultérieures de son dictionnaire. Au fil des guerres anglo-birmanes du XIX^{e} siècle et de la colonisation britannique qui s'ensuivit, l'intérêt pour l'apprentissage du birman devient plus prosaïque, et surtout pratique. Membre de l'armée de la Compagnie britannique des Indes orientales au Bengale, Thomas Latter publie sa grammaire en 1845 (Latter 1845). Au tournant du siècle apparaissent des manuels destinés aux fonctionnaires et marchands de la colonie (St John 1894), dont celui de l'archéologue sino-birman Taw Sein Ko (1898). Ce premier livre de St John constitue le premier *Reader* ou livre de lecture en birman langue étrangère. Son deuxième livre, *Burmese self-taught* (St John 1911), est constitué de listes thématiques de vocabulaire, suivi d'un précis de grammaire et des phrases de conversation courante pour voyageurs. Les premiers cours formels en birman en Grande-Bretagne eurent lieu en 1917 à la School of Oriental Studies (qui devient plus tard la School of Oriental and African Studies, SOAS).

Après l'indépendance du pays en 1948, les raisons pour apprendre la langue se diversifièrent, avec un intérêt particulier pour l'ethnologie, ce qui motiva la

création de l'enseignement du birman à l'Inalco en 1960 par Denise Bernot, auteure du Dictionnaire birman-français en quinze volumes (Bernot 1978-1988 ; Bernot 1988-1992).

En 1974 les premiers cours de birman pour étrangers sont ouverts à l'*Institute of Foreign Languages* (IFL) à Yangon, institut qui deviendra en 1996 la *University of Foreign Languages, Yangon*, YUFL, နိုင်ငံခြားဘာသာတက္ကသိုလ် (ရန်ကုန်) /nàɪɴŋàɴtɕʰábàθàtɛʔkə̰θò (jàɴkòʊɴ)/. Selon l'agence de presse Xinhuanet (2017), en 2017, 1727 étudiants étrangers avaient suivi le cursus à YUFL depuis son inauguration, et cette même année-là 243 étudiants étaient inscrits pour apprendre le birman en Birmanie, soit à YUFL, soit à la *University of Foreign Languages, Mandalay* (MUFL) မန္တလေ နိုင်ငံခြားဘာသာတက္ကသိုလ် /màɴtə̰lè nàɪɴŋàɴtɕʰábàθàtɛʔkə̰θò/. Des formations plus courtes sont aussi proposées en Birmanie, comme celles de l'Institut français de Birmanie ou les cours annuels *Bamazaga*[6] assurés par des professeurs de programmes de BLE basés à l'étranger.

De nos jours des programmes de birman langue étrangère existent dans de nombreux autres pays, notamment aux États-Unis, au Japon, et plus récemment en Thaïlande[7]. Le cas de la Thaïlande est particulièrement intéressant, exposé en détail dans une monographie sur l'enseignement du birman en Thaïlande de Chinnak (2015). L'auteur explique que, à commencer à la fin du XIX^{e} siècle et tout au long du siècle dernier, une religion commune, le bouddhisme theravāda, a favorisé les échanges de maîtres spirituels et de professeurs entre temples en Thaïlande et en Birmanie, surtout dans le nord de la Thaïlande, dans la région de l'ancien royaume Lanna. Ces échanges ont donné lieu à l'enseignement non seulement de la langue sacrée, le pâli, de textes sacrés et de la méditation, mais aussi de la langue birmane. Si la mise en place de programmes universitaires est relativement récente (l'Université de Chiangmai offre un mineur de birman depuis 1997, et l'Université de Naresuan propose un diplôme national de birman depuis 2001), la coopération mutuelle croissante entre les pays de l'Asie du Sud-est promue par l'ASEAN[8] a favorisé l'intérêt pour les langues voisines. Si le birman est ainsi appris comme langue étrangère aussi bien pour des raisons économiques que culturelles, par des étudiants, il y a aussi un intérêt

6. http://lukecorbin.org/bamazaga/

7. Une liste de centres de formation et de ressources pédagogiques est tenue sur le site de *Bamazaga* https://lukecorbin.org/bamazaga/study-resources/

8. *Association des nations de l'Asie du Sud-Est* est le plus souvent nommée par son sigle anglais ASEAN, même en birman et en thaï.

professionnel pour la langue au sein des entreprises, pour profiter du marché birman, ou pour faciliter la communication avec des ouvriers immigrés. Pour améliorer les perspectives des enfants réfugiés de Birmanie en Thaïlande, certaines organisations non gouvernementales proposent des cours de birman à ceux dont la langue maternelle est une langue minoritaire de Birmanie. Le grand nombre de Birmans résidant en Thaïlande a suscité l'intérêt pour l'apprentissage du birman dans d'autres domaines : la langue est apprise depuis longtemps par des militaires, mais de plus en plus de stages de langues sont proposés aux policiers et aux personnels médicaux, surtout dans des régions frontalières.

En ce qui concerne les ressources pédagogiques, du fait de la colonisation britannique de la Birmanie, la plupart de ressources du birman actuellement disponibles pour l'apprenant du birman ont été destinés aux apprenants anglophones. Les manuels d'apprentissage de birman langue étrangère utilisés par la plupart des étudiants étrangers anglophones sont *Colloquial Burmese* (Hnin Tun et McCormick 2015) et les quatre tomes de Okell, U Saw Tun et al. (2010c) – *Burmese : An Introduction to the Spoken Language* en deux tomes, *Burmese : An Introduction to the Literary Style* et *Burmese : An Introduction to the Script. Burmese/Myanmar Dictionary of Grammatical Forms* (Okell et Allott 2017), aussi destiné aux apprenants de BLE. Le choix a été élargi plus récemment par la publication de *Advancing in Burmese : A Drill Book for Intermediate to Advanced Learners* (Yadana Aung 2020) et *Burmese : a cultural approach* (Keeler et Lyan 2021). Pour les francophones, il y a le *Manuel du birman 1* (Bernot, Cardinaud et al. 2001) et son deuxième tome, une grammaire (Bernot, Cardinaud et al. 2010), un dictionnaire de poche (Bernot, Cramerotti et al. 1998) et un grand dictionnaire birman-français en quinze tomes (Bernot 1978-1988 ; Bernot 1988-1992). Le *Myanmar-English Dictionary* မြန်မာ–အင်္ဂလိပ် အဘိဓာန် (abrégé MED) publié par la Myanmar Language Commission en 1993 est devenu le dictionnaire de référence pour les étudiants. Largement piraté, ses entrées se trouvent sur le Wiktionaire birman[9] et sont reproduites avec permission par le projet *Sealang* (Southeast Asian Languages Library 2006).

Bien que les effectifs des étudiants soient peu nombreux, il existe des ressources en ligne en libre accès pour apprendre le birman. *SEASite Burmese (Myanmar)*, développé depuis 1997 par le Center for Southeast Asian Studies à Northern

9. https://my.wiktionary.org

Illinois University, contient des textes et exercices, ainsi que des textes bilingues, destinés aux étudiants de niveaux débutant et intermédiaire. Un *Kit de survie* birman du Centre des langues étrangères de l'institut de langues de la Défense des États-Unis destiné à faciliter l'acquisition de vocabulaire spécifique aux forces armées est aussi disponible librement sur l'internet[10]. Le vocabulaire des cours est en effet très spécifique[11] et probablement d'un intérêt limité pour la plupart des apprenants. Un site web de l'*University of Foreign Studies* au Japon propose des ressources audiovisuelles *TUFS Language Modules*[12], qui comprend des dialogues et un glossaire en ligne. A part les groupes Facebook destinés aux apprenants[13], il existe aussi quelques sites de passionnés pour l'apprentissage du birman, mais la plupart des contenus sont destinés aux grands débutants[14]. Le projet *Sealang* mentionné ci-dessus, en dehors du dictionnaire (interrogeable dans les deux sens), propose une base de phrases alignées anglais-birman qui permet à l'utilisateur de trouver des mots ou des phrases en contexte bilingue. Riche du Wikipédia birman et d'autres sources, *BabelNet* (Navigli et Ponzetto 2012) est devenu un outil de référence très utile non seulement comme dictionnaire, mais aussi comme dictionnaire de noms propres.

Faute de système officiel, une association birmano-japonaise a instauré un test de compétence linguistique en birman langue étrangère appelé le Myanmar Language Test ou MLT. Ce test propose cinq niveaux de difficulté.

1.3 Caractéristiques générales de la langue birmane

Le birman contemporain est une langue du groupe *lolo-birman*, lui-même faisant partie de la branche *tibéto-birmane* de la famille *sino-tibétaine* (Dryer et WALS author team 2011). L'agrégateur d'informations scientifiques sur les langues du monde *Glottolog* (Hammarström et al. 2021) propose une classification plus nuancée, où le birman moderne se placerait dans une branche du *burmish du sud*, les langues *lolo-birman* faisant partie d'un groupe de langues sino-tibétaines

10. https://www.dliflc.edu/elearning/

11. Les cours sont intitulés *Guide linguistique de l'équipage aérien*, *Guide linguistique de base*, *Affaires civiles*, *Langage médicale*, *Langage de la marine* et *Affaires publiques*.

12. http://www.coelang.tufs.ac.jp/mt/my/dmod/

13. Par exemple, Burmese/Myanmar Language Learning, Daily Burmese et ဖတ်လို့ ရလား Can you read it in Burmese ?

14. Comme les sites Asia Pearl Travels Learn Myanmar et Bama Learn Burmese

orientales, appelé *burmo-qiangic*, détails ci-dessous figure 1.2 (nombre de langues entre parenthèses).

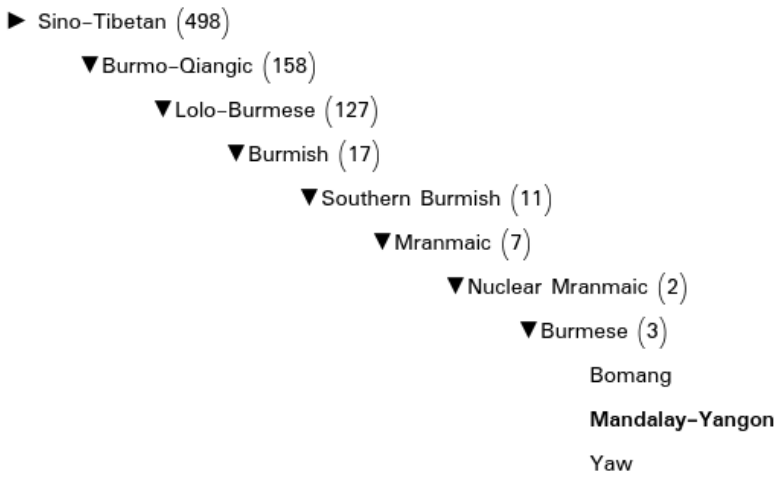


FIG. 1.2 : Classification du birman moderne, le dialecte Mandalay-Yangon. Source : *Glottolog* (Hammarström et al. 2021)

C'est une langue à tons, considérée par les chercheurs comme fondamentalement monosyllabique (Bernot 1980 ; Hnin Tun 2013 ; Myint Soe 1999 ; Brac de La Perrière 1999), avec néanmoins de nombreuses unités lexicales polysyllabiques, soit d'origine birmane, soit des emprunts polysyllabiques de langues indo-européennes, essentiellement du pâli et de l'anglais (Bernot 1980). En dehors des emprunts à ces langues, la langue d'influence religieuse et du colonisateur respectivement, le lexique du birman emprunte des mots à ses langues régionales (le môn et le shan, par exemple) et à ses langues voisines (comme le thaï et le malais), mais aussi à des langues d'origine plus lointaines comme le sanskrit et le hindi (Jenny 2015). La structure de la phrase est caractérisée par un groupe verbal positionné en fin de l'énoncé précédé de ses arguments (Okell 1969 ; Wheatley 1982 ; Stewart 1955), qui n'ont pas d'ordre fixe et qui (surtout en langue parlée) peuvent être facultatifs. Des particules post-positionnées servent de marqueurs syntaxiques, d'autres particules jouent un rôle discursif ou modal.

La langue possède deux registres de langue, appelés *style parlé* et *style écrit* qui sont associés au niveau de langue et genre du discours. Les deux styles se caractérisent principalement par des différences de vocabulaire, notamment

parmi les unités grammaticales, par exemple la conjonction de coordination နှင့် /n̥ɪɴ̃/ (et/avec) en style littéraire est နဲ့ /nɛ̰̃/ en style parlé. Hopple (2003) estime que les trois quarts de la différence entre les deux styles peuvent s'attribuer à l'emploi de particules lexicales et grammaticales, et que la pérennité du style écrit s'explique par un cadre grammatical sous-jacent commun aux deux styles. Malgré leurs noms, les deux styles ne se limitent pas exclusivement à l'orale ou à l'écrit ; le style écrit est utilisé dans des situations plus formelles et officielles, y compris dans les discours officiels, la littérature et les journaux et sur le Wikipédia birman, alors que le style parlé, qui reflète la langue orale de tous les jours, s'emploie à la télévision et à l'écrit sur certains sites web d'actualités. Bien que la différence entre les deux styles (souvent bien plus marquée que dans d'autres langues) soit particulièrement surprenante pour l'apprenant de la langue (Hnin Tun 2006), Allott et al. (1989) notent que la distinction entre ces styles n'est pas stricte, avec un échange de vocabulaire entre les deux. Nous espérons en tirer plus d'informations sur la différence de fréquence d'usage du vocabulaire entre les deux styles lors de nos expériences.

1.4 Le système d'écriture du birman

Attesté depuis le XII^e^ siècle[15], le système d'écriture du birman, tout comme les autres systèmes d'écriture de l'Asie du Sud-est dérivés de l'écriture indienne *brāhmī*, est un alphasyllabaire ou *abugida*. Des variantes de ce système d'écriture, complétées avec des signes complémentaires, sont utilisées pour écrire d'autres langues de la Birmanie telles que le shan, les langues karènes, le môn ou le pâli. Le système est une scriptio continua, qui ne sépare pas explicitement les unités lexicales par des espaces typographiques.

Le texte écrit se lit de gauche à droite, syllabe par syllabe. L'intérieur de la syllabe ne se lit pas dans une seule direction. Selon le principe de l'abugida, on commence à lire la syllabe par sa consonne initiale, les signes accompagnant la consonne initiale formant une syllabe peuvent se placer à gauche, à droite, en dessous, au-dessus ou autour de cette consonne.

Les 33 signes consonantiques de base sont dotés d'un son vocalique inhérent,

15. Le document en langue birmane le plus ancien attesté serait l'inscription lapidaire quadrilingue Myazedi မြစေတီ ကျောက်စာ /mjaˍsètì tɕaʊʔsà/ de l'an 1113 de n. è. environ (Myanmar Language Commission 1993).

FIG. 1.3 : Manuscrit du Maha Niddesa, un texte canonique du bouddhisme, qui montre trois styles d'écriture birmane : anguleux (haut), arrondi (centre), arrondi à contours (bas) en laque rouge. Source : Wellcome Collection. CC BY

le /a̰/ bref[16]. Ce sont des signes indépendants qui peuvent former une syllabe sans l'ajout d'un autre signe. Ici le signe consonantique ∞ seul forme une syllabe complète prononcée /la̰/ :

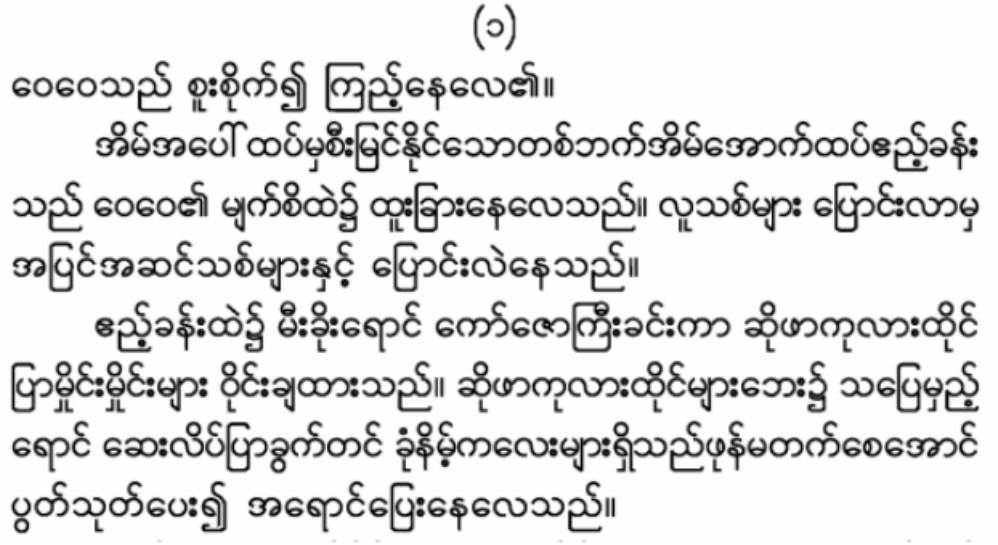
(1) ∞ ⟨l⟩ /la̰/ *mois, lune*

D'autres signes remplacent, modifient ou suppriment le /a̰/ bref inhérent. En voici des exemples pour des syllabes qui terminent par une voyelle ouverte. Ces exemples illustrent l'emplacement des voyelles par rapport à la consonne initiale : à gauche, au-dessus, à droite, autour, en dessous.

(2) ∞ ⟨l⟩ /la̰/ + ɛ◌[17] ⟨e⟩ ⇒ ɛ∞ ⟨le⟩ /lè/ *vent*

16. Rappelons que les caractères entre chevrons, ⟨ et ⟩, représentent les éléments du système d'écriture, la translittération de l'écriture birmane, qui permet de comprendre l'ordre de stockage en mémoire physique des caractères, alors que les caractères entre barres obliques, / et /, représentent approximativement la prononciation, la transcription phonémique. Le tableau A.1 qui détaille le système de translittération est donné en annexe.

17. Par convention, en Unicode le caractère du rond en pointillé ◌ (U+25CC DOTTED CIRCLE)


(၁)

ဝေဝေသည် စူးစိုက်၍ ကြည့်နေလေ၏။

အိမ်အပေါ် ထပ်မှစီးမြင်နိုင်သောတစ်ဘက်အိမ်အောက်ထပ်ဧည့်ခန်းသည် ဝေဝေ၏ မျက်စိထဲ၌ ထူးခြားနေလေသည်။ လူသစ်များ ပြောင်းလာမှ အပြင်အဆင်သစ်များနှင့် ပြောင်းလဲနေသည်။

ဧည့်ခန်းထဲ၌ မီးခိုးရောင် ကော်ဇောကြီးခင်းကာ ဆိုဖာကုလားထိုင် ပြာမှိုင်းမှိုင်းများ ဝိုင်းချထားသည်။ ဆိုဖာကုလားထိုင်များဘေး၌ သပြေမှည့်ရောင် ဆေးလိပ်ပြာခွက်တင် ခုံနိမ့်ကလေးများရှိသည်ဖုန်မတက်စေအောင် ပွတ်သုတ်ပေး၍ အရောင်ပြေးနေလေသည်။

FIG. 1.4 : Exemple de texte en birman contemporain, un extrait du roman de Journal-Gyaw Ma Ma Ley, မုန်း၍မဟူ /mó̰ jwémə̰hù/, traduit en français comme *La mal-aimée*.

(3) လ ⟨l⟩ /lə̰/ + ◌ဲ ⟨ɛ⟩ ⇒ လဲ ⟨lɛ⟩ /lɛ́/ *échanger*

(4) လ ⟨l⟩ /lə̰/ + ◌ာ ⟨a⟩ ⇒ လာ ⟨la⟩ /là/ *venir*

(5) တ ⟨t⟩ /tə̰/ + ◌ော ⟨ea⟩ ⇒ တော ⟨tea⟩ /tɔ́/ *forêt*

(6) ဂ ⟨g⟩ /gə̰/ + ◌ု ⟨ù⟩ ⇒ ဂု ⟨gù⟩ /gù/ *grotte*

En voici un exemple de l'expression des tons en écriture avec la même consonne initiale, mais trois tons différents prononcés /lə̰/, /là/ et /lá/ :

(7) လ ⟨l⟩ /lə̰/ = *mois, lune*

(8) လ ⟨l⟩ /lə̰/ + ◌ာ ⟨a⟩ ⇒ လာ ⟨la⟩ /là/ *venir*

(9) လ ⟨l⟩ /lə̰/ + ◌ာ ⟨a⟩ + ◌း ⟨²⟩ ⇒ လား ⟨la²⟩ /lá/ *marqueur interrogatif*

Notons en passant que pour plus de clarté, parfois le son /à/ s'écrit ◌ါ au lieu de ◌ာ.

est utilisé pour illustrer l'effet d'une *marque de combinaison* (combination mark) telle que les caractères diacritiques, ou comme ici les caractères vocaliques dépendants sur un caractère porteur, en général un caractère consonantique.

Pour indiquer les groupes de consonnes à l'initial, le système d'écriture utilise divers procédés de ligature. Ainsi, certaines signes consonantiques possèdent une forme dépendante pour indiquer que la consonne est une consonne médiale (sans voyelle dépendante propre) qui s'attache à la consonne initiale de la syllabe. Par exemple, la forme dépendante du caractère ဟ ⟨h⟩ est ◌ှ ⟨ɦ⟩ s'utilise ainsi :

(10) လ ⟨l⟩ /la̰/ + ◌ှ ⟨ɦ⟩ ⇒ လှ ⟨lɦ̆⟩ /l̰a̰/ *joli*

Les formes dépendantes de ရ ⟨ɹ̇⟩ et ဝ ⟨w⟩ sont ◌ြ ⟨j̆⟩ et ◌ွ ⟨w̆⟩ respectivement.

(11) မ ⟨m⟩ + ◌ြ ⟨j̆⟩ + ◌ွ ⟨w̆⟩ + ◌ေ ⟨e⟩ ⇒ ◌ေွြ ⟨mj̆w̆e⟩ /mjwè/ *serpent*

Les syllabes qui se terminent par un coup de glotte ou une nasalisation de la voyelle voient la voyelle inhérente du signe consonantique final supprimée par le signe asat ◌် ⟨ˣ⟩, comme ceci :

(12) ဆ ⟨s̈ʰ⟩ + န ⟨n⟩ /na̰/ + ◌် ⟨ˣ⟩ ⇒ ဆန်် ⟨s̈ʰnˣ⟩ /sʰàɴ/ *riz cru*

Certaines syllabes s'écrivent au moyen de combinaisons de ces procédés.

(13) က ⟨k⟩ + ◌ျ ⟨y̆⟩ + ◌ောင် ⟨eaŋˣ⟩ + ◌း ⟨ˀ²⟩ ⇒ ◌ကျောင်း ⟨ky̆eaŋˣˀ²⟩ /tɕáʊɴ/ *monastère, école*

(14) က ⟨k⟩ +◌ိုက် ⟨iukˣ⟩ ⇒ ကိုက် ⟨kiukˣ⟩ /kaɪʔ/ *mordre*

Les emprunts utilisent des consonnes souscrites, qui parfois changent de forme ou d'orientation :

(15) တက္ကသိုလ် ⟨tk̆kθiulˣ⟩ /tɛʔka̰θò/ *université* (la lettre က ⟨k⟩ souscrite)

(16) ပဏ္ဍိတ ⟨pn̈dit⟩ /pàɴdḭta̰/ *un érudit* (la lettre ဍ ⟨d̈⟩ change d'orientation)

(17) နိဿယ ⟨niθ:y⟩ /nḭθa̰ja̰/ *texte pâli glosé* (ဿ ⟨θ⟩ + သ ⟨θ⟩ = ဿ ⟨θ:⟩)

Chaque voyelle possède une forme dépendante et une forme indépendante qui forme une syllabe indépendante. En voici des exemples :

En Unicode ဘ ⟨ʔ⟩ /ʔa̰/ est classé parmi les signes vocaliques, mais étant le caractère qui représente le coup de glotte suivi de /a̰/, ce caractère est souvent traité comme un signe consonantique.

En dehors de ces signes, il y a des symboles logographiques comme ၍ ⟨(sub)⟩ /jwé/ et ၌ ⟨(loc)⟩ /n̰aiʔ/ (en l'occurrence des marqueurs grammaticaux dans la langue écrite pour l'accompli et le locatif) et des chiffres. Il y a également

Forme dépendante	Forme indépendante	Prononciation
◌ု ⟨u⟩	ဥ ⟨u̯⟩	/ʔu̯/
◌ူ ⟨ù⟩	ဦ ⟨ù⟩	/ʔù/
◌ိ ⟨i⟩	ဣ ⟨i̯⟩	/ʔi̯/
◌ော် ⟨ea^{x}⟩	ဧ ⟨ɔ̰⟩	/ʔɔ̀/

TAB. 1.1 : Exemples de formes dépendantes et indépendantes vocaliques

deux signes de ponctuation : une barre verticale simple, ।, qui correspond approximativement à la virgule en français, et une barre verticale double, ॥, qui correspond au point. Les chiffres sont souvent entourés de parenthèses. L'ensemble des systèmes d'écriture comprenant le système d'écriture du birman et ses variantes est appelé *Myanmar* dans la documentation Unicode. Il faut noter également qu'en typographie birmane, il n'y a pas d'équivalent de la différence entre minuscules et majuscules.

1.5 L'informatisation du système d'écriture du birman

Nous faisons ici un bref exposé du fonctionnement du système Unicode pour le birman, avant d'éclaircir les différences entre le système Unicode standard et d'autres systèmes d'encodages du birman, à savoir, les systèmes pré-Unicode et les systèmes basés sur Unicode de façon non standard. Nous expliquerons les défauts de ces deux derniers, qui démontrent pourquoi il est essentiel de convertir les textes en Unicode dans le processus de prétraitement du corpus. Afin de poser les bases pour notre exposé de l'informatisation du birman, nous commençons par des remarques générales sur l'encodage de caractères et le standard Unicode.

1.5.1 *Remarques générales sur l'encodage de caractères*

Dans un ordinateur, tout caractère, qu'il soit une lettre d'un alphabet, un syllabogramme d'un syllabaire ou alphasyllabaire, un idéogramme ou un sinogramme, une marque de ponctuation, même invisible comme une espace typographique, un chiffre ou tout autre symbole, est représenté comme un nombre, appelé point de code. C'est cette représentation de caractères par des nombres qui permet l'encodage des écritures d'un nombre croissant de langues humaines nécessaire

Glyphes			Point de code	
Verdana	Libertine	Lobster Two	décimal	hexadécimal
A **A** *A* ***A***	A A *A* ***A***	A **A** *A* ***A***	65	41
a **a** *a* ***a***	a a *a* ***a***	a **a** *a* ***a***	97	61
K **K** *K* ***K***	K K *K* ***K***	K **K** *K* ***K***	75	4B

TAB. 1.2 : Les glyphes d'un même caractère, de polices différentes, ont le même point de code en italiques et en gras, mais des points de codes différents pour les majuscules et les minuscules

à leurs divers usages écrits, leur conservation et leur manipulation informatique. Les points de code du standard Unicode (abordé brièvement ci-après) peuvent être stockés physiquement en machine selon divers formats d'encodage, UTF-8, UTF-16 ou UTF-32, l'UTF-8 étant le format le plus répandu. Toutefois, l'encodage des caractères en UTF-8 ne garantit pas, comme nous verrons plus loin, leur conformité avec le standard Unicode.

La police de caractères permet le décodage et une visualisation correcte des caractères d'un document, à l'écran ou sur papier. Le caractère apparié au point de code n'est en fait qu'une abstraction[18] : ses différentes représentations graphiques, ses glyphes, réfèrent au même caractère, avec le même point de code. Ce faisant, les différences dans la représentation visuelle, telles que le corps (la taille), la graisse (l'épaisseur du trait), l'inclinaison (la distinction romain/italique), sont déterminées uniquement par les divers styles d'affichage d'un document. Par exemple, ‹ g › et ‹ g › sont deux glyphes du même caractère ‹ g ›. Les majuscules et minuscules sont par contre considérées comme des caractères distincts et donc encodés avec des points de code différents. Voir tableau 1.2 pour des exemples de glyphes d'un même caractère, et la figure 1.5 pour une description schématique de la relation entre les points de code et les polices.

18. La notion abstraite de caractère est appelée *grapheme* dans la documentation Unicode, mais puisque *graphème* en français est plutôt utilisé pour indiquer la transcription d'un phonème, par souci de clarté, nous employons le terme *caractère*.

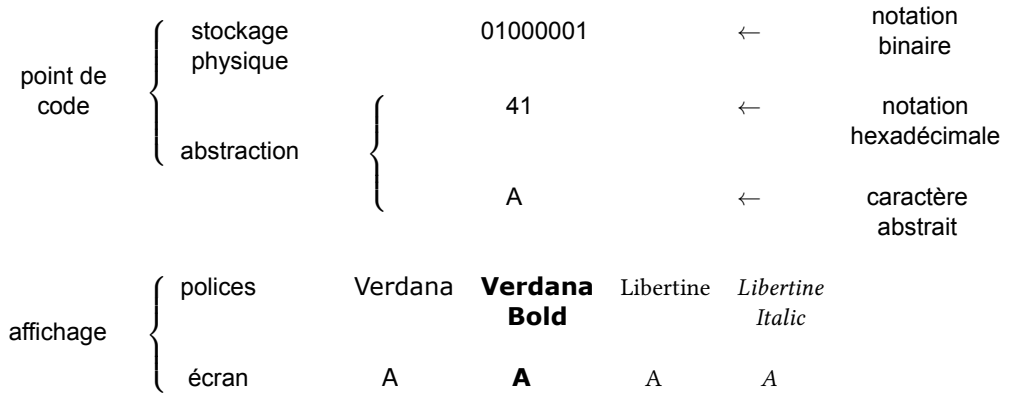


FIG. 1.5 : Relation schématique entre le point de code d'un caractère et son affichage avec des polices différentes

1.5.2 *Le standard Unicode*

Actuellement à la version 15.0, encodant 149 186 caractères, Unicode est devenu le standard international de référence pour le codage numérique des caractères[19]. Le principe sous-jacent est d'attribuer un nom unique et un point de code unique pour chaque caractère de chacun des systèmes d'écriture de toutes les langues naturelles, y compris les caractères de langues mortes et les caractères obsolètes, pour un nombre croissant de symboles et notations. Un seul système d'encodage universel permet ainsi l'échange et le stockage d'informations indépendamment du format de stockage, du système d'exploitation, de l'application, du matériel ou de la langue. Le standard concerne bien plus que le nom et le point de code des caractères, car à chaque caractère sont associées des propriétés qui le décrivent et constituent la référence pour son traitement informatique (Unicode Consortium 2019d). La première version du standard a été publiée en 1991, puis à chaque nouvelle version, des écritures supplémentaires ont été rajoutées progressivement et d'autres complétées.

Unicode est maintenu par le consortium Unicode, une organisation sans but lucratif, dont le site unicode.org héberge la documentation concernant l'Unicode : un glossaire, les tableaux de point de code, des bases de données concernant les caractères (comme la Unicode Character Database et la Unihan Database), des rapports techniques, des propositions de modification du standard.

19. Le standard Unicode et la norme ISO-10646, le Jeu universel de caractères codés (JUC) sont maintenus conjointement pour assurer la compatibilité (ISO/IEC 2017)

Caractère	Appellation Unicode	Point de code
A	LATIN CAPITAL LETTER A	U+0041
ဘ	MYANMAR LETTER A	U+1021

TAB. 1.3 : Une seule police Unicode (Padauk) peut afficher des caractères de plusieurs systèmes d'écriture, car les points de code des caractères sont uniques.

1.5.2.1 Points de code et polices

Les points de code Unicode disponibles sont disposés sur dix-sept plans de points de code, contenant chacun au maximum environ 65 000 caractères, mais dont seulement trois sont actuellement utilisés. Le premier, le plan 0, ou *plan multilingue de base*, est le plus utilisé. Les points de code Unicode sont toujours cités en notation hexadécimale, précédés par U+ pour spécifier qu'il s'agit d'un point de code Unicode. Cette notation hexadécimale a été choisie parce qu'elle est plus lisible et plus facile à manipuler par un humain que les notations binaire ou décimale, sachant que la machine ne manipule par nature que des données en format binaire. Théoriquement, l'Unicode encode les écritures (*scripts* en anglais) plutôt que les systèmes d'écriture de langues individuelles, parce que certains sous-ensembles de caractères peuvent s'utiliser dans plus d'un système d'écriture. Autrement dit, les caractères partagés par des langues différentes ne sont pas dédoublés. Dans la documentation Unicode, les points de code de chaque écriture sont disposés dans les plans de code en tables de code par écriture, appelés *blocks* (chaque bloc désigné par un nom unique en anglais, comme *Armenian*, *Arabic* ou *Devanagari*). Chaque bloc est classé (numériquement) à côté d'écritures similaires, le premier bloc du premier plan de code (le *plan multilingue de base*, numéro 0) étant *Basic Latin (ASCII)*. Les points de code de ce premier bloc du standard sont partagés avec un grand nombre de systèmes non Unicode ou antérieurs à l'Unicode.

Le principe d'encodage d'écritures (comme l'alphabet latin) au lieu de systèmes d'écriture pour des langues spécifiques (comme l'anglais ou le français) est important pour éviter qu'un caractère puisse avoir deux points de code. Toutefois, le standard prend en compte les usages historiques, et considère les caractères similaires des alphabets grec, latin et cyrillique comme distincts, et attribue un point de code différent pour des caractères équivalents. Le tableau 1.4 illustre

Caractère	Appellation Unicode	Point de code
A	LATIN CAPITAL LETTER A	U+0041
a	LATIN SMALL LETTER A	U+0061
Α	GREEK CAPITAL LETTER ALPHA	U+0391
α	GREEK SMALL LETTER ALPHA	U+03B1
А	CYRILLIC CAPITAL LETTER A	U+0410
а	CYRILLIC SMALL LETTER A	U+0430

TAB. 1.4 : Points de code Unicode distincts attribués à des caractères ressemblants

comment ces caractères peuvent porter confusion et effectivement il existe des cas de mélanges d'alphabets, un texte russe avec quelques lettres latines au lieu de leurs équivalents cyrilliques, par exemple.

Contrairement à ces alphabets, les sinogrammes (les caractères chinois) qui forment le jeu de caractères *Unihan*, bien qu'utilisés pour écrire des langues différentes, sont traités comme formant une seule écriture en Unicode, car les sinogrammes ont été considérés historiquement comme un seul système d'écriture. Le bloc Unihan contient des caractères pour les sinogrammes partagés par des langues qui utilisent ce système, comme le chinois (des *hànzì*), le japonais (des *kanji*) et le coréen *hanja*, mais aussi des variantes et caractères des sinogrammes spécifiques à certaines langues, encodés séparément, comme les caractères *chữ Nôm* vietnamiens (utilisés avant 1920), les *kokuji* japonais et les caractères *Sawndip* utilisés pour écrire la langue zhuang et d'autres langues en Chine et au Vietnam. Les sinogrammes simplifiés (adoptés en Chine continentale et à Singapour) ne sont pas considérés comme identiques à leurs équivalents traditionnels (toujours utilisés à Hong Kong, à Macao, à Taiwan et par la diaspora chinoise) car il n'existe pas de correspondance biunivoque entre les deux ; un sinogramme (tel que 台 U+53F0) peut être la forme simplifiée de plus d'un sinogramme traditionnel (檯 U+6AAF, 臺 U+81FA ou 颱 U+98B1)[20]. Un caractère d'un même point de code utilisé dans des langues différentes sera affiché, quelle que soit la police compatible Unihan choisie, mais selon la tradition esthétique la forme du glyphe peut varier légèrement ; il est donc conseillé de choisir une police conçue selon la tradition de la région, ou la langue véhiculée par l'écriture. Dans

20. Ces quatre caractères se prononcent tous tái, sauf lorsque 台 est utilisé comme le nom de famille Yí.

ces cas, les différences seront gérées au niveau de l'affichage par la police. Par exemple, le caractère U+9AA8 affiché par une police conçue pour les sinogrammes simplifiés affichera le composant interne en haut du sinogramme positionné à gauche 骨 (/gǔ/, *os* en français[21]), alors que ce même caractère Unicode sera affiché 骨 (avec le composant interne en haut du caractère positionné à droite et le radical 月 *viande* au lieu du radical 月 *lune*) par une police conçue pour le chinois traditionnel[22] et 骨 par une police japonaise[23], le haut correspondant à l'esthétique du chinois traditionnel et le bas du chinois simplifié.

Peu de polices pourvoient des glyphes pour l'ensemble des 143 859[24] caractères spécifiés dans la norme Unicode, car, outre le travail immense que dessiner un tel nombre de glyphes représente, une police qui couvre tous les caractères peut s'avérer trop lourde et ingérable pour certaines applications. En général, les concepteurs de polices visent à couvrir un ou plusieurs jeux de caractères et le bloc Basic Latin. Quand un caractère ne s'affiche pas correctement, le jeu de caractères auquel il appartient (et ainsi la police nécessaire pour son affichage) peut s'identifier à partir du point de code du caractère[25]. Si le système d'exploitation n'affiche pas directement le point de code, il est possible d'utiliser une *fallback font* (police de repli) comme l'*Unicode BMP Fallback Font*[26] qui couvre à peu près la moitié des caractères Unicode, soit presque toutes les écritures de toutes langues vivantes.

De telles polices à spectre large sont souvent utilisées comme police par défaut par certains logiciels, ce qui peut entraîner des inconvénients pour l'utilisateur. A titre d'exemple, avec la police *Unicode BMP Fallback Font* installée, l'utilisateur du logiciel de textométrie TXM (Heiden et al. 2010) peut rencontrer des problèmes d'affichage, même après avoir configuré soigneusement TXM avec la police adaptée à la langue de son corpus. Dans la figure 1.6 l'affichage du concordancier est correct, mais le texte birman de la colonne de gauche est affiché dans la police

21. Affiché ici avec la police *NotoSansCJKsc-Regular*. Toutes les polices de la famille *Noto* créées par Google, y compris celles pour l'écriture Myanmar, sont disponibles ici : https://www.google.com/get/noto/.

22. Affiché ici avec la police *NotoSerifCJKtc-Regular*

23. Affiché ici avec la police *NotoSerifCJKjp-Regular*

24. Unicode version 13.0

25. Le bloc Unicode peut s'identifier en interrogeant la base de données http://www.fileformat.info et l'utilisateur peut dénicher les polices appropriées sur le site du Consortium Unicode https://www.unicode.org/resources/fonts.html

26. *Unicode BMP Fallback Font* est une police de Sil International https://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&id=UnicodeBMPFallbackFont

Unicode BMP Fallback Font, car elle est la police la plus importante installée qui comprend des glyphes pour les points de code birmans. Une fois la police

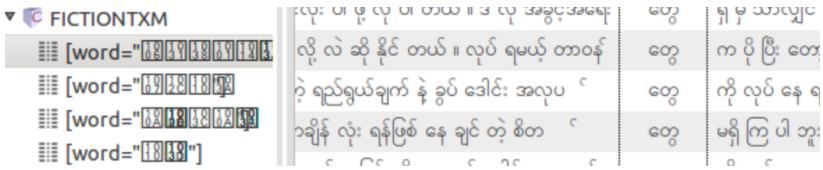


FIG. 1.6 : Exemple d'affichage dans TXM erroné, dû à l'installation d'une police « gloutonne ». La colonne à gauche n'est pas affichée dans la police choisie à l'import du corpus si une police plus importante est installée

Unicode BMP Fallback Font désinstallée, la police par défaut pour les points de code birmans choisie par le logiciel devient la police la plus importante qui comprend ces points de code, une police birmane, appelée Padauk[27], et le texte birman de la colonne à gauche s'affiche correctement (figure 1.7). Cette anomalie

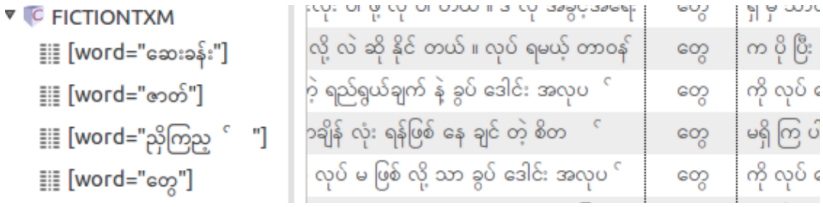


FIG. 1.7 : Affichage correct sans la police *Unicode BMP Fallback Font* installée

dépend des choix du développeur du logiciel, mais peut être surmontée par une gestion méticuleuse des polices du système d'exploitation.

1.5.2.2 L'ordre canonique et propriétés de caractères

Le standard Unicode ne spécifie pas uniquement l'encodage, mais pour certaines écritures, il stipule aussi le bon ordre de stockage des caractères, appelé *ordre canonique*. Un exemple simple qui démontre l'importance du placement fixe concerne le stockage des caractères accentués. Les caractères **e** et ˆ dans l'ordre ˆ + **e** sont stockés comme un seul caractère accentué, **ê** (LATIN SMALL LETTER E WITH CIRCUMFLEX U+00EA), alors que dans l'ordre **e** + ˆ ils sont stockés

27. La police *Padauk* de SIL International, est disponible ici : https://software.sil.org/padauk/. Padauk (ပိတောက် /pḭtaʊʔ/) est la fleur nationale de l'Union de Myanmar (Ryan 2017)

comme deux caractères, **e** (LATIN SMALL LETTER E U+0045) et ˆ (CARET U+2038).

(18) ^ + e ⇒ ê
U+2038 + U+0045 ⇒ U+00EA

(19) e + ^ ⇒ e^
U+0045 + U+2038 ⇒ U+0045 U+2038

Seul l'ordre canonique Unicode permet la composition de caractères en un seul, appelé un caractère précomposé. Cet ordre est implémenté par la désignation de propriétés pour chaque caractère dans la base de données des caractères Unicode[28], qui comprend des propriétés telles que la largeur, la direction de l'écriture et la catégorie de caractère (type de lettre, de ponctuation ou chiffre, par exemple) (Unicode Consortium 2019b). Certains caractères qui se combinent avec d'autres possèdent les propriétés de *combinant avec espacement* (*spacing, combining*) ou *sans espacement* (*nonspacing*), c'est-à-dire sans espace horizontale supplémentaire, comme le caractère ˆ (CARET U+2038) de l'exemple 18. Il faut noter toutefois que les propriétés de *combinant avec espacement* ou *sans espacement* n'impliquent pas forcément que les caractères qui se combinent à l'affichage doivent être stockés comme un seul. Dans l'exemple 20, les caractères préservent leurs propres points de code, bien que ◌̆ ⟨j̆⟩ soit un caractère avec la propriété *combinant avec espacement* :

(20) ə ⟨kʰ⟩ + ◌̆ ⟨j̆⟩ ⇒ ə̆ ⟨kʰj̆⟩ /tɕʰạ/ *termite*
U+1001 + U+103C ⇒ U+1001 U+103C

Nous verrons plus d'exemples de caractères avec les propriétés *combinant avec espacement* et *sans espacement* par la suite, car l'encodage Unicode du système d'écriture du birman en fait usage pour afficher les voyelles et consonnes dépendantes.

Dans Unicode l'agencement des codes de caractères est particulièrement pertinent pour les écritures dérivées de l'écriture brāhmī, dont les consonnes peuvent se combiner en ligatures (voir exemples 10 et 11) et la disposition de voyelles n'est pas toujours linéaire (comme l'exemple 2). L'ordre de caractères stipulé par les règles Unicode reflète aussi étroitement que possible l'ordre phonétique, mais pour certaines écritures, comme le thaï, l'ordre d'écriture de gauche à droite a

28. *Unicode Character Database* http://www.unicode.org/reports/tr44/.

été maintenu pour préserver la compatibilité avec les systèmes antérieurs. Dans l'exemple 21, on constate que l'ordre de stockage suit la translittération de l'écriture ⟨e̱mḵʰ⟩, et non pas la transcription de la prononciation /mèːk/, le caractère vocalique ⟨e⟩ (ici en vert) étant stocké avant le caractère consonantique, bien que la voyelle se prononce après la consonne.

(21) ⟨e⟩ + ⟨m̱⟩ + ⟨ḵʰ⟩ ⇒ ⟨e̱mḵʰ⟩ /mèːk/ *nuage*
U+0E40 + U+0E21 + U+0E06 ⇒ U+0E40 U+0E21 U+0E06

Quel que soit l'ordre canonique Unicode pour une écriture, c'est l'immuabilité du système qui importe le plus, car l'agencement figé normatif des caractères est essentiel pour garantir l'uniformité d'affichage, l'intégrité de stockage et la récupération de données fiable. Afin de satisfaire les exigences de continuité du standard Unicode, une fois normalisé, l'ordre de caractères stipulé par le standard ne peut être changé. Ceci garantit la validité persistante des données textuelles encodées (Unicode Consortium 2019d).

1.5.3 *L'encodage du birman en Unicode*

La partie du bloc *Myanmar*[29] qui contient les caractères nécessaires pour le système d'écriture du birman est inchangée depuis la version 5.1 de l'Unicode datant de 2006. L'implémentation en Unicode des caractéristiques de l'écriture des langues de Myanmar est résumée par la figure 1.8. Le bloc principal *Myanmar* s'étend du point de code U+1000 à U+109F (un partie du plan multilingue de base, le plan 0), et comprend les caractères nécessaires à la représentation du birman, mais aussi d'autres langues de la Birmanie : le môn, le karène sgaw, le karène pwo occidental, le karène pwo oriental, le karène geba, le kayah, le shan, le palaung rumai, le shan khamti, l'aiton et le phake, ainsi que des caractères pour le pâli et le sanskrit. Ces deux dernières langues mortes, bien que notées dans d'autres écritures ailleurs ont une tradition écrite importante en écriture Myanmar (voir figure 1.9). Deux blocs supplémentaires, *Myanmar Extended-A*[30] qui s'étend du point de code U+AA60 au point de code U+AA7B et *Myanmar Extended-B*[31] (U+A9E0 à U+A9FF) comprennent des caractères supplémentaires pour le shan khamti, l'aiton, le karène pa'o, le tai laing, le shwe palaung et l'écriture shan

29. Voici le lien permanent vers le bloc *Myanmar* : http://www.unicode.org/charts/PDF/U1000.pdf
30. http://www.unicode.org/charts/PDF/UAA60.pdf
31. http://www.unicode.org/charts/PDF/UA9E0.pdf

Nom de l'écriture	*Myanmar*
Nombre de caractères	223 (73)
Caractères combinant	62 (20)
Combinaisons de caractères multiples	oui
Position	relative contextuelle
Signes vocaliques	oui
Écriture cursive	non
Sensible à la casse	non
Formes contextuelles	oui
Direction	gauche à droite
Segmentation lexicale	non
Chiffres propres	oui

FIG. 1.8 : Résumé des caractéristiques relatives à l'écriture *Myanmar* implémentée en Unicode (Informations relatives au système d'écriture du birman entre parenthèses)

du pâli. Le site *scriptsource.org*[32] liste vingt-neuf langues qui utilisent l'écriture des blocs *Myanmar*. Nous voyons donc que les blocs *Myanmar* concernent une écriture utilisée pour plusieurs langues, dont le système d'écriture de la langue birmane est un sous-ensemble (de la même manière que le jeu de caractères Unihan est partagé par plusieurs langues). C'est pour cette raison qu'en Unicode le terme *Myanmar* est réservé pour l'écriture (le nom du pays), et le terme *Burmese* pour la langue.

Comme nous avons évoqué dans la section précédente, le standard Unicode précise plus que le code d'un caractère et son nom, et ces aspects supplémentaires sont particulièrement importants pour le système d'encodage de l'écriture de la langue birmane. Les propriétés (appelées *catégories générales* dans la documentation Unicode, Unicode Consortium (2019b)) attribuées à chaque caractère déterminent les relations entre caractères et comment ils devraient être affichés à l'écran par les polices. L'implémentation fidèle ou partielle du système a une influence considérable sur l'intégrité des données textuelles en birman. Le tableau 1.5 résume les caractères utilisés pour représenter la langue birmane en Unicode avec certaines de leurs catégories et propriétés.

Le standard Unicode précise les principes de codage dans des notes techniques.

32. https://www.scriptsource.org/cms/scripts/page.php?item_id=script_detail&key=Mymr

Catégories Unicode officielles	**Caractères**	**Codes propriétés**	**Points de code**
Consonants	ကခဂဃငစဆဇဈညဋဌဍဎဏတထဒဓနပဖ ဗဘမယရလဝသဟဠ	Lo	U+1000–U+1020
Consonant	ဿ	Lo	U+103F
Independent vowels	အဣဤဥဦဧဩဪ	Lo	U+1021, U+1023–U+1027, U+1029, U+102A
Dependent vowel signs	ါာေ ိီုူဲ	Mc Mn	U+102B, U+102C, U+1031 U+102D–U+1030, U+1032
Various signs	ံ့ း	Mn Mc	U+1036, U+1037 U+1038
Virama and killer	္်	Mn	U+1039, U+103A
Dependent consonant signs	ျြ ွှ	Mc Mn	U+103B, U+103C U+103D, U+103E
Digits	၁၂၃၄၅၆၇၈၉၀	Nd	U+1040-U+1049
Punctuation	၊ ။	Po	U+104A, U+104B
Various signs	၌၍၎၏	Po	U+104C-U+104F

TAB. 1.5 : Les caractères du système d'écriture du birman en Unicode et leurs propriétés générales. Sources : Unicode Consortium 2019b ; Unicode Consortium 2019e
Clé : Lo = lettre, autre Mn = marque sans espacement Mc = marque combinante avec espacement Nd = nombre, chiffre décimal Po = ponctuation, autre.

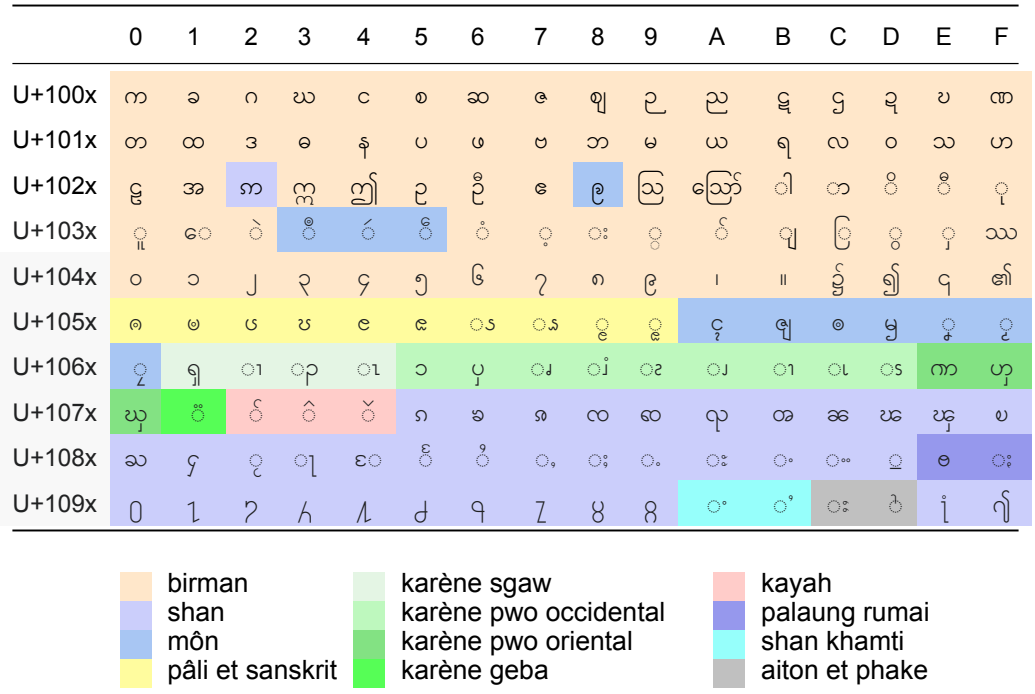


FIG. 1.9 : Table de caractères Unicode (version 13.0) du bloc de base Myanmar.

Celle rédigée par Martin Hosken (2012) précise tous les principes d'encodage du bloc *Myanmar*. Nous résumons ici les principes concernant le birman les plus importants pour nos explications ultérieures.

Premier principe : les caractères sont stockés dans l'ordre de prononciation. Contrairement à l'écriture thaïe, les systèmes d'écriture du bloc Myanmar sont pourvus de ligatures complexes, il n'était donc pas question d'utiliser l'ordre d'écriture ; ceci aurait laissé la voie ouverte pour plusieurs ordres de stockage de caractères. Voici des exemples de la position du caractère vocalique ေ◌ ⟨e⟩, point de code U+1031, qui se place *devant* la consonne bien que la voyelle qu'il représente est prononcée après :

(22) နေ ⟨ne⟩ /nè/ U+1014 U+1031 *soleil*

(23) မြွေ ⟨mǰw̆e⟩ /mjwè/ U+1019 U+103C U+103D U+1031 *serpent*

Deuxième principe : les formes libres vocaliques sont encodées par un seul point de code, alors que les signes vocaliques analytiques (les matras), qui peuvent être composés de deux caractères, sont codés par plusieurs points de code. Ainsi

ေ◌ော, le signe vocalique composé qui entoure le signe consonantique, prononcé /-ɔ́/, est composé de ေ◌ ⟨e⟩ U+1031 + ◌ော ⟨a⟩ U+102C :

(24) ေတာ ⟨tea⟩ /tɔ́/ U+1010 U+1031 U+102C *forêt*

(25) ဩ ⟨ɔ̣⟩ /ʔɔ́/ U+1029 MYANMAR LETTER O

Troisième principe : l'encodage des formes dépendantes et indépendantes d'une consonne ou voyelle donnée sont encodées avec des points de code distincts (ရ ⟨ṙ⟩ est U+101B, ◌ျ ⟨ǰ⟩ est U+103C), alors que les différentes formes visuelles d'un même caractère ne sont pas codées différemment, mais sculptées par les procédés de rendu 'intelligents' des polices, qui choisissent la forme nécessaire au contexte[33]. Ainsi, une police supportant une écriture donnée contiendra bien plus de glyphes que les caractères de cette même écriture présentés dans la documentation Unicode, car celle-ci dernier ne définit que des caractères abstraits. Ces formes graphiques qui dépendent de la position relative d'un caractère à d'autres caractères contigus sont appelées des *formes de présentation*. Pour illustrer ce phénomène, voici des exemples de formes différentes du caractère ◌ျ ⟨ǰ⟩ U+103C, rendues par des glyphes légèrement différents :

(26) ကျာ ⟨kǰa⟩ /tɕà/ U+1000 U+103C U+102C *durer longtemps*

(27) ခ္ခ်ီး ⟨kʰǰmˣ²⟩ /tɕʰáɴ/ U+1001 U+103C U+1019 U+103A U+1038 *diviser*

(28) ကြဲ ⟨kǰɛ⟩ /tɕɛ́/ U+1000 U+103C U+1032 *éparpiller*

L'exemple 26 du glyphe du caractère ◌ျ ⟨ǰ⟩ montre une forme large pour pouvoir entourer le caractère ◌ာ ⟨k⟩. Le même caractère est rendu avec un glyphe moins large dans l'exemple 27 qui entoure un caractère plus étroit, ခ ⟨kʰ⟩. Dans l'exemple 28, le glyphe choisi est coupé sur la partie supérieure droite pour laisser place au caractère vocalique ◌ဲ ⟨ɛ⟩. Dans chaque cas, le caractère est le même, encodé avec le code U+103C.

Les différentes formes de présentation du caractère ◌ု ⟨ù⟩ U+1030 sont affichées de la même façon avec des glyphes différents :

(29) ဖြူ ⟨pʰǰù⟩ /pʰjù/ U+1016 U+103C U+1030 *blanc, pur, porc-épic*

(30) မူး ⟨mù²⟩ /mú/ U+1019 U+1030 U+1038 *étourdi, ivre*

33. Les polices Unicode utilisent les technologies de rendu OpenType, Pango ou Graphite.

Une exception est faite pour les deux formes de la voyelle /à/, ◌ာ ⟨a⟩ (U+102C MYANMAR VOWEL SIGN AA) et ◌ါ ⟨ä⟩ (U+102B MYANMAR VOWEL SIGN TALL AA). Les deux formes de présentation sont encodées avec deux points de code différents, alors qu'en birman, il s'agit du même caractère abstrait, dont la réalisation typographique varie selon la forme de la consonne précédente[34], voir les exemples 31 et 32. L'encodage sur deux points de code (au lieu d'un seul avec un procédé de sélection entre deux glyphes selon le contexte) a été choisi pour prendre en considération l'usage de l'écriture Myanmar par d'autres langues, comme le karène sgaw, qui ne font usage que de la forme variante ◌ါ ⟨ä⟩ U+102B (Hosken 2012) comme l'illustre l'exemple 33.

(31) ဿ ⟨śa⟩ U+1005 U+102C *lettre*

(32) ခါ ⟨kʰä⟩ U+1001 U+102B *secouer*

(33) KARÈNE SGAW လါ ⟨lä⟩ U+101C U+102B *vert, mois, lune*
BIRMAN လာ ⟨la⟩ U+101C U+102C *venir*

Les consonnes souscrites en birman sont pour la plupart considérées comme des formes d'un même caractère et à ce titre n'ont pas de codes distincts non plus. On utilise le dispositif du virāma invisible[35], le caractère ◌္ ⟨◌̣⟩ (U+1039 MYANMAR SIGN VIRAMA) pour superposer des consonnes et inhiber la voyelle inhérente de la consonne indépendante supérieure de la superposition. On constate que, là encore, il y a deux formes de présentation d'un même glyphe က (U+1000), une plus petite que l'autre, comme dans l'exemple 34.

(34) စက္ကူ ⟨śk̇kù⟩ /sɛʔkù/ U+1005 U+1000 U+1039 U+1000 U+1030 *papier*

Cette superposition a l'effet d'inhiber la voyelle inhérente de la première consonne, comme nous avons vu précédemment avec les consonnes finales en fin de syllabe fermée (voir l'exemple 12)[36]. S'il n'y avait pas de virāma invisible ◌္, l'exemple 34 serait écrit စက်ကူ ⟨śkˣkù⟩. Pour faciliter les procédés de combinaison en ligature, les consonnes médiales, ◌ျ ⟨y̌⟩, ◌ြ ⟨ǰ⟩, ◌ွ ⟨w̌⟩ et ◌ှ ⟨ȟ⟩ sont encodés sur leurs propres points de code (U+103B à U+103E, MYANMAR CONSONANT SIGNS

34. Dans le système d'écriture du birman, la forme ◌ါ ⟨a :⟩ est employée après les caractères consonantiques ခ ⟨kʰ⟩, ဂ ⟨g⟩, ငဂ ⟨ŋ⟩, ဒ ⟨d⟩, ဓ ⟨d̪ ⟩, ဝ ⟨p⟩, et ဝ ⟨w⟩

35. Tout seul, ce caractère s'affiche ◌္, mais devient invisible en combinaison avec d'autres caractères.

36. Pour rappel, le symbole utilisé dans ce cas est ◌် ⟨ˣ⟩ U+103A MYANMAR SIGN ASAT, exemple d'usage : ထင် ⟨tʰŋˣ⟩ /tʰɪ̀n/ U+1011 U+1004 U+103A *penser*

MEDIAL YA, RA, WA et HA respectivement)[37]. L'exemple 23 montre une ligature complexe à trois caractères.

Le quatrième principe concerne l'ordre de stockage physique en mémoire des caractères : il ne peut y avoir qu'un seul ordre de stockage. L'ordre est relatif. Cet aspect est très important, car ce principe assure qu'il ne peut y avoir qu'un seul ordre de tri. Les polices conçues pour suivre strictement le standard Unicode pour le birman n'acceptent pas l'ordre non canonique[38], et indiquent une séquence de points de code mal formée à l'aide du glyphe ◌, un cercle pointillé[39]. L'exemple 35 montre une erreur d'ordre de saisie et sa correction. Le caractère vocalique dépendante ေ◌ ⟨e⟩ (U+1031 MYANMAR VOWEL SIGN E) étant mal placée par rapport à la consonne initiale.

(35) *ေ◌ရ ⟨er⟩ U+1031 U+101B
ေရ ⟨re⟩ U+101B U+1031 *eau*

A cet égard, un seul aspect du standard Unicode pour l'écriture du birman peut porter à confusion. La note technique de Hosken (2012) explique que l'ordre de deux caractères, la marque de ton ့ (U+1037 MYANMAR SIGN DOT BELOW ou *aukmyit*) et le virāma visible ် (U+103A MYANMAR SIGN ASAT) doit toujours être normalisé avec la marque de ton en première position. L'exemple 36 montre l'ordre canonique, ့ (U+1037) + ် (U+103A)

(36) ဖ့်် ⟨pʰŭŋ[1x]⟩ U+1016 U+103D U+1004 U+1037 U+103A *ouvrir*

(37) *ဖ့်် ⟨pʰŭŋ[x1]⟩ U+1016 U+103D U+1004 U+103A U+1037 *ouvrir*

Ceci diffère de l'ordre de frappe naturel (exemple 37), et signifie que les caractères des marques de ton ◌း (U+1038 MYANMAR SIGN VISARGA) et ့ *aukmyit* ne sont pas ordonnés par rapport au caractère *asat* ် (U+103A) suivant la même logique, comme le montre les exemples 38 et 39.

37. Contrairement aux apparences, ◌္ U+1039 MYANMAR SIGN VIRAMA n'est pas utilisé pour superposer consonnes indépendantes et ◌ွ ⟨w̆⟩, bien que cette consonne médiale à l'air d'une consonne souscrite.

38. Toutefois, il faut noter que certains systèmes de saisie et certains sites web normalisent automatiquement l'ordre des caractères pour certaines combinaisons, ce qui peut donner l'impression qu'un ordre non canonique n'est pas indiqué par la police en question.

39. Il faut bien distinguer le glyphe du caractère codé. Ce cercle pointillé utilisé de cette manière apparaît uniquement à l'affichage et n'est pas encodé dans le texte. Si toutefois on a besoin d'utiliser le *caractère* du cercle pointillé, il s'agirait du point de code U+25CC, nommé DOTTED CIRCLE.

(38) ချင်း ⟨kʰɨ̌ŋ[xz]⟩ U+1001 U+103B U+1004 U+103A U+1038 *gingembre*

(39) ချင့် ⟨kʰɨ̌ŋ[1x]⟩ U+1001 U+103B U+1004 U+1037 U+103A *mesurer*

On constate dans les deux exemples 36 et 37, qu'il n'y a pas ici de différence visible entre ces l'ordre canonique et l'ordre non canonique. L'utilisateur n'a donc pas de retour visuel d'une éventuelle faute de frappe (au moins avec cette police, Padauk). Les claviers birmans devraient normaliser cet ordre en ordre canonique, mais nous avons trouvé un mélange d'ordres canonique et non canonique dans nos corpus, ce qui laisse croire que beaucoup de claviers ne sont pas conçus pour empêcher la saisie de caractères en ordre non canonique. Selon Hosken (2012), il s'agit d'une erreur dans la vérification du standard qui, pour des raisons de pérennité de l'intégrité de données, ne peut être corrigée. Toutefois, au chapitre dédié aux langues de l'Asie du Sud-est de la dernière version du standard (Unicode Consortium 2019a), il est recommandé que les deux ordres soient considérés comme équivalents, mais en réalité ce n'est pas toujours le cas. Par exemple, le moteur de recherche de Google et le Wiktionnaire birman considèrent les deux ordres comme équivalents, mais le dictionnaire Sealang Burmese n'accepte que l'ordre canonique. En ce qui concerne le traitement automatique, le taliste doit choisir quel ordre utiliser pour stocker des données, sachant que l'ordre canonique n'est pas forcément le plus pratique. En effet, dans certains cas l'ordre canonique n'est pas souhaitable en ce qui concerne la segmentation en unités d'analyse (la tokenisation), que nous discuterons plus en détail dans les chapitres qui suivent. Ces principes d'encodage ne concernent que le stockage — il est possible de permettre à l'utilisateur de saisir les caractères dans l'ordre qu'il veut, mais le standard Unicode stipule que le créateur du clavier doit prendre en compte les règles pour que les caractères soient correctement enregistrés dans l'ordre standard. Comme nous verrons, parmi les claviers conçus pour la saisie de données selon le standard Unicode, tous ne font pas la normalisation de l'ordre de stockage. En outre, certaines polices qui respectent les correspondances glyphe-code Unicode ne respectent pas l'ordre canonique, ou ne l'appliquent pas strictement.

1.5.3.1 Les encodages pré-Unicode et leurs polices

Jusqu'ici nous avons vu un seul système d'encodage de caractères, nommé *Unicode*, dont le texte peut s'afficher grâce à une myriade de polices telles

que *Verdana*, *Padauk*, *Libertine*, *Lobster Two*, *NotoSansCJKsc*, *NotoSerifCJKjp* ou *NotoSerifCJKjp*. En ce qui concerne le birman, les systèmes qui précédaient Unicode pour la plupart ne faisaient pas de distinction entre le nom du système d'encodage et le nom de la police qu'il fallait pour l'interpréter, car souvent ces systèmes ne fonctionnaient qu'avec une seule police. Pour cette raison, on appelle souvent à ce genre d'encodage *font encoding* ou *encodage à polices*. D'une façon moins technique, ces systèmes sont aussi appelés *legacy encodings* (encodages *patrimoniaux*), appellation qui peut se référer aux encodages spécifiques à un système d'exploitation, à un standard national ou international antérieur à l'Unicode ou bien aux encodages créés par des entreprises ou des particuliers sans référence à un standard particulier. Puisque jusqu'au présent il n'y a pas eu de standard national officiel de l'encodage de l'écriture Myanmar (Unicode Consortium 2019a), il existe de nombreux types de *legacy encodings*.

Les premières tentatives d'informatisation du système d'écriture du birman étaient basées sur les points de code de l'encodage ASCII, remplaçant les glyphes des caractères latins par les glyphes représentant les caractères birmans. Ainsi, les correspondances entre caractère abstrait et point de code sont complètement différentes de celles de l'Unicode. Il existe plusieurs encodages différents basés sur ASCII, donc les correspondances entre glyphe et code diffèrent aussi entre ces systèmes. L'exemple 40 montre le birman encodé par le système Win Innwa[40].

(40) police *wininnwa.ttf* : စန္ဒကူး *santal*
police système (Unicode) : pe´ul ;
points de code : U+70 U+65 U+B4 U+75 U+6C U+3B

On constate que ces points de code sont bien inférieurs à ceux du bloc *Myanmar*, car ils proviennent de la plage ASCII du premier bloc (*Basic Latin*) au début du premier plan Unicode, le *plan multilingue de base*. Voici le même exemple, mais cette fois avec un autre encodage à base d'ASCII, Kannaka.

(41) police *KANNAKA_TTF* : စန္ဒကူး
police *wininnwa.ttf* : ꧋ည꧌း
police système (Unicode) : sn »k¨ ;
points de code : U+73 U+6E U+BB U+6B U+A8 U+3B

40. Bien qu'il ne s'agit pas d'un encodage Unicode, nous citons les points de code Unicode pour faciliter la comparaison.

Nous voyons que les encodages Win Inwa et Kannaka sont incompatibles et qu'il est impossible de déchiffrer le texte sans la police adéquate.

Ce genre de polices n'encode pas simplement les caractères abstraits, mais des formes des glyphes, y compris différentes formes contextuelles. Cela veut dire que les correspondances ne sont pas entre caractères et points de code, mais entre glyphes et points de code. Il ne s'agit donc pas d'encodages de caractères comme Unicode, mais d'encodages de glyphes. Le tableau 1.6 montre les différents points de code des glyphes d'un même caractère pour l'encodage Win Innwa. Les cas de ligature comme l'exemple du deuxième ligne du tableau, où un point de code est attribué à une combinaison de deux signes consonantiques, compliquent le tri alphabétique, ce qui constitue un des défauts majeurs des encodages à base de glyphes.

Encodage Win Innwa Police *wininnwa.ttf*	**Point de code**
ဂ	U+6A
ဂ္ဂ	U+3E
ဂ်ဂ	U+42
ဂ	U+4E

TAB. 1.6 : Encodage de glyphes au lieu de caractères abstraits par le système Win Innwa. Les différentes formes contextuelles d'un même caractère sont encodées sur des points de code distincts

Par contre, en encodage Unicode, un seul code est utilisé, la forme du glyphe est choisie automatiquement lors du rendu selon le contexte. De ce fait, le jeu de caractères nécessaire pour encoder le birman en Unicode est considérablement plus restreint que les systèmes encodant les glyphes. Dans le tableau 1.7, les différentes formes contextuelles ont le même point de code.

L'inconvénient de ces encodages est qu'ils sont tous différents et incompatibles entre eux et aucun de ces systèmes n'a émergé comme standard (même de facto) — l'utilisateur doit spécifier la police qui a été utilisée pour saisir les données s'il veut partager son fichier. En outre, avec ce genre d'encodage mono-écriture (le seul birman), la manipulation de documents multilingues (franco-birmans par

Encodage Unicode Police *Padauk*	**Points de code**
ကြ	U+1001 U+103C
ကျ	U+1001 U+103C U+103D
ကြံ့	U+1000 U+103C U+1032
ပြုံး	U+1015 U+103C U+102E U+1038

TAB. 1.7 : Glyphes différents d'un seul caractère Unicode. Le caractère ◌ U+103C MYANMAR CONSONANT SIGN MEDIAL RA prend des formes différentes selon son contexte immédiat.

ex.) est très fastidieuse, car ceux-ci nécessitent au moins une deuxième police compatible pour les caractères non birmans. Sans les deux polices, une partie du texte (soit les caractères birmans, soit les caractères latins) sera mal décodée, comme ci-dessous dans les deux lignes à police unique de l'exemple 42.

(42) polices *Libertine* + *AvaLaser* : Charles လာ၊ *C'est Charles ?*
police *Libertine* : Charles la ;"
police *AvaLaser* : ဦးဟာရုလစေ လာ၊

Ainsi, tout traitement automatique de textes en birman encodé ASCII doit subir un traitement spécifique pour identifier la partie birmane à re-encoder en Unicode, car les textes birmans comprennent souvent des mots étrangers en caractères latins. Pour de petites quantités de texte à encodage unique (sans texte en caractères latins), on peut utiliser un convertisseur en ligne, comme l'outil *Burmese Font Converter*[41] ; les stratégies plus élaborées pour traiter le problème de déchiffrage de texte encodé en ASCII et à encodages multiples seront abordées dans le chapitre suivant.

1.5.3.2 Les encodages quasi Unicode à base d'Unicode

Après la création par le consortium Unicode du bloc *Myanmar* dans le jeu universel de caractères, la création de polices pour le birman a connu une période d'implémentation quelque peu désordonnée. Ceci a été dû au manque de prise en charge de procédés de rendu intelligent des glyphes nécessaires à un affichage conforme aux normes Unicode par les logiciels et systèmes d'exploitation les plus

41. https://burglish.my-mm.org/latest/trunk/web/fontconv.htm

couramment utilisés (comme Microsoft Windows ou les logiciels Adobe), mais aussi par manque de compréhension ou d'attention aux particularités du standard. Puisque les créateurs de ces polices revendiquent que leurs polices sont des polices Unicode, il n'y a pas de nom spécifique pour désigner les modes d'encodage employés lors de leur création, on ne peut qu'utiliser les noms de polices elles-mêmes. C'est le cas des systèmes Zawgyi-One[42] et Ayar (ဧရာ), le premier étant le système d'encodage de loin le plus répandu au Myanmar (Pann Yu Mon et al. 2011). Dans la documentation Unicode sur les écritures de Myanmar (Unicode Consortium 2019c), ils sont appelés des *encodages ad hoc*. Ces deux encodages ad hoc sont souvent revendiqués comme des polices Unicode, car ils utilisent les points de code Unicode. Néanmoins, ils ne respectent pas tous les principes d'encodage Unicode que nous venons d'examiner, à savoir : les correspondances caractère-code Unicode, l'ordre de stockage (l'ordre de prononciation), l'encodage des caractères et non pas des glyphes, et la rigidité de l'ordre de stockage.

Le système Zawgyi-One utilise les points de code du bloc Unicode Myanmar, mais ne respecte pas entièrement les correspondances caractère-point de code Unicode (voir la figure 1.10). Ceci est dû au fait que les polices du système Zawgyi-One n'utilisent pas des procédés d'affichage complexes, seules capables de donner des formes contextuelles aux glyphes ; ces formes contextuelles sont codées avec des points de code distincts à l'instar des encodages à base d'ASCII comme Win Innwa (voir tableau 1.7). Cela veut dire que le principe de code unique par caractère abstrait n'est pas suivi. Par exemple, en encodage Zawgyi-One les différentes formes de présentation du caractère ◌ျ ‹ǰ›, codé uniquement `U+103C` en Unicode, sont codées sur `U+103B` et les points de code allant de `U+107E` à `U+1084`. Le même principe d'encodage de glyphes est employé pour la superposition de consonnes, les formes consonantiques souscrites recevant ses propres codes, au lieu d'utiliser le dispositif du virāma invisible (`U+1039`). L'exemple 43 compare les points de code nécessaires pour l'encodage Zawgyi-One (ici affichés avec la police *Zawgyi-One.ttf*) et l'encodage Unicode (affichés avec la police *Padauk*) d'un même texte. On constate que les caractères en Zawyi-One ဃ et ် n'ont pas les mêmes points de code (`U+1019` et `U+107C` respectivement).

42. Zawgyi ဇော်ဂျီ /zɔ̀dʑì/ est un personnage surnaturel, magicien et alchimiste du folklore birman.

(43) Zawgyi-One : ဓမ္မာ U+1013 U+1019 U+107C U+102C *dharma*
Unicode : ဓမ္မာ U+1013 U+1019 U+1039 U+1019 U+102C

Le système d'encodage Zawgyi-One a donc besoin de beaucoup plus de points de code que le système d'encodage Unicode, utilisant les points de code réservés pour les autres langues de Birmanie utilisant la même écriture que le birman, avec quelques lettres supplémentaires. Il s'ensuit que les autres langues sont exclues de tout texte encodé en Zawgyi-One, et il est donc impossible d'avoir un texte multilingue à l'instar des systèmes à base d'ASCII. Autre inconvénient majeur d'une police de plus grande taille, la police *Zawgyi-One.ttf* a tendance à s'imposer comme police par défaut quand elle est installée côte à côte avec des polices Unicode, tout comme notre exemple du *Unicode Fallback Font*, figure 1.6.

L'encodage Zawgyi-One n'utilise pas l'ordre de prononciation comme ordre de stockage. Ceci est illustré par l'exemple suivant où la position relative du stockage du caractère ဘ ⟨e⟩ /è/, reflétée dans la translittération ⟨eb́²⟩, diffère de sa prononciation /bè/. En Unicode, l'ordre de stockage (et donc la translittération) et la prononciation sont les mêmes.

(44) Zawgyi-One : ဘေး ⟨eb́²⟩ /bé/ U+1031 U+1018 U+1038 *côté*

Unicode : ဘေး ⟨b́e²⟩ /bé/ U+1018 U+1031 U+1038

Le principe de la rigidité de l'ordre de stockage n'est pas pris en compte par l'encodage Zawgyi-One. Prenons l'exemple de ခိုး /kʰó/ (*voler, cambrioler*). En Unicode, ce verbe est encodé suivant l'exemple 45, où ို est stocké dans l'ordre ိ ⟨i⟩ + ု ⟨u⟩. Cet ordre est aussi valable en encodage Zawgyi-One.

(45) Unicode : ခိုး ⟨kʰiu²⟩ U+1001 U+102D U+102F U+1038 ိ + ု

Zawgyi-One : ခိုး ⟨kʰiu²⟩ U+1001 U+102D U+102F U+1038 ° + ု

Or, à l'inverse d'Unicode, un ordre de stockage différent, ု ⟨u⟩ + ိ ⟨i⟩, produit aussi un affichage satisfaisant en encodage Zawgyi-One :

(46) Unicode : ခုိး U+1001 U+102F U+102D U+1038 ု + ိ

Zawgyi-One : ခိုး U+1001 U+102F U+102D U+1038 ု + °

Bien que superficiellement identiques à l'affichage avec la police *Zawgyi-One*, les exemples 45 et 46 sont traités comme différents par l'ordinateur, car il ne s'agit plus de la même chaîne de caractères.

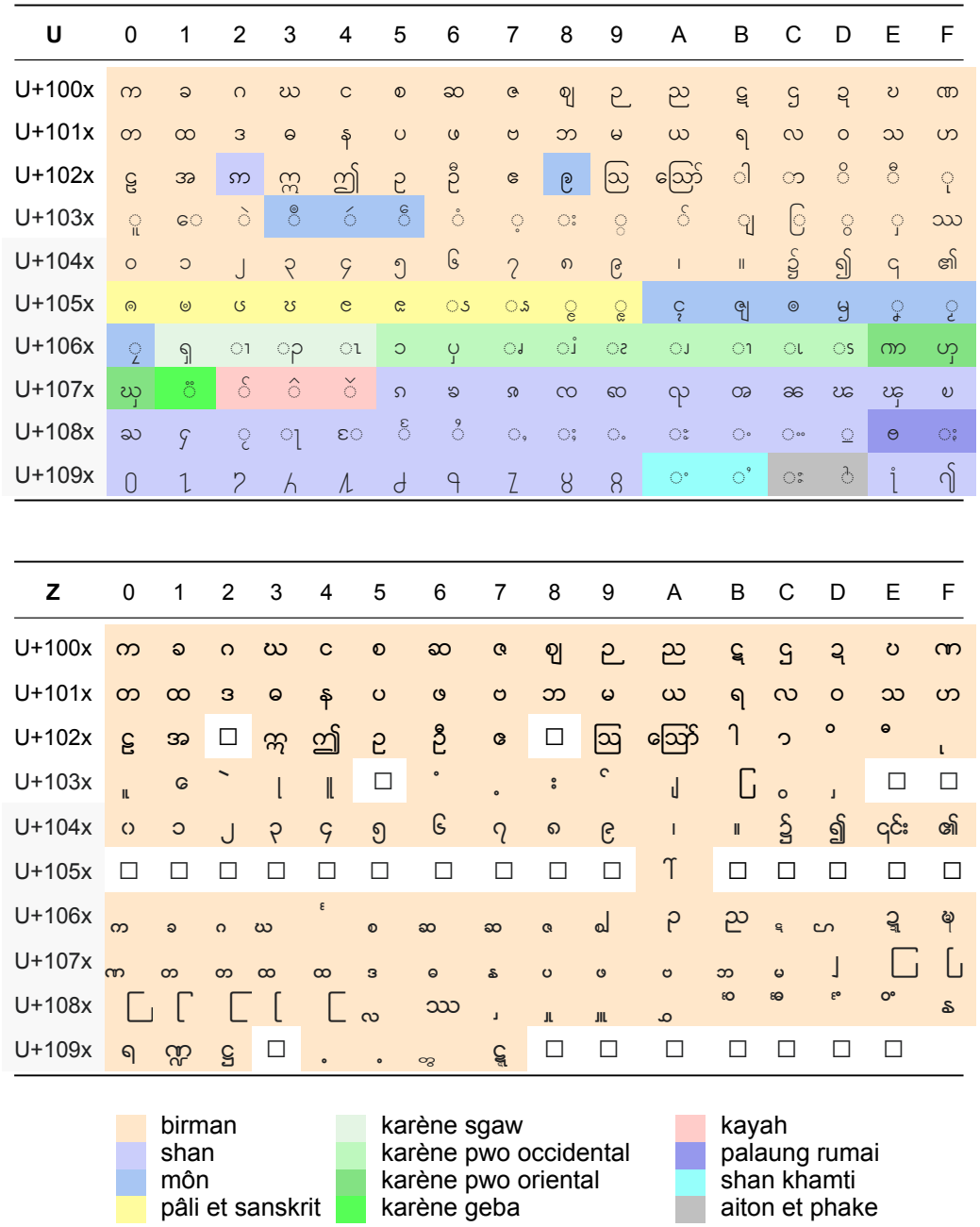


FIG. 1.10 : Les points de code du bloc Myanmar utilisés par l'Unicode (en haut) et le système Zawgyi-One (en bas) pour l'encodage du birman.

L'encodage Zawgyi-One est en fait un encodage de glyphes, comme les encodages à base d'ASCII pré-Unicode. Il diffère de l'encodage Ayar, qui est un encodage de caractères comme Unicode, mais pas tout à fait conforme à celui-ci. Le système d'encodage utilisait les mêmes points de code qu'Unicode (voir figure 1.9), mais sans en respecter l'ordre de stockage. Autrefois, certains logiciels (de graphisme notamment) ne gérant pas tous les aspects d'un affichage complexe Unicode, cela a inspiré la création de polices *Ayar*. L'exemple 47 en encodage Ayar utilise la police *ayar.ttf*. On constate que pour le même résultat d'affichage que l'exemple Unicode, les caractères doivent être stockés dans un ordre différent.

(47) Ayar : ေကြာင် U+1031 U+103C U+1000 U+102C U+1004 U+103A
ေ ြ က ာ င ်

(48) Unicode : ေကြာင် U+1000 U+103C U+1031 U+102C U+1004 U+103A
က ြ ေ ာ င ်

Le système Ayar, qui se revendiquait une police Unicode, était source de confusion supplémentaire dans le paysage déjà accidenté de l'encodage birman. L'internaute était invité à télécharger une police supplémentaire qui coexistait mal avec la police *Zawgyi-One* sur des systèmes d'exploitation. Autrefois utilisé, entre autres, par le site d'actualités libres, la Democratic Voice of Burma, l'encodage Ayar a presque disparu de l'Internet, et n'ayant pas pu s'imposer comme standard, le système Ayar semble tombé en désuétude.

D'un point de vue pratique, la coexistence des systèmes Zawgyi-One et Unicode sur l'internet pose problème, car l'utilisateur est souvent confronté au phénomène de mojibake, texte brouillé suite au mauvais décodage. Cet exemple montre un texte encodé en Zawgyi-One, d'abord affiché avec la police Unicode *Padauk.ttf* comme exemple de mojibake, ensuite affiché correctement avec la police *Zawgyi-Onev4.ttf*.

(49) Encodage Zawgyi-One, police Unicode :
ျမစ္ဝကြၽန္းေပၚေဒသဟာ ျမန္မာႏိုင္ငံ ရဲ့ စပါးကို ေဒသ ျဖစ္ပါတယ္။
Encodage Zawgyi-One, police *Zawgyi-Onev4.ttf* :
မြစ်ဝကျွန်းပေါ်ဒေသဟာ မြန်မာနိုင်ငံ ရဲ့ စပါးကို ဒေသ ဖြစ်ပါတယ်။
La région du delta est le grenier de la Birmanie.

VOA Burmese

Ceci peut survenir pour plusieurs raisons. Souvent l'utilisateur n'a que des polices d'un seul système installées, et n'a pas donc accès aux textes encodés au moyen de l'autre système. Cela arrive souvent sur les réseaux sociaux. Pour palier ce problème certains internautes n'hésitent pas à poster le même texte deux fois dans deux encodages différents, ou poster une image du texte, pour éviter que le lecteur ne soit obligé d'avoir recourt à un convertisseur. Sur certains systèmes d'exploitation, l'installation de polices des deux systèmes d'encodage ne garantit pas un affichage correct si des logiciels comme des navigateurs ne sont pas configurés correctement, car comme nous l'avons expliqué ci-dessus, une police *Zawgyi-One* peut parfois s'imposer par défaut en vertu de sa taille supérieure à certaines polices Unicode. Les concepteurs de sites web, très conscients du problème, intègrent une police qui correspond à l'encodage du texte de leurs pages dans le site même, garantissant ainsi un affichage correct sur l'écran de l'utilisateur, quelles que soient les polices installées sur sa machine ou son appareil. Il ne s'agit pas d'un transcodage ou conversion d'encodage proprement dit, ce que constate l'utilisateur quand il copie-colle le texte ailleurs. S'il n'a pas la police nécessaire installée sur sa propre machine, le texte devient illisible.

D'une manière générale, les retombées négatives de la persistance de Zawgyi-One pour les birmanophones ne sont pas négligeables. Parfait comme police de machine à écrire (texte destiné à être imprimé), le système Zawgyi-One n'assure pas l'intégrité ou la perpétuité des données textuelles. D'abord, l'inconvénient majeur de systèmes utilisant l'ordre d'écriture pour le stockage et non pas l'ordre phonétique concerne le tri automatique, qui ne serait plus alphabétique. Ensuite, le fait de pouvoir entrer un même texte de façons différentes ne permet pas la récupération ou l'analyse fiable de données (requêtes de recherche, traitement statistique). Ceci a des conséquences négatives dans tous les domaines où l'usage de l'informatique est indispensable — commerce, gouvernement, éducation et ainsi de suite. De surcroît, l'utilisation par Zawgyi-One des points de code Unicode destinés aux autres langues crée un obstacle pour les locuteurs des langues minoritaires, qui souvent doivent se contenter de communiquer sur l'internet uniquement en birman.

1.5.3.3 Les inconvénients des polices Unicode

La relation entre police et encodage en Unicode se veut assez simple : la police se charge de l'affichage d'un système d'encodage. Dans la pratique, les utilisateurs sont guidés par l'affichage des caractères par les polices dans la saisie de texte, de la même manière que la correction automatique signale les erreurs. Quand l'utilisateur tape des séquences de caractères qui ne sont pas prévues par le système d'affichage de la police, il peut y avoir une correction automatique prise en charge par le clavier (actuellement peu commune) ; au minimum, il devrait y avoir un retour visuel qui indique l'erreur : des caractères mal formés, et comme nous l'avons évoqué, l'usage d'un cercle pointillé (comme pour l'exemple 35 ci-dessus). Toutes les polices Unicode pour le birman ne prévoient pas de retour visuel d'erreur dans les mêmes cas. Nos exemples ici utilisent les polices *Padauk* (*Padauk.ttf*), *Pyidaungsu* (*Pyidaungsu-1.8.2_Regular.ttf*[43]), *Noto Sans Myanmar* (*NotoSansMyanmarUI-Regular.ttf*), *Myanmar3* (*Myanmar3-2018.ttf*) et *Myanmar Census* (*myanmarcensus.ttf*, une ancienne version de *Pyidaungsu* toujours en circulation[44]).

Police	**Ordre canonique**	**Ordre non canonique**
Padauk	မ့်င်	မ့်င်
Pyidaungsu	မ့်င်	မ့်င်
Noto Sans Myanmar	မ့်င်	မ့်င်
Myanmar3	မ့်င်	မ့်င်
Myanmar Census	မ့်င်	မ့်င်

TAB. 1.8 : Affichage identique entre l'ordre canonique *aukmyit*+*asat* et l'ordre non canonique *aukmyit*+*asat* toutes polices confondues

Nous avons vu que les polices capables d'afficher du texte encodé selon les normes Unicode sont très souples concernant l'affichage de la combinaison de la marque du ton *aukmyit*, ့ ⟨[1]⟩, (U+1037 MYANMAR SIGN DOT BELOW) et du caractère virāma visible *asat*, ် ⟨[x]⟩, (U+103A MYANMAR SIGN ASAT) — l'ordre de

43. Les polices Pyidaungsu de la Myanmar Computer Federation (MCF) ainsi que la police Myanmar3 sont disponibles ici : https://www.mcf.org.mm/download.htm.

44. https://mcf.org.mm/myanmar-unicode/384-font-history.html consulté le 2 février 2019.

stockage de ces caractères doit en principe toujours être *aukmyit+asat* ◌့+◌် ⟨[1x]⟩, mais dans les faits les caractères s'affichent de façon identique, quel que soit l'ordre de stockage, canonique ou non canonique. Ceci est vrai pour toutes les polices Unicode que nous avons testées. Le tableau 1.8 en donne quelques exemples d'affichage identique pour မ့င် (⟨mǰŋ[1x]⟩ et ⟨mǰŋ[x1]⟩, *haut*), alors que l'on aurait pu s'attendre à မ့င်◌့ (⟨mǰŋ[x1]⟩) pour l'affichage de l'ordre non canonique.

De cette façon, toutes les polices confirment la validité des deux ordres avec un retour positif à l'affichage. Bien que peu pratique, ceci ne peut être considéré comme choquant, car le standard Unicode recommande dans ce cas précis de traiter les deux ordres comme équivalents. Ceci est le seul cas où ordre canonique et ordre non canonique devraient être traités comme identiques. Nous examinons quelques cas d'ordres non canoniques qui devraient être considérés comme des erreurs d'encodage. Toutes les polices considérées ici affichent correctement du texte encodé rigoureusement selon le standard Unicode, mais certaines affichent correctement aussi du texte mal encodé. Prenons d'abord l'ordre relatif de ေ◌ (U+1031 MYANMAR VOWEL SIGN E) que nous avons examiné dans l'exemple 35 ေရ (*eau*). Nous voyons dans le tableau 1.9 que seule la police *Myanmar Census* affiche l'ordre non canonique (U+1031 U+101B) de la même manière que l'ordre canonique (U+101B U+1031) et ne signale pas l'ordre non standard comme erroné au moyen du cercle pointillé.

Police	**Ordre canonique**	**Ordre non canonique**
Padauk	ေရ	ေ◌ရ
Pyidaungsu	ေရ	ေ◌ရ
Noto Sans Myanmar	ေရ	ေ◌ရ
Myanmar3	ေရ	ေ◌ရ
Myanmar Census	ေရ	ေရ

TAB. 1.9 : Affichages avec différentes polices des ordres canonique et non canonique du caractère ေ◌ MYANMAR VOWEL SIGN E.

Il est donc possible d'afficher du texte encodé selon le standard Unicode avec toutes ces polices, mais la police *Myanmar Census* est plus permissive, dans la mesure où elle affiche aussi du texte non conforme. Si l'utilisateur qui affiche

son texte avec *Myanmar Census* entre du texte dans l'ordre non canonique ၎၍, et envoie son texte à un autre utilisateur qui utilise une autre police telle que *Noto Sans Myanmar*, le destinateur verra ၎◌်၍. Cela a pour effet de ne pas encourager les utilisateurs à dactylographier selon l'ordre standard voire de manière uniforme.

Dans la même veine, si l'ordre canonique de ခို် ⟨kʰiu⟩ /kʰò/ (*pigeon* U+1001 U+102D U+102F) s'affiche correctement dans toutes les polices mentionnées ci-dessus, l'ordre non canonique ⟨kʰui⟩ (U+1001 U+102F U+102D) s'affiche ခု◌ိ် dans toutes ces polices sauf *Myanmar Census*, avec laquelle il s'affiche ခို့. Seul l'ajout d'un caractère, comme ◌း (U+1038 MYANMAR SIGN VISARGA pour écrire ခို်း /kʰó/, *voler, cambrioler*) permet de distinguer l'erreur dans l'ordre des caractères qui le précèdent ခို့◌း, ce qui est déconcertant pour l'utilisateur, qui voit une erreur intempestive difficile à élucider.

Le manque de rigueur de *Myanmar Census* pose en effet problème non seulement pour l'échange d'informations, mais aussi pour le traitement automatique, car des textes dactylographiés à l'aide de cette police présentent des erreurs souvent difficiles à détecter de manière systématique. Or, elle n'est pas la seule police qui présente ce genre de défaut. Un des défauts du système *Zawgyi-One* est que les utilisateurs ne font pas de distinction entre la lettre ဝ ⟨w⟩ /wa̰/ (U+101D MYANMAR LETTER WA) et le chiffre birman zéro ၀ (U+1040 MYANMAR DIGIT ZERO), bien que les glyphes en *Zawgyi-One* à l'affichage ne ressemblent pas totalement : ဝ et ၀. Les nouvelles polices Unicode devraient non seulement faire une distinction entre les glyphes, mais aussi ne pas permettre des combinaisons erronées avec le chiffre. De notre échantillon de polices, trois d'entre elles ne font pas de distinction perceptible entre les glyphes pour WA et pour ZÉRO : *Padauk* (WA : ဝ, ZÉRO : ဝ), *Noto Sans Myanmar* (WA : ဝ, ZÉRO : ဝ), et *Myanmar Census* (WA : ဝ, ZÉRO : ဝ). Seule la police *Pyidaungsu* distingue entre WA et ZÉRO dans l'analyse de la combinatoire de caractères. Par exemple, la combinaison du chiffre ZÉRO ၀ avec d'autres caractères dépendants ne devrait pas être permise en Unicode, comme l'exemple 50 de ZÉRO en combinaison avec le caractère ◌ိ (U+102D MYANMAR VOWEL SIGN I).

(50) *Pyidaungsu* : ၀◌ိ U+1040 U+102D
၀ ◌ိ

Toutes les autres polices affichent comme valide cette séquence erronée : *Padauk*

၀, *Noto Sans Myanmar* ၀, *Myanmar3* ၀ et *Myanmar Census* ၀. Ceci explique pourquoi la confusion entre le chiffre zéro et la lettre wa persiste dans les textes en birman.

Nous voyons donc que cette confusion entre caractères peut s'éviter en exploitant pleinement les propriétés des caractères, mais il faut aussi garder à l'esprit que l'affichage dépend aussi du logiciel utilisé. Par exemple, une faute de frappe courante, le double *asat* ်် ⟨ˣˣ⟩, qui apparaît en fin de syllabes comme ခြောင် (⟨kǰeaŋˣ⟩ /tɕàʊɴ/ *chat*), est invisible affichée en *Noto Sans Myanmar* ခြောင်, mais perceptible en *Pyidaungsu* ခြောင် ⟨kǰeaŋˣˣ⟩ (corrigée en ခြောင် ⟨kǰeaŋˣ⟩). Avec *Padauk*, dans un texte compilé avec XeLaTeX, la faute ခြောင် ⟨kǰeaŋˣˣ⟩ ne diffère pas visuellement de la forme corrigée ခြောင် ⟨kǰeaŋˣ⟩, mais dans LibreOffice Writer (et aussi dans certains éditeurs de textes) il s'affiche ခြောင်် ⟨kǰeaŋˣˣ⟩.

Dans l'attente de polices parfaites dédiées aux caractères du bloc Myanmar, il est nécessaire de bien connaître les caractéristiques des polices que l'on utilise, pour éviter les erreurs certes, mais aussi pour vérifier les textes. Lors de la correction manuelle de texte, il est parfois nécessaire de changer la police du texte à vérifier entre relectures.

1.6 Résumé

Ce chapitre introductif à la langue birmane et à sa numérisation, qui ne se veut pas exhaustif, se concentre sur les aspects utiles pour mieux appréhender l'exposition de notre travail. Nous avons vu que la langue birmane est peu apprise comme langue étrangère en dehors de l'Asie du Sud-est, mais elle occupe une place importante en tant que deuxième langue en Birmanie, et aussi gagne de l'importance comme langue étrangère dans les pays limitrophes comme la Thaïlande, surtout grâce au potentiel économique du pays. En abordant les caractéristiques de la langue elle-même, nous avons mis l'accent sur les principes de son système d'écriture, pour mieux cerner ensuite son encodage informatique. Avant d'exposer les vicissitudes de l'encodage et son implémentation par des polices de caractères, nous avons tenté de faire un résumé du système Unicode et d'éclaircir quelques notions clé, telles que glyphe, point de code et police. Au premier abord pédante, la distinction entre ces notions est essentielle à comprendre pour effectuer de manière appropriée le prétraitement d'un corpus

birman, car la confusion entre ces notions est responsable non seulement d'un bon nombre des fautes systématiques, mais aussi a été responsable de la mauvaise implémentation du système d'encodage du birman en Unicode, qui, comme nous avons vu, est tout à fait performant pour la création de corpus en birman.

Chapitre 2

Corpus et ressources birmans

> ဝမရှိဘဲ၊ ဝိမလုပ်နိုင်။
> Without the letter "Wa" in Burmese, the word "Wi" cannot be spelt.
> (Without a strong foundation, no structure can be built.)
> *Aucun édifice ne peut se construire sans fondations solides.*
>
> *Myanmar Proverbs in Myanmar and English* (Hla Thamein 2000)

La constitution de corpus en birman, d'une manière générale, se heurte à de nombreuses difficultés, la principale étant le recueil de données qui répondent à une hypothèse de travail spécifique. Dans le cas qui nous concerne, l'exploration de la fréquence lexicale de la langue en général, demanderait normalement un corpus de très grande taille qui contiendrait des échantillons de langue de différentes sortes. Les difficultés techniques font que la création d'un corpus de très grande envergure demanderait des moyens humains conséquents, car il faudrait non seulement trouver une variété de textes au format numérique, mais s'assurer de leur qualité de rédaction par une correction manuelle laborieuse. En effet, les hypothèses de travail doivent en quelque sorte s'inspirer des données disponibles. Compte tenu de la difficulté à collecter un large échantillonnage de textes de divers types en langue birmane et de la nécessité d'un prétraitement complexe, nous avons renoncé à la création d'un corpus de référence, c'est-à-dire, un corpus de très grande envergure qui représenterait la langue générale (Bowker et Pearson 2002). Biber (1993) soutenant que la représentativité d'un corpus, l'efficacité du corpus en tant qu'échantillon de la langue, dépend plus de la définition des textes choisis que de la taille du corpus, et que l'échantillonnage à partir de genres spécifiques est toujours plus représentatif que des échantillons tirés d'un grand corpus général, il n'est donc pas question d'amalgamer tous les textes que nous avons rencontrés. A la place, nous avons opté pour des corpus multiples de taille moins ambitieuse, afin de représenter la fréquence

lexicale de genres textuels spécifiques. Les genres textuels choisis sont des genres régulièrement rencontrés par les apprenants de langue étrangère (la presse, la littérature…). Ce choix méthodologique nous permet de focaliser sur les genres utiles pour des apprenants, mais aussi apercevoir des aspects de langue plus générale en comparant les différents corpus et les comparer à nos corpus didactiques. Tous ces corpus ont aussi nourri notre réflexion sur le traitement automatique du corpus birman, car nous voulons que les listes de fréquence qui en résultent soient compatibles avec les besoins d'un apprenant du birman langue étrangère.

Cette section sert principalement à introduire nos corpus birmans, à la fois les corpus que nous avons construits et les corpus que nous avons utilisés créés par d'autres. L'étude emploie deux catégories de corpus : des corpus de textes authentiques et des corpus de textes didactiques. Les premiers sont utilisés comme références pour calculer et examiner la fréquence lexicale de la langue birmane. La description de chaque corpus et les méthodes de recueil de données seront accompagnées de statistiques descriptives et d'une explication des prétraitements simples, alors que le prétraitement concernant la segmentation des textes birmans en tokens sera détaillé dans le chapitre suivant. Nous commençons cette partie par un aperçu de l'état de l'art des corpus du birman, avant de décrire nos corpus.

2.1 Etudes précédentes

Le birman contemporain est actuellement peu étudié, et jusqu'au présent les études à base de corpus qui visent spécifiquement le birman ont été peu nombreuses. Parmi les études linguistiques récentes faites sur de grands corpus, nous citons d'abord Hnin Tun (2006) et Hnin Tun (2013), deux thèses en linguistique à base de corpus de la langue orale dont le contenu a été transcrit dans un système exclusif de transcription et segmenté en syllabes pour permettre une analyse quantitative. Le corpus qui a servi pour la thèse de Vittrant (2004) ne concerne pas exclusivement la langue orale, mais aussi le style parlé du birman, car il contient des textes moins spontanés. Celui-ci dans Vittrant (2005), décrivant l'élaboration de son corpus, cite la taille par nombre de phrases, ce qui nous mène à croire que le corpus n'a pas subi de traitement quantitatif. En ce qui concerne le birman écrit, l'élaboration du *Burmese/Myanmar Dictionary*

of Grammatical Forms (Okell et Allott 2017) s'est faite à partir d'un corpus d'environ 750 000 tokens qui réunit des textes écrits de style écrit et parlé. Aucun de ces corpus mentionnés ci-dessus n'a vocation à être utilisé en TAL ou disponible dans un format facilement exploitable par les talistes. Seul le dernier est disponible en libre accès, mais en format brut sans métadonnées[1].

Il existe quelques corpus birmans soigneusement préparés disponibles via une interface en ligne uniquement. Très connu parmi les apprenants de birman, le site SEAlang[2], la Southeast Asian Languages Library, intègre plusieurs corpus. Ces corpus birmans différents (Southeast Asian Languages Library 2006), tous encodés en Unicode, sont accessibles via deux interfaces : une qui affiche les contextes d'un élément recherché et ses cooccurrences, *Corpus*, et une autre qui affiche des phrases alignées anglais-birman qui résultent d'une requête, *Bitexts*. Ces interfaces ne permettent pas d'extraire facilement de l'information sur la fréquence lexicale, fonctionnalité qui présupposerait une segmentation en tokens des textes. En fait, le contenu n'est pas segmenté à dessein, afin que le système n'ait pas à essayer de prédire la segmentation attendue par l'utilisateur. Face à l'inexistence de règles de segmentation prévisibles pour l'utilisateur, les auteurs du site ont décidé de privilégier la quantité de résultats sur la qualité, en utilisant ce qu'ils appellent *peephole segmentation* (segmentation à fenêtre). Le système cherche simplement les chaînes de caractères qui correspondent à celle de la requête de l'utilisateur. Ces requêtes se font sur des phrases entières, ce qui entraîne un rappel plus élevé et une précision plus faible. Les résultats de requêtes sur l'interface *Bitexts* consistent en des phrases isolées, dans l'ordre soit birman-anglais, soit anglais-birman, selon la langue de la requête, en taille croissante. Le mode de présentation des résultats SEAlang ne permet pas de voir un contexte large, se limitant à la phrase pour des requêtes avec *Bitexts* et à une fenêtre d'une trentaine de caractères de part et d'autre de la chaîne de caractères recherchée pour les requêtes avec *Corpus*. Le contenu de *Bitexts* provient des ressources birmanes du SAY Project de la New Mexico State University Computing Research Laboratory. Le contenu de *Corpus* est composé d'échantillons de textes birmans publiés sur l'internet, divisés en *Misc text* (textes divers, d'environ seize millions de caractères) et *Literature* (textes littéraires, cinq millions de caractères). Ces corpus sont exclusivement à usage didactique.

1. https://doi.org/10.5281/zenodo.1202324 (Okell 2018)
2. http://sealang.net/library/

Dans le domaine du traitement automatique, les projets de recherche sur le birman s'appuient souvent sur des corpus propriétaires créés spécifiquement pour l'étude en question. A titre d'exemple, un corpus parallèle anglais-birman de 45 963 mots a été créé comme ressource de support au travail de Nyein Thwet Thwet Aung et Ni Lar Thein (2011) sur un système de désambiguïsation de sens des mots pour la traduction automatique. Le corpus est segmenté en mots, mais le système de segmentation employé n'est pas décrit. Hla Hla Htay et Narayana Murthy (2008) ont développé pour les besoins de leur recherche en traitement automatique du birman de corpus de grande envergure, un corpus monolingue de plus de deux millions de phrases et un corpus bilingue anglais-birman de 80 000 phrases, les deux composés provenant de contenu de sites internet d'actualités, de magazines en ligne et de livres électroniques, tous re-encodés au format *WinInnwa*. Win Win Thant et al. (2011) ont construit un corpus d'environ 45 000 tokens constitué de manuels scolaires au niveau collège, de manuels de grammaire et de sites internet (non précisés) afin de mener des expériences sur l'étiquetage syntaxique. Deux corpus oraux avec transcription ont été créés pour la recherche en reconnaissance automatique de la parole. Le premier, de Aye Nyein Mon et al. (2017), vingt heures d'enregistrements à partir de nouvelles en ligne est composé de 7 332 énoncés, ne semble pas être librement disponible pour le recherche. Le deuxième de Yin May Oo et al. (2020), un corpus de 2 530 phrases enregistrées par des volontaires est librement disponible avec sa transcription[3]. Depuis son édition de 2018 les tâches de traduction automatique du Workshop on Asian Translation (WAT) comporte une tâche de traduction birman-anglais, basée sur le corpus ALT (Riza et al. 2016 ; Ding, Utiyama et al. 2018 ; Ding, Aye et al. 2019) (voir plus loin) et le *UCSY corpus* birman-anglais créé par des chercheurs du *NLP Lab*[4] à l'UCSY (University of Computer Studies, Yangon), corpus parallèle qui, dans son édition 2019, comporte 204 839 phrases non segmentées, provenant de domaines variés tels que des articles de presse et des manuels scolaires (Nakazawa et al. 2019). Ce corpus UCSY est disponible exclusivement pour les chercheurs qui participent à la tâche de traduction de l'atelier[5]. Pour cet atelier, les chercheurs ont intégré leurs expériences des corpus monolingues de grande envergure, par exemple Marie et al. (2019) et Wang, Sun

3. http://www.openslr.org/80/
4. http://www.nlpresearch-ucsy.edu.mm/
5. http://lotus.kuee.kyoto-u.ac.jp/WAT/my-en-data/

et al. (2019) ont utilisé des phrases en birman de Common Crawl[6] aspiré en avril 2018 et la totalité du Wikipédia birman téléchargé le 1er janvier 2017. Chen et al. (2019) quant à eux, ont utilisé une partie du Common Crawl Myanmar combiné avec le jeu de données de Buck et al. (2014). Ces grands corpus sont disponibles en accès libre, mais dans un format brut, parfois avec des encodages multiples, très bruités ou sans segmentation, parce que pour la plupart ils font partie de corpus multilingues, qui ne prennent pas en compte les particularités du traitement du birman.

Le premier corpus d'envergure de ce type à apparaître en accès libre, semble être celui de Christodouloupoulos et Steedman (2014), un corpus parallèle disponible en libre accès[7], au format XML, construit à partir de traductions de la Bible en entier dans cent langues, y compris le birman, encodé en Unicode. Les alignements sont au niveau des versets, 30 928 en tout. Le texte birman n'est pas (encore) segmenté en tokens. La source du texte birman n'est pas spécifiée, mais nous avons identifié la version comme étant la traduction du missionnaire Adoniram Judson de 1840, moins représentative du birman moderne que les traductions les plus récentes.

Celui-ci n'est pas le seul corpus multilingue à inclure le birman. Le site de corpus parallèles du projet OPUS (Tiedemann 2012) recense quatre corpus parallèles qui incluent le birman. Tous les textes sont segmentés et balisés au niveau de la phrase et du mot, les fichiers séparés d'alignement permettent l'alignement. Deux d'entre eux concernent des traductions d'interfaces de systèmes d'exploitation informatiques (Ubuntu v14.10 et GNOME v1) dont l'intérêt est limité, car outre, son domaine très spécifique, tant leurs alignements que leur segmentation du birman semble ne pas avoir été vérifiés, d'autant que l'on y observe une sur-segmentation surprenante, par exemple ဖိုင်များ /pʰàɪnmjá/ ⟨pʰiuŋˣmy̆aˀ⟩ (*files*) se découpe en trois tokens : ဖို, ိင် et များ (⟨pʰiuŋ⟩, ⟨ˣ⟩ et ⟨my̆aˀ⟩). Ces défauts rendent ces corpus inexploitables. Il en va de même avec les deux versions du Global Voices Parallel Corpus sur OPUS, dont la première version de 2015 possède les mêmes défauts que les deux corpus précédents. Sa dernière version datant de 2017 est censée aussi être alignée par *mots*, mais la segmentation reste très inégale. La plupart du corpus est sous-segmenté, mais l'on y observe aussi des sur-segmentations, par exemple အင်္တာနက် /ʔɪ̀ɴtànɛʔ/

6. Un corpus libre et gratuit d'exploration Web https://commoncrawl.org/
7. http://christos-c.com/bible/

(*internet*) est scindé en trois tokens chacun balisé comme étant un mot. Le corpus d'origine de Global Voices[8] est un corpus parallèle de textes d'actualités de la blogosphère mondiale tirés du site de l'ONG du même nom. Une cinquantaine de langues sont concernées, mais seules 2426 phrases sont alignées par phrase avec les textes birmans originaux (sans balisage des *mots*, donc sans segmentation en tokens). Ce corpus, qui a été aligné automatiquement sans vérification manuelle (Braune et Fraser 2010) présente peu d'erreurs d'alignement. Le corpus OSCAR [9] (Ortiz Suárez et al. 2019) est un corpus multilingue de grande taille obtenu par identification de langue et filtrage du contenu du corpus Common Crawl. Les données sont distribuées au format brut par langue et au format dédupliqué. Le corpus dédupliqué[10] de 1,1 Go est composé de textes en encodage mixte (Zawgyi et Unicode), et n'est pas segmenté. Le corpus c4 (Raffel et al. 2019) avec 813 530 échantillons[11] semble présenter les mêmes inconvénients. Nous avons aussi trouvé les mêmes problèmes avec un autre corpus basé sur des données de Common Crawl, le corpus CCAligned[12] (El-Kishky et al. 2020), la partie birmane nous semblant particulièrement bruitée par du contenu dans d'autres langues. Nous avons constaté aussi des problèmes d'encodage dans le corpus Paracrawl Burmese (Esplà-Gomis et al. 2019)[13], un corpus de 31 374 phrases non segmenté en tokens. Un corpus plus modeste, mais de meilleure qualité, le corpus FLORES-200 (Goyal et al. 2021)[14], est composé de 3001 phrases (non-segmentées) extraites de pages diverses du Wikipédia anglais et traduites en 101 langues, dont le birman. Un corpus traduit donc, et sans segmentation en tokens.

Deux corpus segmentés (et étiquetés avec des étiquettes morpho-syntaxiques), le *Asian Language Treebank* (Riza et al. 2016 ; Ding, Utiyama et al. 2018 ; Ding, Aye et al. 2019) et *myPOS* (Khin War War Htike et al. 2017), ont été mis à disposition de la communauté scientifique. Ces corpus sont détaillés dans la section suivante qui introduit les corpus et ressources utilisés dans le présent travail.

8. Global Voices Parallel Corpus https://metatext.io/datasets/global-voices-parallel-corpus
9. (Open Super-large Crawled ALMAnaCH coRpus) https://oscar-corpus.com/
10. téléchargé le 20 décembre 2020
11. www.tensorflow.org/datasets/catalog/c4
12. http://www.statmt.org/cc-aligned/
13. https://paracrawl.eu/ mis à jour en avril 2021.
14. https://github.com/facebookresearch/flores/tree/main/flores200

2.2 Corpus de textes authentiques

Si notre choix de textes pour les corpus authentiques que nous avons constitués nous-mêmes a été grandement influencé par l'accessibilité de données, notre premier critère a été l'utilité du contenu pour un apprenant de niveau avancé. Il aurait été inutile de prendre un genre textuel qu'un apprenant ne va que rarement rencontrer dans la vie réelle. Ainsi, nous avons choisi des textes de la presse écrite, des textes littéraires et des sous-titres de conférences informelles, qui, bien qu'à la forme écrite, sont représentatifs de la langue parlée soignée. Un problème potentiel réside dans le fait que ces textes sont souvent des traductions (voir tableau 2.1), ce qui pourrait avoir un effet sur l'authenticité de la langue utilisée (voir tableau 2.1). Nous justifions notre choix par l'omniprésence de textes traduits en birman depuis l'anglais, et proposons une comparaison entre les textes du corpus ALT (section 2.2.1) qui sont tous des traductions, et les textes du corpus BBC Burmese (section 2.3), qui sont censés être (au moins pour la plupart) écrits en birman.

Nous utilisons ces corpus essentiellement pour créer des listes de fréquence spécifiques à chaque corpus, afin de les comparer entre eux et avec la fréquence du vocabulaire des corpus didactiques (décrits dans la section 2.3).

2.2.1 *Corpus ALT Wikinews Myanmar*

Nous nous intéressons uniquement au volet birman du Asian Language Treebank (ci-après appelé ALT), un corpus ouvert parallèle pour dix langues d'Asie (Riza et al. 2016 ; Ding, Utiyama et al. 2018 ; Ding, Aye et al. 2019) mis à disposition en ligne[15] sous la licence Creative Commons[16] par le National Institute of Information and Communications Technology au Japon (NICT). Il s'agit de contenu de pages en anglais du site Wikinews[17], site de journalisme collaboratif, traduit en birman (Utiyama et Sumita 2015) au style écrit et annoté par des membres du NLP Lab, University of Computer Studies, Yangon (UCSY). La taille en tokens du corpus est donné comme 866 593 tokens et le corpus est originellement encodé en Unicode.

Chacune des 20 106 phrases est accompagnée par un identifiant unique qui

15. http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/
16. Attribution 4.0 International (CC BY 4.0)
17. https://en.wikinews.org

Corpus	Contenu	Langue d'origine
ALT Wikinews	presse	anglais
myPOS Wikipédia	encyclopédie	birman
BBC Burmese	presse	birman/anglais
Lotaya Litterature	littérature	birman
TED Talk Subtitles	conférences	anglais+
BBE Burmese by Ear	didactique	birman
MdB1 Manuel de birman 1	didactique	birman
MFG Myanmar Flower Grammar	didactique	birman
CB Colloquial Burmese	didactique	birman
AiB Advancing in Burmese	didactique	birman
MNR Myanmar Newspaper Reader	didactique	birman
MLT Vocabulary	didactique	birman
SS Seasite Lessons	didactique	birman
TUFSD TUFS Dialogues	didactique	birman
TALPCo TALPCo/TUFS Vocabulary	didactique	japonais

TAB. 2.1 : Corpus et ressources birmans utilisés

permet de retrouver non seulement l'URL de la page originale anglaise, mais aussi l'ordre des phrases dans le texte. L'encodage suit les normes Unicode, y compris l'ordre du caractère ့ (U+1037, nom Unicode MYANMAR SIGN DOT BELOW, appelé par son nom birman *aukmyit* par les créateurs du corpus) et *asat* ် (U+103A), normalisé en U+1037 U+103A (*aukmyit asat*) suivant Hosken (2012). Pour prendre en charge les parenthèses typographiques, **)** et **(**, celles-ci sont remplacées par -RRB- et -LRB- respectivement[18].

Deux versions sont fournies : *Myanmar ALT with Basic POS Tags* et *Myanmar ALT with NOVA*. La première comporte une version segmentée avec des étiquettes morphosyntaxiques. Ces étiquettes son basées sur les parties du discours employées dans le *Myanmar-English Dictionary* မြန်မာ–အင်္ဂလိပ် အဘိဓာန် de la MLC, la Myanmar Language Commission (1993). Le tableau 2.2 détaille les étiquettes morphosyntaxiques des corpus que nous avons utilisés en comparaison avec les catégories employées dans le *Myanmar-English Dictionary* မြန်မာ–အင်္ဂလိပ် အဘိဓာန် et le jeu d'étiquettes morphosyntaxiques de Universal Dependancies (Nivre et al. 2017). *Myanmar ALT with NOVA* possède un jeu plus réduit d'étiquettes : **n** (nom), **v** (verbe), **a** (adjectif), **o** (autre modificateur), **1** (chiffre), . (ponctuation)

18. RRB = ***R**ight **R**ound **B**racket*, LRB = ***L**eft **R**ound **B**racket*

et + (token avec rôle syntaxique indéfini). Un guide d'annotation (Ding, Hnin Thu Zar Aye et al. 2016) donne des exemples d'usage de ce jeu d'étiquettes et le système d'association de tokens avec des crochets. Une tentative a été faite pour réconcilier ces étiquettes NOVA avec les étiquettes Universal Dependencies par Su Su Yee et al. (2019), mais cette implémentation présente un certain nombre d'incohérences (les démonstratifs peuvent être étiquetés ADJ ou DET) et semble également influencée par les parties du discours dans le *Myanmar-English Dictionary* မြန်မာ–အင်္ဂလိပ် အဘိဓာန်. Le nouveau jeu d'étiquettes ne fait pas usage de certaines étiquettes disponibles dans le jeu de Universal Dependencies telles que SCONJ pour les subordonnants, AUX pour les auxiliaires (catégories décrites par Bernot, Cardinaud et al. (2001), par exemple).

2.2.2 *Corpus myPOS (Myanmar Part-of-Speech)*

De son nom entier *myPOS (Myanmar Part-of-Speech) Corpus for Myanmar NLP Research and Developments* (Khin War War Htike et al. 2017), est un corpus spécialement constitué pour les besoins en traitement automatique du birman et mis à disposition[19] sous licence Creative Commons[20]. Le corpus contient environ 11 000 phrases extraites des pages birmanes de l'encyclopédie en ligne Wikipédia[21] qui sont composées en style écrit. Selon la description fournie, le corpus est composé de 264 920 *mots* ou 242 865 *mots composés* segmentés et étiquetés à la main, avec le jeu d'étiquettes morphosyntaxiques spécifié dans le tableau 2.2, système qui diffère légèrement du système de *Basic POS-tags* utilisé pour ALT. Le système d'annotation de *myPOS* prend en compte les composés, la jonction entre morphèmes formant un composé étant indiquée par une barre verticale, | . Par exemple, la chaîne de caractères suivante အစားထိုးမှု /ʔəsátʰóʊ̯mʊ̯/ (*substitution*) est annotée ainsi :

(51) အစားထိုး/v|မှု/part

Les créateurs du corpus incluent aussi une liste intéressante des 8627 mots composés uniques qui apparaît dans le corpus. Le corpus (et son contenu d'origine) est originellement encodé en Unicode, respectant à la lettre les normes Unicode. Ce corpus nous sert surtout pour l'évaluation de notre outil de segmentation de

19. https://github.com/ye-kyaw-thu/myPOS
20. Attribution-NonCommercial-Share Alike 4.0 International (CC BY-NC-SA 4.0)
21. ဝီကီပီးဒီးယား /wìkìpídíjá/ https://my.wikipedia.org/

Universal POS-tags	**MLC**	**ALT**	**myPOS**
			abb (abbreviation)
ADJ (adjective)	adjective	ADJ	adj
ADP (adposition)	post-positional marker	PPM	ppm
ADV (adverb)	adverb	ADV	adv
AUX (auxiliary)			
CCONJ (coordinating) conjunction	conjunction	CONJ	conj
DET (determiner)			
		FOR (foreign etymology)	fw (foreign script)
INTJ (interjection)	interjection		int
NOUN (noun)	noun	N	n
NUM (numeral)		NUM	num
			tn (text number)
PART (particle)	particle	PART	part
			part_neg (negative particle)
PRON (pronoun)	pronoun	PRON	pron
PROPN (proper noun)			n
PUNCT (punctuation)	punctuation	PUNC	punc
SCONJ (subordinating conjunction)			
SYM (symbol)			sb
VERB (verb)	verb	V	v
X (other)		?	

TAB. 2.2 : Systèmes d'étiquettes morphosyntaxiques

Rubique en birman	Equivalent anglais
မြန်မာ့ရေးရာ /mjànmàjéjà/	Burma
နိုင်ငံတကာ /nàɪnŋànta̰kà/	World
ဆောင်းပါး /sʰábʊnpá/	In Depth
အားကစား /ʔáka̰sá/	Sport
ကုန်သွယ်စီးပွား /kòʊnθwɛ̀sípwá/	Economy
ဓာတ်ပုံများ /daʔpòʊnmjá/	Photo Galleries

TAB. 2.3 : Rubriques du corpus BBC

textes en unités lexicales[22].

2.2.3 *Corpus de presse BBC Burmese*

Ce corpus est composé d'articles du site de BBC Burmese ဘီဘီစီ မြန်မာ /bìbìsì mjànmà/[23] qui sont rédigés en style parlé, que nous avons téléchargés par un aspirateur de site en 2016. Le site est originellement encodé en Unicode. Il s'agit essentiellement d'articles de l'année 2016 (l'aspirateur a téléchargé aussi des pages d'archives), 2144 pages HTML, qui ont été nettoyées du balisage par nos scripts Perl après l'extraction des métadonnées (URL, date, rubrique). Le contenu textuel est stocké en format texte, puis segmenté par l'outil de segmentation *Motor* détaillé dans le chapitre suivant. La taille moyenne de chaque article est de 403 syllabes (environ 267 tokens) avec une taille très variable, allant de moins de cent syllabes pour des pages qui n'ont que les légendes de photos jusqu'à 5026 syllabes pour un article approfondi. Les noms des rubriques que nous avons gardées (avec leurs équivalents en anglais) sont listés dans le tableau. Les rubriques အသံဖိုင်များ /ʔa̰θànpʰàɪnmjá/ (*audio*) et ဗီဒီယိုများ /bìdìjòmjá/ (*videos*) ont été volontairement écartées, car elles ont peu de contenu textuel.

Nous avons parfois remarqué que l'article birman est parfois une traduction d'un article originellement en anglais sans que cela soit mentionné dans l'article. Ce phénomène semble se limiter à une poignée de textes, un nombre suffisamment limité pour que nous puissions considérer le corpus dans son ensemble d'être composé de textes écrits en birman.

22. La version 3.0 du corpus myPOS a été publiée en juin 2021. L'augmentation de la taille du corpus est décrite par Hlaing, Thu et al. (2020)

23. https://www.bbc.com/burmese

2.2.4 *Corpus littéraire Lotaya*

Bien qu'il existe des sites et pages de Facebook partageant de grandes quantités de livres en birman au format électronique[24], l'histoire mouvementée de l'informatisation de la langue fait que le format préféré de partage est devenu le format PDF, qui ne nécessite pas de système de décodage de caractères spécifique pour être visualisé par le lecteur. Face à la panoplie de systèmes d'encodages des polices historiquement utilisées pour le birman, et le fait que bien souvent, les fichiers PDF contiennent des images de pages plutôt que les caractères des textes, les perspectives d'extraction automatique massive de textes à partir de fichiers PDF ne sont pas encourageantes, d'autant plus que la qualité de ces images, souvent des photographies ou photocopies scannées de livres au papier dégradé, nécessiterait l'utilisation d'un système de reconnaissance optique de caractères (OCR) très performant. Or, à l'heure actuelle, il n'existe pas de tel système conçu spécifiquement pour le birman qui soit capable d'extraire du texte à partir d'images de mauvaise qualité[25] Ce manque d'océrisation est bien décevant pour les créateurs de corpus, car de prime abord, on pourrait croire que la disponibilité d'un grand nombre de textes[26], faciliterait la création d'un corpus littéraire de grande taille.

Notre corpus bien plus modeste est composé de textes littéraires écrits majoritairement en style écrit, des chapitres de romans et des nouvelles d'auteurs divers, téléchargés en octobre 2017 du site လိုတရ[27] /lòtə̰jə̰/, qui ne sont plus disponibles en ligne. Le choix de site comme contenu de corpus a été motivé par l'homogénéité de l'encodage des textes et la taille similaire des textes. Les pages ont été aspirées au moyen d'un aspirateur de site écrit en Perl, avec extraction de métadonnées concernant l'auteur et le titre de chaque texte. Le contenu textuel total comprend 841 886 tokens, la taille moyenne de chaque texte étant de 2854 tokens. Ce contenu (texte et métadonnées), encodé entièrement

24. Comme par exemple *Free Myanmar Books Download* sur Facebook, https://www.facebook.com/mmebooks/, ou le site *Myanmar Book Download*, http://www.mmbookdownload.com/

25. Toutefois, la fonctionnalité d'océrisation du birman fournit par Google Cloud Services, bien qu'au stade expérimental, est assez performant sur les images nettes. Nous avons utilisé ce service de façon limitée à l'intérieur de *Google Drive*, activé à l'ouverture d'un fichier au format PDF avec *Google Documents*.

26. Par exemple, *Myanmar Books Download* recense plus de cinq mille romans et plus de sept cents magazines.

27. http://novels.lotayamm.com

par le système Zawgyi, a été ensuite transcodé en Unicode à l'aide de l'outil *burmeseunicode* (Lwin Moe 2016). Les titres et auteurs de chacun des 295 textes sont listés en birman (le tableau C.1) en annexe.

2.2.5 *Corpus de sous-titres de conférences TED Talks*

Le contenu de ce corpus, des sous-titres birmans de TED Talks (conférences TED) a été téléchargé du site Amara[28], une plateforme d'hébergement et édition libres de sous-titres. Les TED Talks sont une série de conférences courtes (moins de dix-huit minutes) sur des sujets divers organisées partout dans le monde par une fondation à but non lucratif appelée *The Sapling Foundation* et diffusées sur l'internet[29]. La plateforme héberge des sous-titres dans une centaine de langues, traduits par des bénévoles de la langue d'origine de la conférence ou traduits à partir de la traduction anglaise. La langue originale de l'écrasante majorité des 713 textes traduits en birman que nous avons téléchargés est l'anglais ; trois sont traduits de l'allemand, deux du français, deux du japonais et un texte chacun de l'arménien, de l'italien, de l'espagnol et du russe. La liste des titres originaux de ces conférences est fournie en annexe C.2. Trois textes sont composés de sous-titres de conférences présentées en birman (bien que les titres soient en anglais). Toutes les traductions en birman sont en style parlé.

La plateforme Amara permet de télécharger les fichiers de sous-titres dans plusieurs formats, notamment le format XML DFXP et le format texte TXT. Nous avons écrit un aspirateur de site pour télécharger tous les fichiers voulus en une seule fois. Les fichiers TXT ont été utilisés pour notre corpus, les fichiers DFXP servent à nous orienter entre les traductions lors des étapes de correction manuelle et d'identification des entités nommées. La taille moyenne de chaque texte birman est de 1618 tokens.

Le contenu du corpus est originellement encodé en Unicode, mais il a été nécessaire de décoder les chaînes de caractères des noms de fichiers encodé en ASCII *percent-encoding* vers l'encodage Unicode standard lisible (voir section 2.5.3.3 pour plus de détails).

28. https://amara.org
29. https://www.ted.com

2.3 Corpus et ressources didactiques

Ces corpus servent comme matière d'étude de la langue birmane telle qu'elle est enseignée en tant que langue étrangère. Au lieu de nous appuyer uniquement sur les études linguistiques et lexicographiques, nous présupposons que l'expérience des enseignants dans l'enseignement de langue, et la réflexion nécessaire à la préparation de supports pédagogiques que ces enseignants ont menées doivent fournir le socle de toute analyse de la langue qui porte sur l'apprentissage.

Ensemble, ces corpus ne constituent pas un corpus exhaustif du birman didactique, car nous n'avons pas inclus tous les manuels imprimés disponibles. Notre choix est motivé en premier lieu par la facilité de traitement automatique. A ce titre, nous avons privilégié les textes qui étaient déjà au format numérique (sites Web affiliés à des institutions d'enseignement), ou pourraient en être transformés sans recourt à la dactylographie (autrement dit, des textes au format PDF créé à partir de texte et non pas à partir d'images et donc transformable en texte). En dépit de cette restriction, nous avons toutefois inclus cinq manuels de langue de forme plus classique, *Burmese by Ear : or Essential Myanmar* (Okell 2014), *Manuel de birman : Langue de Myanmar Volume 1* (Bernot, Cardinaud et al. 2010), คู่มือการสอนภาษาพม่าแนวปริชาน *[Burmese Language Teaching Guide] Myanmar Flower Grammar* (Niyomtham et al. 2017), *Colloquial Burmese* (Hnin Tun et McCormick 2015), *Advancing in Burmese : A Drill Book for Intermediate to Advanced Learners* (Yadana Aung 2020), et aussi l'ouvrage de Luzoe (1996), *Myanmar Newspaper Reader*. Notre deuxième motivation a été d'examiner des supports destinés à des apprenants de langues maternelles différentes, pour tenter d'élucider si la langue maternelle des apprenants influence la façon de présenter le vocabulaire. Si la plupart des ressources du birman disponibles pour l'apprentissage du birman ont été destinés aux apprenants anglophones, il existe un nombre croissant de ressources destinées aux apprenants de langue autre que l'anglais, dont nous avons pu inclure certaines dans notre étude, les ouvrages plus traditionnels en français (Bernot, Cardinaud et al. 2010) et en thaï (Niyomtham et al. 2017) cités ci-dessus, et également les supports web destinés aux apprenants japonais (du *Tokyo University of Foreign Studies*). Cette diversité de formats et de sources, nous encourage à croire que nos corpus peuvent prétendre à une certaine représentativité du domaine tel qu'il existe actuellement. Néanmoins, la grande faiblesse de nos corpus didactiques provient de la paucité de textes aux

Code corpus	Nom corpus	Textes	Vocabulaire
BBE	Burmese by Ear	—	glossaire
MdB1	Manuel de birman 1	—	glossaire
MFG	Myanmar Flower Grammar	60 leçons	glossaires
CB	Colloquial Burmese	16 leçons	glossaire
AiB	Advancing in Burmese	34 chapitres	glossaire
MNR	Myanmar Newspaper Reader	52 textes	glossaires
MLT	Myanmar Language Test	—	5 listes
SS	Seasite Lessons	niveau débutant : 23 leçons	glossaires
		niveau intermédiaire : 16 leçons	glossaires
		littérature : 14 textes	—
TUFSD	TUFS Dialogues	40 leçons	balisage lexical
TALPCo	TALPCo/TUFS Vocabulary	1372 phrases isolées	segmentation lexical

TAB. 2.4 : Corpus et ressources didactiques

niveaux plus avancés. Seuls le récent manuel *Advancing in Burmese : A Drill Book for Intermediate to Advanced Learners* (Yadana Aung 2020) et l'ouvrage de Luzoe (1996) sont véritablement destinés aux apprenants de niveau intermédiaire et avancé. En effet, l'apprentissage du birman souffre du manque de supports pédagogiques pour les niveaux supérieurs. C'est pour cette raison, que nous considérons que cette étude de la langue birmane telle qu'elle est présentée aux apprenants ne peut qu'être considérée comme préliminaire.

Notre étude porte à la fois sur la présentation du vocabulaire (les modèles de structure de lexique à partir de formes lexicales et parties de discours) et la fréquence de ce vocabulaire dans les leçons. Certaines sources nous ont fourni les deux, liste(s) de vocabulaire et texte(s), d'autres consistent en une seule ressource (voir tableau 2.4).

2.3.1 *Vocabulaire BBE du manuel* Burmese by Ear

Cette liste de 1166 unités provient de la méthode de birman pour débutants anglophones *Burmese by Ear, or Essential Myanmar* (Okell 2014) . La méthode est composée de fichiers audio, accompagnés par un livret en PDF[30]. Notre ressource

30. Autrefois téléchargeables gratuitement sur le site de la SOAS, actuellement ils sont toujours disponibles sur demande via le site Bamazaga https://lukecorbin.org/bamazaga/

inclut le vocabulaire présenté en annexe du livre, dont le texte a été extrait et convertit depuis l'encodage AvaLaser en Unicode (voir plus loin). Nous avons pris le vocabulaire des seules annexes *Topical Vocabulaires* et *General Vocabulary English-Burmese*, éliminé les doublons et les phrases exemples. Pour plus de cohérence, nous avons considéré les unités verbales sans la syllabe တယ် /tɛ̀/ ⟨ty^{x}⟩, qui est caractérisé comme suffixe de fin de phrase dans l'annexe grammaire de la méthode.

2.3.2 *Vocabulaire MdB1 de* Manuel de birman, Volume 1

Le *Manuel de birman : Langue de Myanmar Volume 1* de Bernot, Cardinaud et al. (2010) a été pendant longtemps utilisé comme manuel de base pour les étudiants de birman à l'Inalco de premier niveau. Il est donc destiné aux apprenants francophones débutants et utilise exclusivement le style parlé. La reconnaissance optique de caractères étant peu performante sur la police birmane des vingt-six leçons, nous n'avons gardé que le vocabulaire birman du glossaire birman-français en annexe qui regroupe tout le vocabulaire de l'ouvrage. Pour ce faire, nous avons numérisé les pages du glossaire en images, ensuite le contenu textuel a été extrait au moyen de la fonction de reconnaissance de caractères à l'intérieur du traitement de texte en ligne *Google Docs*, suivi d'une correction manuelle pour produire une liste de vocabulaire de 860 unités.

2.3.3 *Vocabulaire et corpus MFG du manuel* Myanmar Flower Grammar

Le manuel pour débutants destiné aux apprenants thaïs de (Niyomtham et al. 2017), คู่มือการสอนภาษาพม่าแนวปริชาน /kʰûːmɯːkaːnsɔ̌ːnpʰaːsă:pʰámâːnɛːw-prìʔtɕʰaːn/ (*Guide d'enseignement de la langue birmane*), sous-titré joliment *Myanmar Flower Grammar* en référence aux schémas syntaxiques explicatifs dans chacun de ses soixante chapitres, ne se limite pas à une grammaire. Les soixante leçons du livre sont réparties en quatre sections : Vocabulaire, Phrases, Lecture et Dialogues, et au fur et à mesure des leçons la méthode propose des exercices de pratique d'écriture, d'appariement (de vocabulaire avec image, de questions et réponses), de reconstitution de phrases à partir de schémas et de pratique de l'oral. Le contenu est destiné aux débutants avec des textes simples tirés des manuels scolaires birmans de niveau primaire. Tous les textes utilisent exclusi-

Abbr.	POS	Traduction	Exemples
ก.	กริยา /kàrìjaː/	verbe	စား လာ ရှိ
กช.	กริยาช่วย /kàrìjaːtɕʰûaj/	verbe auxiliaire	နေ ချင် ထား
ว.	วิเศษณ์ /wísèːt/	adverbe et adjectif	ခဏခဏ သေသေချာချာ ပူ အေး မြန်
น.	นาม /naːm/	nom	မုန့် ငွေ သူ
ส.	สรรพนาม /sàppʰanaːm/	pronom	ကျွန်တော်
ล.	ลักษณนาม /làksànànaːm/	classificateur	ခွက် ကောင် ခု
ภ.	วิภัตติ /wípʰát/	suffixe, postposition	ကို လား မယ် နော် ဘယ်သူ...လဲ တွေ
สัน.	สันธาน /sănʰtʰaːn/	conjonction	ဒါပေမဲ့ ပြီးတော့ ဒါဆို
อุ.	อุทาน /ùtʰaːn/	interjection	ရော့

TAB. 2.5 : Étiquettes morphosyntaxiques thaïes employées par Niyomtham et al. (2017)

vement le style parlé. L'explication en thaï des parties de discours birmanes se sert des étiquettes morphosyntaxiques exposées dans le tableau 2.5.

Nous avons créé une liste de vocabulaire à partir des glossaires de vocabulaire des leçons (572 unités lexicales), mais ces glossaires sont aussi intégrés à notre corpus MFG qui comprend aussi des phrases et textes, schémas grammaticaux et exercices. L'extraction du texte s'est fait de la même manière que le manuel précédent : numérisation de pages en images et reconnaissance de caractères à l'intérieur du traitement de *Google Docs*, suivi d'une correction manuelle.

2.3.4 *Vocabulaire et corpus CB du manuel* Colloquial Burmese

Cette ressource est composée du vocabulaire et textes birmans du manuel de birman destiné aux apprenants débutants *Colloquial Burmese* (Hnin Tun et McCormick 2015) pour apprenants anglophones. Le style de langue est la langue parlée. Le manuel est divisé en seize leçons (quinze, plus une leçon d'introduction), qui sont composées de dialogues, d'exercices écrits et audio, de listes de vocabulaire, d'explications concernant la langue (*Language Point*) et la culture.

La liste de vocabulaire du manuel que nous avons utilisée est composée du glossaire et des listes thématiques en fin d'ouvrage : les points cardinaux, les jours de la semaine, des noms propres, les interrogatifs, les pronoms, des nombres cardinaux (en chiffres et en toutes lettres) et des nombres ordinaux.

Tout comme les textes des quinze leçons, le vocabulaire du glossaire en fin d'ouvrage a été transcodé depuis l'encodage AvaLaser en Unicode par la méthode

détaillée plus loin. Nous avons pris soin de ne pas éliminer les homographes listés avec des sens distincts.

2.3.5 *Vocabulaire et corpus AiB du manuel* Advancing in Burmese

Le manuel *Advancing in Burmese : A Drill Book for Intermediate to Advanced Learners* de (Yadana Aung 2020) se présente comme un exemplier de formes grammaticales du birman, destiné à améliorer les compétences en compréhension (surtout orale) du style parlé des apprenants de niveaux intermédiaires et avancés. Chacun des trente-quatre chapitres fournit une quinzaine d'exemples pour une forme donnée en birman et en traduction anglaise, suivie d'un glossaire et quelques notes explicatives, souvent culturelles. Le texte que nous avons reçu de l'auteure étant encodé en Unicode, le prétraitement n'a concerné que la vérification de respect des normes Unicode, avant d'extraire le vocabulaire du glossaire (873 unités lexicales), et le texte entier comme corpus AiB.

2.3.6 *Vocabulaire et corpus MNR du* Myanmar Newspaper Reader

L'ouvrage de Luzoe (1996), *Myanmar Newspaper Reader* est la seule ressource didactique de notre corpus qui emploie exclusivement des textes authentiques en langue écrite, à savoir, cinquante-deux articles de presse et éditoriaux publiés entre mai et octobre 1995 par les journaux quotidiens *The New Light of Myanmar* မြန်မာ့အလင်း /mjànmà.ʔəlíɴ/ et *The Mirror* ဓကြီးမုံ /tɕémòʊɴ/, tous deux des organes du gouvernement birman, et à ce titre, compte tenu de leur contenu les chapitres sont généralement d'une lecture fastidieuse. Chaque chapitre se compose du texte, d'un glossaire exhaustif et quelques notes. La traduction complète de tous les textes est donnée en fin d'ouvrage. Nous en avons extrait les textes et les glossaires, utilisant la méthode d'océrisation employée pour *Myanmar Flower Grammar* et *Manuel de birman, Volume 1* détaillée ci-dessus. Les glossaires ont été combinés, les entrées avec des chiffres écartées et les doublons éliminés pour aboutir à une liste de vocabulaire de 2688 unités lexicales.

2.3.7 *Vocabulaire du* Myanmar Language Test MLT

Les listes de vocabulaire (Okano et al. 2016) au format PDF disponibles sur le site du Myanmar Language Test[31] pour aider les candidats à se préparer

31. http://www.mlt-myanmar.com

Code niveau	Compétences testées	% orale % écrit	Vocabulaire
MB	Alphabet birman, vocabulaire et grammaire simples pour la vie quotidienne	100% orale	200 à 300
M1	Vocabulaire et grammaire de base pour survivre et communiquer dans la vie quotidienne	100% orale	500 à 700
M2	Communication de la vie quotidienne, lecture de paragraphe simple	100% orale	1200 à 1400
M3	Communication avec locuteur natif, lecture de paragraphe normale	70% orale 30% écrit	2500 à 2700
M4	Communication commerciale, lecture de paragraphe difficile	50% orale 50% écrit	5000+

TAB. 2.6 : Définition des niveaux du Myanmar Language Test

pour le test, partiellement en libre accès et partiellement payantes, sont réparties en niveaux de difficulté selon les niveaux de compétence de langue testés par la *Myanmar Langauge Test Association*. Les niveaux et tailles du vocabulaire attendus des candidats au test annoncés par l'association sont détaillés dans le tableau 2.6. Les listes elles-mêmes contiennent 296 unités pour le niveau MB[32], 353 unités pour le niveau M1, 647 unités pour le niveau M2, 1366 unités pour le niveau M3 et 3340 unités pour le niveau M4. Le texte birman encodé au système *Zawgyi* a été séparé du reste du contenu textuel (transcription de la prononciation, traductions en anglais et partiellement en japonais) puis transcodé en Unicode standard avant une vérification manuelle.

2.3.8 *Vocabulaire et corpus SS du site* Seasite Burmese Lessons

Le site internet Seasite[33], mentionné dans le chapitre précédent, contient une introduction au système d'écriture du birman (*Script*), et deux niveaux de leçons : *Beginning Burmese* (23 leçons) et *Intermediate Burmese* (16 leçons). Chaque leçon est divisée en plusieurs parties : *Objective*, *Text*, *Translation*, *Glossary*, *New Words*, *Syntax*, *Drills*, *Keys*, *Quizzes*, *Audio*. Nous n'avons inclus dans notre corpus que le contenu textuel en birman des pages *Text* (le texte de la leçon), *Glossary* et *New Words* qui présentent le vocabulaire, *Syntax* (l'explication de la grammaire), *Drills* (les exercices) et *Keys* (les réponses aux questions). Chaque niveau constitue un

32. MB = *Myanmar Beginner*
33. http://www.seasite.niu.edu/burmese/

Sous-corpus	Contenu	Textes	Tokens
Beginning	Text, Glossary & New Words, Syntax, Drills	23	9421
Intermediate	Text, Glossary & New Words, Syntax, Drills	16	18 814
Literature	Burmese Short Stories + Burmese Legends	14	40 268
	Burmese Poems	2	189
Seasite Corpus		55	68 692

TAB. 2.7 : Contenu du corpus Seasite Burmese Lessons

sous-corpus.

En plus de ces leçons, le site contient des pages de textes littéraires bilingues birman-anglais, *Burmese Literature*, divisés en *Burmese Short Stories* (six textes), *Burmese Legends* (huit textes) et *Burmese Poems* (dix textes, dont seulement deux ne sont pas en format image). La plupart des textes littéraires sont aussi écrits au style parlé. Nous avons regroupé les quatorze textes de *Burmese Short Stories* et *Burmese Legends* en un seul sous-corpus. Les textes littéraires comprennent des nouvelles de Theippan Maung Wa သိပ္ပံမောင်ဝ (1899–1942), Nay Win Myint နေဝင်းမြင့် (1952–), Zawgyi ဇော်ဂျီ (1907–1990), Win Pe ဝင်းဖေ (1935–) et Yin Yin Nu (Mandalay) ယဉ်ယဉ်နု(မန္တလေး) (une auteure contemporaine). Les légendes sont tirées de Maung Maung Pye (1952), et Khin Myo Chit (1984) et les deux poèmes traditionnels ont été composés par la reine Ma Mya Le အနောက်နန်း မမြလေး (1809–1809) et le poète Shin Maha Thilawuntha ရှင်မဟာသီလဝံသ (1453–1518). Nous avons numéroté ces textes dans le corpus en ordre de présentation sur le site[34], voir le tableau 2.8.

Nous avons téléchargé les pages du site et nettoyé les textes de tout balisage et contenu anglais. Le contenu du site est originellement encodé en Unicode.

2.3.9 *Le site didactique* TUFS Language Modules

Il s'agit des textes et transcriptions de cours en ligne créés par le *Center of Usage-Based Linguistic Informatics (UBLI)* à Tokyo University of Foreign Studies (TUFS) (Kawaguchi 2007). Nous disposons de deux corpus pédagogiques TUFS,

34. http://www.seasite.niu.edu/Burmese/literature/Literature.htm

N° de texte	Titre en anglais	Titre en birman	Genre
LIT01	Ma Le's Bracelets	မလဲ့လက်ကောက်	nouvelle
LIT02	Tha Dun	သာဒွန်း	nouvelle
LIT03	Ma Sa-Oo From Pantha and Me	ပန်းသာမစာအုန့်ဦ ကိုဒေဝ	nouvelle
LIT04	Wednesday Nan	ဗုဒ္ဓဟူးနံ	nouvelle
LIT05	Innwa	အင်းဝ	nouvelle
LIT06	His Spouse	သူမယား	nouvelle
LIT07	Kyaikhtiyo	ကျိုက်ထီးရိုး	légende
LIT08	Maha Myat Muni	မဟာမြတ်မုနိ	légende
LIT09	Maung Pauk Kyine	မောင်ပေါက်ကျိုင်း	légende
LIT10	The Story of Maung Tint De	မောင်တင့်တယ်	légende
LIT11	Nga Tat Pya, the Daring Robber	သူခိုး ငတက်ပြား	légende
LIT12	Shin-Mway-Loon and Min-Nanda	ရှင်မွေးလွန်း မင်းနန္ဒ	légende
LIT13	Shwedagon	ရွှေတိဂုံဘုရား	légende
LIT14	Taungpyone	တောင်ပြုန်း	légende
LITPOE15	Contrition	–	poésie
LITPOE16	Evil Pleasures	စက်ဝဠုချမ်းသာ	poésie

TAB. 2.8 : Textes littéraires Seasite, numérotés dans le corpus en ordre de présentation sur le site.

l'un tiré des transcriptions de cours audiovisuels que nous avons organisées en corpus *TUFS Dialogues*, et l'autre un corpus de cours axés sur l'acquisition du vocabulaire, *TALPCo-TUFS Vocabulary* (Nomoto et al. 2018). Les textes dont sont tirés ces deux corpus sont encodés en Unicode.

2.3.9.1 Vocabulaire et corpus TUFSD des *TUFS Myanmar Dialogues*

Ces modules, destinés à être utilisés en classe comme support pédagogique, visent le développement de quarante compétences communicatives spécifiques dont la liste figure dans le tableau 2.9 de descripteurs ci-dessous. Il s'agit de dialogues entre deux ou trois personnes scénarisés mais non traduits, la langue est donc apparemment proche de la langue parlée spontanée. Sur le site, le contenu audiovisuel peut s'accompagner de la transcription en écriture birmane, en transcription phonémique et/ou de la traduction en japonais, au choix de l'apprenant.

Après téléchargement des textes des modules de conversation (*Dialogue modules*)

du site *TUFS Language Modules*[35] avec un aspirateur de site Perl, nous avons nettoyé les pages de leur balisage HTML, et stocké le contenu au format texte, avant de procéder à l'étape de segmentation en tokens décrite dans le chapitre suivant.

Nous avons remarqué que l'HTML de ces pages contient un balisage fin du texte au niveau du vocabulaire, bien que ces informations ne semblent pas encore servir à nourrir l'interface. Dans la figure 2.1 le token[36] est balisé avec le sens 成功 /seikō/ (*succès*) et l'étiquette morphosyntaxique 名詞 /meishi/ (*nom/substantif*) :

FIG. 2.1 : Balisage HTML du vocabulaire de TUFS Dialogues

Nous avons décidé d'extraire cette information afin d'examiner le modèle de vocabulaire pédagogique tel qu'il a été conçu par les créateurs de ce site. Ceci nous a permis d'extraire une liste de vocabulaire de 1 186 unités. Les principales étiquettes utilisées sont données dans le tableau 2.10.

2.3.9.2 Vocabulaire et corpus *TALPCo - TUFS Myanmar Vocabulary*

Le corpus TALPCo (TUFS Asian Language Parallel Corpus)[37], préparé par Nomoto et al. (2018), est un corpus parallèle quadrilingue (japonais, anglais, malais et birman) qui utilise la langue japonaise comme langue pivot. Le contenu provient du module d'acquisition de vocabulaire (Vocabulary Module) sur le site des *TUFS Language Modules*[38] qui a été créé à partir de phrases d'exemples isolées portant sur les 799 éléments de vocabulaire japonais considérés comme formant le vocabulaire du niveau le plus bas du test d'aptitude en japonais

35. http://www.coelang.tufs.ac.jp/mt/my/dmod/
36. Les champs *token* et *type* sont identiques partout dans le corpus, nous ne gardons que le contenu du champ *token*.
37. https://github.com/matbahasa/TALPCo
38. http://www.coelang.tufs.ac.jp/mt/my-en/vmod/

TAB. 2.9 : Leçons du corpus TUFS Dialogue Module

Leçon	Compétence
01	Greeting someone
02	Thanking
03	Attracting someone's attention
04	Introducing yourself
05	Saying sorry
06	Offering something
07	Saying goodbye
08	Asking the price
09	Asking about someone's experience(s)
10	Saying you intend to do something
11	Asking about the degree of something (price, etc.)
12	Asking about time
13	Asking about figures
14	Asking about the way to do something
15	Asking about someone's ability to do something
16	Asking for the location of a place
17	Asking about the characteristics of something
18	Giving your opinion
19	Expressing likes
20	Saying what you like doing
21	Stating a procedure
22	Asking how someone is
23	Setting conditions
24	Comparing
25	Suggesting
26	Giving reasons
27	Requesting
28	Giving examples
29	Reaching a compromise
30	Asking for permission
31	Saying someone is obliged to do something
32	Saying someone must not do something
33	Telling someone how to do something
34	Telling someone not to do something
35	Saying someone need not do something
36	Inviting someone
37	Advising someone
38	Asking someone to do something
39	Saying what you hope will happen
40	Introducing someone

TAB. 2.10 : Principales étiquettes du balisage de TUFS Dialogues

POS	Traduction	Exemples
副詞 /fukushi/	adverbe	မြန်မြန်, အရမ်း
副詞相当句 /fukushi sōtō-ku/	expression adverbiale	အခုနောက်ပိုင်း
助動詞 /jodōshi/	verbe auxilliaire	ရ, နိုင်, ကြ
格助詞 /kaku joshi/	particule de cas	ကို, မှာ, က
助数詞 /josūshi/	classificateur	ယောက်, ခု
複合表現 /fukugō hyōgen/	expression composée	အဲဒီလောက်, ဘာမှ
複合動詞 /fukugō dōshi/	verbe composé	ယူသွား, ပေါ်လာ
接続詞 /setsuzokushi/	conjunction	ဒါနဲ့, ဒါပေမယ့်
接続助詞 /setsuzoku joshi/	particule connective	ပြီး, မှ
指示詞 /shijigo/	démonstratif	ဘယ်, ဒီ
指示代名詞 /shiji daimeishi/	pronom démonstratif	ဒါ, အဲဒါ
終助詞 /shūjoshi/	particule de fin de phrase	နော်, ပေါ့
慣用句 /kan'yōku/	locution	တတ်အားသလောက်
間投詞 /kantōshi/	interjection	ဟုတ်ကဲ့, ဩော်
名詞化接尾辞 /meishi-ka setsuoji/	particule de substantivation	ဘက်, စရာ, မှု
数詞 /sūshi/	chiffre	၉
人称代名詞 /ninshō daimeishi/	pronom personnel	ကျွန်တော်, ကျွန်မ
助詞 /joshi/	particule	ပါ
人称代名詞・斜格形 /ninshō daimeishi shakaku-gata/	pronom personnel forme oblique	ကျွန်တော့်
成句動詞 /seiku dōshi/	verbe à syntagme	ကျေးဇူးတင်
接頭辞 /settō ji/	préfixe	မ
固有名詞 /koyūmeishi/	nom propre	ရန်ကုန်, ဂျပန်
数量表現 /sūryō hyōgen/	expression de quantité	တစ်ခွက်, တစ်ခု
接尾辞 /setsuoji/	suffixe	လေး, တွေ
敬称 /keishō/	honorifique	မောင်, ဦး
動詞 /dōshi/	verbe	ဆွဲ, ပေး
動詞文標識助詞 /dōshi bun hyōshiki joshi/	particule marqueur de phrase verbale	တယ်, ပြီ, ဘူး

standard, le *Japanese-Language Proficiency Test* abrégé JLPT, phrases qui ont été traduites dans les quatre langues. Nous ne gardons que les phrases en birman, 1372 phrases au total. Les auteurs du corpus ont fourni une version segmentée utilisant le tiret pour indiquer la séparation entre les unités lexicales, ce qui nous a permis d'en extraire une liste de vocabulaire unique (de 939 unités). Ainsi, l'indication de frontières entre syntagmes par des espaces typographiques est préservée. Une version étiquetée avec un jeu d'étiquettes morphosyntaxes que les auteurs qualifient de « provisoire » est aussi fournie.

Nomoto et al. (2018) précisent que les phrases d'origine seraient celles de la vie quotidienne, mais sont au registre le plus soutenu de la langue parlée japonaise, aspect qu'ils tentent de préserver dans les traductions. Les phrases birmanes sont traduites en style parlé. L'utilisation de la langue japonaise comme langue source serait un atout pour ce type de corpus parallèle selon les auteurs, car les traductions sont souvent influencées par la structure et le lexique de la langue source. Ils donnent comme exemple le fait que le birman, tout comme le japonais, n'exprime pas obligatoirement tous les arguments d'un prédicat, alors qu'en anglais (ou en français), les omettre serait agrammatical :

(52) ဈေး-သက်သာ-ရင် ဝယ်-မယ်။
zé-θɛʔθà-jìɴ wè-mɛ̀
prix-pas.cher-si acheter-V.IR
'S'ils sont bon marché, je les achèterai.'

Les auteurs fournissent trois fichiers pour le birman : un fichier avec le texte brut, composé d'une liste de phrases chacune identifiée par un numéro de référence à quatre chiffres permettant de retrouver la phrase correspondante dans les autres fichiers, un fichier avec le texte segmenté et un fichier d'étiquetage morphosyntaxique. La segmentation entre unités lexicales est indiquée par des traits d'union ou des espaces typographiques déjà présentes dans le texte brut comme dans l'exemple 52. Les étiquettes morphosyntaxiques utilisées sont listées dans le tableau 2.11.

2.4 Statistiques comparatives

Nous verrons par la suite que la définition de tokens pour le texte birman peut varier d'une étude à l'autre, selon le modèle de langue choisi et les

Etiquette	Partie du discours
n	nom
suf	suffixe
v	verbe
adv	adverbe
dem	démonstratif
pr	nom propre
postp	post-position
clf	classificateur
num	nombre
pron	pronom

TAB. 2.11 : Étiquettes morpho-syntaxiques du corpus *TALPCo*

besoins de l'étude en question. Quand il s'agit de corpus en birman, il est donc prudent de ne pas se contenter de fournir le nombre de tokens pour un corpus, mais aussi de le décrire d'une manière figée et facilement reproductible, afin de permettre la comparaison quantitative entre corpus qui proviennent d'études différentes. Pour ce faire, nous suivons la méthode de découpe en syllabes, déjà en usage pour la translittération de mots d'origine birmane (Okell 1971) et utilisée aussi en linguistique de corpus birmane par Hnin Tun (2013, p. 86). Nous utilisons l'outil de segmentation en syllabes `sylbreak` qui se sert d'expressions régulières pour déterminer les frontières de syllabes. L'outil, de Ye Kyaw Thu, existant dans de nombreux langages de programmation (avec Swan Htet Aung et Chan Mrate Ko Ko pour les versions Java et Javascript), est disponible sur github[39]. Le décompte ne concerne que les syllabes en birman : nous écartons toute ponctuation (y compris la ponctuation birmane) et les caractères étrangers. Cette étape s'accomplit aisément par moyen d'expression régulière utilisant les catégories Unicode identifiées par `\p{}` – nous remplaçons tous les caractères qui ne correspondent pas aux caractères du bloc Myanmar avec un p majuscule `\P{Myanmar}`.

Le tableau récapitulatif 2.12 fourni des statistiques de base sur nos corpus. Il faut noter que nous n'avons inclus que les informations pertinentes à chaque corpus pour donner un aperçu de leur taille générale et leur structure. Certains des corpus ne contiennent que des phrases isolées, d'autres, comme certains

39. https://github.com/ye-kyaw-thu/sylbreak

Code corpus	Taille en syllabes	Nombre total de phrases	Nombre de textes	Taille moyenne des textes en syllabes (écart type)	Nombre moyen de phrases par texte (écart type)
ALT	1 124 939	20 106	—	—	—
myPOS	350 811	11 021	—	—	—
BBC	812 940	16 394	2019	403 (σ 396)	8 (σ 10)
Lotaya	1 016 624	44 939	295	3435 (σ 1980)	152 (σ 96)
TED	1 488 429	49 859	713	2088 (σ 1364)	70 (σ 51)
MFG	9302	—	60	155 (σ 112)	—
CB	18 275	—	16	1142 (σ 301)	—
AiB	17 885	1163	34	526 (σ 340)	34 (σ 4)
MNR	23 838	413	52	458 (σ 444)	8 (σ 10)
SS déb	10 624	—	23	462 (σ 190)	—
SS int	23 124	—	16	1445 (σ 430)	—
SS lit	49 535	2323	14	3538 (σ 2012)	166 (σ 93)
TUFSD	6577	552	40	164 (σ 66)	14 (σ 5)
TALPCo	15 728	1372	—	—	—

TAB. 2.12 : Statistiques comparatives

corpus didactiques, incluent souvent des listes de vocabulaire et contiennent comparativement peu de phrases complètes. Pour les corpus didactiques, nous considérons chaque leçon dans son intégralité. Pour les besoins de notre analyse donc, chaque leçon entière (avec ses textes, listes de vocabulaire, exercices, explications grammaticales) est considérée comme un « texte ». Puisque les corpus ALT, myPOS et TALPCo ne sont constitués que de phrases isolées, il n'y a pas lieu de les décrire selon le nombre de textes.

La figure 2.2 illustre la disparité apparente entre nos corpus authentiques, au moins quand la taille est mesurée en syllabes. Ce déséquilibre n'est pas gênant en soi, car il est d'usage de citer la fréquence non pas par fréquence brute, mais normalisée à une base, fréquence par million de tokens par exemple. Nous verrons par la suite si cette disparité sera aussi prononcée quand les textes seront segmentés en tokens. Pris ensemble, ces trois corpus authentiques ont une taille de plus de trois millions de syllabes, mais cela est loin d'être suffisant pour en faire un corpus de référence[40]. Il nous semble donc plus intéressant de

40. A titre de comparaison, la taille du British National Corpus (BNC, https://www.english-corpora.

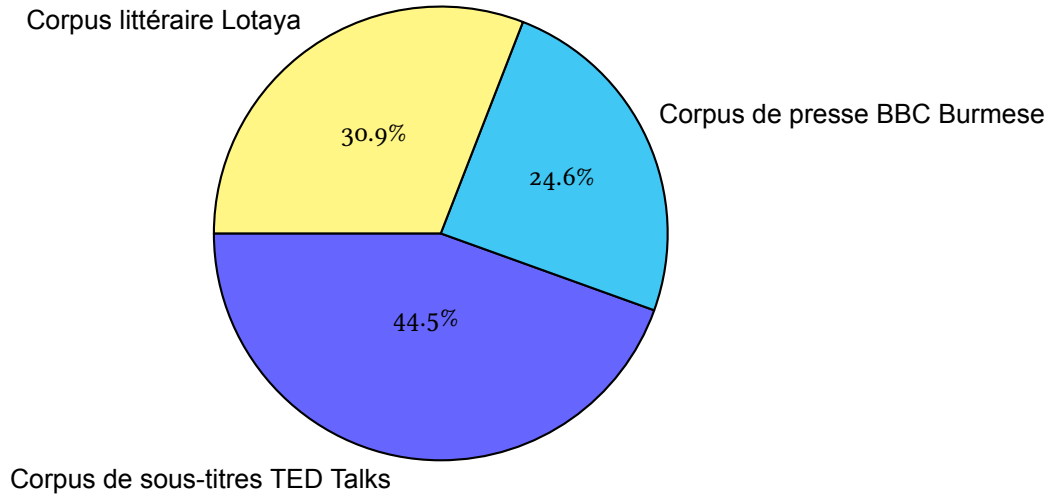


FIG. 2.2 : Comparaison quantitative de nos corpus authentiques par nombre de syllabes

considérer ces corpus séparés, chacun représentatif de leur genre textuel respectif, sous-titres de conférences, presse et littérature.

2.5 Prétraitements simples

Pour que les données soient traitées de manière uniforme par les mêmes procédés, il est nécessaire de normaliser les textes bruts. Cela évite d'inventer des traitements différents pour les textes d'origines diverses (Heitz 2006). Cette normalisation prend des formes différentes — on peut parler d'uniformisation ou de correction selon les cas. Ici, nous aborderons les prétraitements simples du texte birman qui influencent la précision du prétraitement de la segmentation, et doivent obligatoirement être fait en amont de celui-ci. D'autres traitements qui doivent être considérés en même temps que la segmentation, seront traités au chapitre suivant.

2.5.1 *La ponctuation et les chiffres*

Avant de procéder à une analyse statistique, il faut décider comment traiter les marques de ponctuation et les nombres représentés en chiffres, car ces choix ont des effets non négligeables sur la distribution de fréquences des occurrences.

org/bnc/) est supérieure à 100 millions de tokens (Burnard 1995).

Les marques de ponctuation constituent un ensemble fini dont chaque membre possède un taux d'occurrence élevé, alors que si les nombres, particulièrement ceux représentés en chiffres, apparaissent souvent dans les textes, ils sont membres d'un ensemble potentiellement infini dont chaque membre apparaît rarement. Puisque nous nous intéressons à la fréquence *lexicale* et à la fréquence relative d'une occurrence par rapport à un corpus entier, nous pouvons sans réticence éliminer la ponctuation de nos calculs quand il s'agit de ponctuation de structure, comme les caractères qui indiquent la fin de la phrase. Quant aux nombres, nous avons décidé de garder les nombres écrits en toutes lettres comme faisant partie du vocabulaire des textes, mais nous avons décidé de traiter les nombres en chiffres comme de la ponctuation et donc de les éliminer. Ce choix (fait aussi par Baroni (2008)) de ne pas compter les chiffres comme des tokens est justifiable dans la création de listes de fréquence lexicale, mais plus critiquable en ce qui concerne le calcul de difficulté relative de textes, car les nombres font partie du vocabulaire du texte. Les informations véhiculées par des nombres en chiffres sont totalement transparentes pour le lecteur, ne lui posant aucune difficulté de compréhension. Si un texte contient beaucoup de chiffres, il est donc plus 'facile', c'est-à-dire, sa lisibilité est forcément plus élevée. La manière de traiter ce problème lors du calcul de la lisibilité sera considérée dans la partie ultérieure sur la lisibilité de textes birmans.

Le texte d'origine en birman n'utilise pas les espaces typographiques pour séparer ses mots, mais ce caractère est bien utilisé pour organiser le texte. Selon Hopple (2007), l'emplacement de l'espace n'est pas en fait libre, mais, en combinaison avec les particules post-positionnelles, le segment bordé par les espaces forme une unité conceptuelle intermédiaire entre le mot et la phrase, une unité de sens cognitive. Afin de ne pas perdre cette information qui peut s'avérer utile pour certains traitements, nous remplaçons l'espace typographique (SPACE U+0020) dans le texte d'origine par un tiret bas _ (U+005F LOW LINE).

Voir tableau 2.13 pour la liste des principaux chiffres et marques de ponctuation birmans. À ces caractères s'ajoutent des marques de ponctuation occidentales, telles que %, $ ou le point, qui peuvent accompagner les chiffres.

Les marques de ponctuation à l'intérieur des mots requièrent un traitement particulier pour qu'elles ne soient pas considérées comme des séparateurs de

Forme en birman	Equivalent français	Point de code	Appellation Unicode
၀	0	U+1040	MYANMAR DIGIT ZERO
၁	1	U+1041	MYANMAR DIGIT ONE
၂	2	U+1042	MYANMAR DIGIT TWO
၃	3	U+1043	MYANMAR DIGIT THREE
၄	4	U+1044	MYANMAR DIGIT FOUR
၅	5	U+1045	MYANMAR DIGIT FIVE
၆	6	U+1046	MYANMAR DIGIT SIX
၇	7	U+1047	MYANMAR DIGIT SEVEN
၈	8	U+1048	MYANMAR DIGIT EIGHT
၉	9	U+1049	MYANMAR DIGIT NINE
၊	,	U+104A	MYANMAR SIGN LITTLE SECTION
။	.	U+104B	MYANMAR SIGN SECTION

TAB. 2.13 : Principaux chiffres et ponctuation birmans

tokens. C'est le cas en français d'un token comme *chou-fleur* par exemple, qu'il faut se garder de scinder en deux. En birman, les parenthèses ‹(› et ‹)› sont employées à l'occidentale comme ponctuation de structure pour isoler un groupe de mots à l'intérieur d'une phrase, mais aussi pour indiquer des caractères qui ne sont pas prononcés, au moins en birman. A titre d'exemple, *Wikileaks* est écrit ဝီကီလိ(ခ်) ⟨wïkïli(k^{hx})⟩ /wìkìlḭ/ avec en finale le caractère ခ ⟨k^{h}⟩ et l'asat ် ⟨^{x}⟩ (qui annule la voyelle inhérente) entre parenthèses, pour représenter la lettre latine ⟨k⟩ qui n'est pas prononcée en birman. Notre outil de segmentation, *Motor*, détaillé plus loin, nous permet de considérer les parenthèses comme des diacritiques quand elles apparaissent à l'intérieur des mots, les autres occurrences de parenthèses sont toujours considérées comme faisant partie du jeu de signes de ponctuation et traitées comme tels.

2.5.2 *Les caractères invisibles*

Le guide des spécifications du standard Unicode (Unicode Consortium 2019a) recommande l'usage du caractère invisible U+200B ZERO WIDTH SPACE (ZWSP), si l'utilisateur souhaite spécifier le placement de sauts de ligne ou de coupures entre deux caractères, normalement entre les mots.

Nous avons remarqué aussi la présence d'espaces invisibles insécables, princi-

palement le caractère U+00A0 NO-BREAK SPACE, qui a l'effet inverse d'interdire une coupure entre deux caractères. Ceci est employé pour ne pas segmenter un mot en deux. Dans nos corpus nous avons rencontré aussi d'autres types de caractères d'espaces insécables : U+200C ZERO WIDTH NON-JOINER (ZWNJ), U+200D ZERO WIDTH JOINER (ZWJ), U+2060 WORD JOINER (WJ), U+FEFF ZERO WIDTH NON-BREAKING SPACE, et même U+202F NARROW NO-BREAK SPACE, normalement utilisé en mongol.

Tous ces caractères invisibles sont potentiellement source d'erreur dans le traitement automatique et sont éliminés systématiquement lors du prétraitement après le transcodage en Unicode.

2.5.3 *L'uniformisation de l'encodage en Unicode*

2.5.3.1 Pourquoi Unicode ?

Choisir un seul système pour encoder tous nos textes de tous nos corpus est nécessaire pour pouvoir comparer les données, mais pourquoi choisir Unicode ? Dans le chapitre précédent, nous avons présenté certains aspects d'autres systèmes d'encodage qui présentent des inconvénients techniques pour le stockage, le tri et la récupération de données que nous résumons ici.

En somme, il y a trois raisons principales de favoriser l'Unicode sur les autres systèmes d'encodage pour encoder un corpus birman. Premièrement, les systèmes d'encodage pré-Unicode qui utilisent les points de code de l'alphabet latin, comme l'encodage Avalaser, ne permettent pas de stocker des données à écritures multiples (un texte avec des lettres latines à l'intérieur d'un texte birman, par exemple). Deuxièment, les systèmes qui utilisent plusieurs glyphes pour les différentes formes contextuelles d'un seul caractère[41] sont moins précis dans le tri de tokens. Finalement, il est primordial de respecter un seul ordre de stockage de caractères, pour s'assurer de la fiabilité de l'appariement automatique de tokens lors de calculs statistiques.

2.5.3.2 Détecter l'encodage

Une méthode très simple pour créer un corpus exclusivement encodé en Unicode serait de ne choisir que du contenu déjà encodé en Unicode. Ceci

41. Rappelons l'exemple, valable et pour l'encodage Zawgyi et pour l'encodage AvaLaser, de l'usage de points de code distincts pour les glyphes multiples tels que ◌ and ◌ du caractère ◌.

réduirait considérablement le contenu disponible, car l'encodage Zawgyi est toujours largement majoritaire sur l'internet birman, estimé à 85% des pages par Pann Yu Mon et al. (2011), mais probablement encore plus répandu depuis, car les polices *Zawgyi-One* sont préinstallées sur la plupart des smartphones vendus en Birmanie (Liao 2017).

Ecarter le contenu non Unicode implique aussi l'identification de l'encodage du contenu. Ceci n'est pas toujours possible à l'œil nu, car les créateurs de pages internet en birman souvent incorporent des polices pour s'assurer d'un affichage correct des caractères. En ce qui concerne la détection automatique d'encodage, la mention `UTF-8` dans le code HTML n'est pas une garantie d'encodage Unicode, car le contenu de pages à encodages basés sur les points de code Unicode comme l'encodage Zawgyi est aussi stocké en `UTF-8`. Pour ces raisons, il est souvent nécessaire d'avoir recours à un outil spécifique pour détecter l'encodage Zawgyi automatiquement.

L'incorporation de contenu qui n'est pas encodé en Unicode réclame aussi une étape de détection d'encodage pour pouvoir les transcoder en Unicode. S'il existe des outils pour la détection automatique de Zawgyi[42], nous n'en avons trouvé aucun pour d'autres systèmes. Une méthode d'identification d'encodage simple, mais fastidieuse est de copier-coller un échantillon de texte dans le bloc-notes de l'ordinateur et tester des polices différentes jusqu'à l'apparition de l'affichage correct du texte. Or, cette méthode présuppose que les polices sont déjà installées sur l'ordinateur, ce qui n'est pas toujours le cas. Pour les fichiers PDF contenant du texte (et non pas des images de texte), l'outil en ligne de commande `pdffonts` affiche la liste de polices incorporées dans le fichier. Une fois l'encodage détecté, il existe de nombreux outils en ligne de conversion d'un encodage à l'autre (le transcodage), le plus complet semble être l'application *Burmese Font Converter*[43].

2.5.3.3 Le transcodage du percent-encoding en Unicode

L'encodage en pourcent (*percent-encoding*), qui n'est pas spécifique au birman, est une représentation en caractères ASCII de codes de caractères. Elle est en

42. Google a mis à disposition sur GitHub un détecteur-convertisseur pour l'encodage Zawgyi, appelé *Myanmar Tools*, basé sur un modèle d'apprentissage automatique https://github.com/google/myanmar-tools.

43. http://burglish.my-mm.org/latest/trunk/web/fontconv.htm

Corpus	Style	Encodage d'origine
ALT Wikinews	écrit	Unicode
myPOS Wikipédia	écrit	Unicode
BBC Burmese	parlé	Unicode
Lotaya littérature	écrit	Zawgyi
TED Talk sous-titres	parlé	Unicode
BBE Burmese by Ear	parlé	Avalaser
MdB1 Manuel de birman 1	parlé	—
MFG Myanmar Flower Grammar	parlé	—
CB Colloquial Burmese	parlé	Avalaser
AiB Advancing in Burmese	parlé	Unicode
MNR Myanmar Newspaper Reader	écrit	—
MLT Myanmar Langauge Test	parlé	Zawgyi
SS Seasite Lessons	parlé	Unicode
TUFSD TUFS Dialogues	parlé	Unicode
TALPCo/TUFS Vocabulary	parlé	Unicode

TAB. 2.14 : Styles et encodages de corpus et ressources birmans. Les corpus sans mention d'encodage ont été extraits à partir d'images

général utilisée pour permettre dans les noms de ressources (URI) des caractères normalement interdits (ponctuations, caractères accentués ou non latins). Les noms de fichier des fichiers TED Talks étaient dans ce format. Les noms de fichier y ont souvent recours pour pouvoir insérer des espaces typographiques alors représentés par la chaîne %20, ou des caractères non latins, l'URI est alors illisible et souvent extrêmement longue (voir l'exemple de la figure 2.3).

```
%E5%AE%9A%E5%B9%B4%E5%BE%8C%E3%81%AF%E7%A4%BE%E4%BC%9A%E3%81%B8%E6
%81%A9%E8%BF%94%E3%81%97%E3%82%8F%E3%81%9F%E3%81%97%E3%81%AE%E8%87
%AA%E7%94%B1%E6%99%82%E9%96%93%E3%81%AE%E4%BD%BF%E3%81%84%E6%96%B9
%20%20%E9%B3%A5%E5%B1%85%20%E8%82%87%20%20TEDxAnjo.my
```

FIG. 2.3 : Nom de fichier en encodage à pourcent avant transformation

Nous avons utilisé la fonction `uri_unescape` dans le module Perl `URI ::Escape` pour décoder les chaînes de caractères des noms de fichiers en encodage à pourcent (*percent-encoding*) en Unicode standard dans le sous-programme (Burke 2002) suivant :

```
sub smartdecode {
  use URI::Escape qw( uri_unescape );
  use utf8;
  my $x = my $y = uri_unescape($_[0]);
  return $x if utf8::decode($x);
  return $y;
}
```

Le sous-programme est appelé :

```
$decoded = smartdecode( $chaîneàdécoder )
```

Après décodage, et remplacement des espaces par les tirets bas _ permis dans les URIs, le nom de fichier dans la figure 2.3 s'avère être
定年後は社会へ恩返しわたしの自由時間の使い方__鳥居_肇__TEDxAnjo.my

2.5.3.4 Le transcodage de AvaLaser en Unicode

Comme nous avons vu au chapitre précédent, l'encodage AvaLaser se différencie en tant qu'encodage de glyphes et non de caractères. Ainsi, le caractère g affiché avec la police Arial, est transformé en ဂ avec la police AvaLaserT1A. Pour convertir de l'encodage AvaLaser en Unicode, nous devons donc identifier les parties du texte qui doivent être affichées en AvaLaser et ne convertir que ces parties. Ceci n'est pas impossible, surtout avec un fichier au format Word ou OpenOffice, mais s'avère moins propre avec les fichiers au format PDF. Nous effectuons le transcodage du système AvaLaser en Unicode à l'aide de l'outil `burmeseunicode` (Lwin Moe 2016)[44].

Tout texte en format Word (`.docx`) peut se transformer en XML via le format *Flat OpenDocument Text* (extension `.fodt`). Ceci est possible soit en ligne de commande soit avec l'interface graphique de LibreOffice ou OpenOffice. A partir du format XML, il est possible de choisir le texte écrit dans une certaine police et ainsi choisir quelles parties doivent être transcodées. Nos scriptes Perl choisissent les parties balisées en AvaLaser à convertir avec `burmeseunicode`.

Le transcodage de texte d'un fichier au format PDF comme celui de *Burmese by Ear* est potentiellement fastidieux, car nous n'avons pas trouvé de moyen de

44. Un outil en ligne pour le transcodage du système AvaLaser vers Unicode est disponible ici https://er-tim.fr/~stuck/JW/conversion/index.html

préserver les informations concernant les polices spécifiées lors de la transformation du PDF en texte, et nous ne disposions pas de suffisamment de textes pour entraîner un modèle de détection de la langue. Pour illustrer ce problème, voici l'affichage dans le fichier au format PDF d'une ligne de texte du manuel, comportant une transcription de la prononciation, suivie du texte en écriture birmane et de sa traduction en anglais, chacun dans une police différente :

(53) texte : ăye အရည် juice
police : Helvetica AvalaserT1A Palatino

Le texte brut extrait du PDF avant transcodage ressemble à ceci :

(54) texte : a˜ye Arv\ juice
police : système système système

On remarque que l'encodage du texte en écriture birmane est basé sur l'écriture latine. Fort heureusement, le texte en birman qui nous intéressait était entouré de tabulations (U+0009 CHARACTER TABULATION), et donc facile à extraire avec une expression régulière.

2.5.3.5 Le transcodage de Zawgyi en Unicode

Comme pour les textes en encodage Avalaser, le transcodage du système Zawgyi en Unicode s'est fait à l'aide de l'outil `burmeseunicode` (Lwin Moe 2016) intégré dans nos scripts Perl d'extraction de texte. L'affichage de texte encodé en Zawgyi masque souvent les fautes de frappe, telles que les instances multiples du signe *asat* ် (U+103A) qui ne se manifestent qu'après transcodage en Unicode. Notre script de nettoyage de ces fautes récurrentes intervient systématiquement juste après transcodage.

2.5.4 *Normalisation orthographique et typographique*

La normalisation de l'orthographe et la correction de fautes typographiques se font par rapport à une norme orthographique décidée par l'usage ou une autorité quelconque, les deux sources de normes orthographiques étant répertoriées dans des ouvrages de référence tels que les dictionnaires. Pour le traitement automatique, nous devons aussi prendre en compte les normes d'encodage établies par l'Unicode. La normalisation de l'orthographe permet à l'ordinateur d'identifier deux mots orthographiés différemment comme le même mot, que ce soit de

véritables variantes orthographiques ou de simples fautes corrigées. Nous avons voulu corriger les fautes, mais nous avons toutefois décidé de ne pas corriger avec zèle la plupart des variantes orthographiques, principalement pour des raisons pratiques ; la grande majorité de variantes concerne les noms propres pour lesquels souvent il n'y a pas d'orthographe standard.

2.5.4.1 L'orthographe du birman

Selon Nishi (1998), la première description écrite qui atteste d'une préoccupation pour la perfection du code écrit de la langue birmane par les pouvoirs centraux date de la dynastie Konbaung (1752-1885). Un édit royal de 1783 admoneste les sujets du roi sur leur orthographe, qu'ils consultent des livres d'orthographe ou suivent l'usage orthographique des textes de bonzes savants et de hauts fonctionnaires. Il semblerait que cet intérêt royal pour la normalisation de l'écriture de la langue remonte même au royaume de Pagan (849–1297), et pour le démontrer (Nishi 1998) reconstruit certains aspects de l'évolution la standardisation de l'orthographe notamment au moyen d'inscriptions lapidaires. Bradley (2011) note plusieurs réformes qu'il qualifie de récentes, la suppression de ိုင် ⟨iuw^{x}⟩, ုင် ⟨w^{x}⟩ et ဉ် ⟨u^{x}⟩ après 1878, la réintroduction de ဉ် ⟨u^{x}⟩ en 1970 (Bradley 2009), et la normalisation de ဖ ⟨p^{h}⟩ ဘ ⟨b̥⟩ en 1986 pour mieux représenter la prononciation moderne. La MLC a publié un guide de l'orthographe en 1978 (Allott 1985 ; Myanmar Language Commission 2003), élaboré sur la base de quatre principes : la préservation de l'orthographe traditionnelle, l'attestation littéraire, la signification en contexte et l'usage courant (Ne Win San 2012). Cette norme de la MLC est actuellement exemplifiée par le dictionnaire monolingue မြန်မာအဘိဓာန် /mjàɴmà?ə̰bḭdaɴ/ de la Myanmar Language Commission (2008), publié pour la première fois en 1991 et son équivalent bilingue birman-anglais မြန်မာ–အင်္ဂလိပ်အဘိဓာန် /mjàɴmà-ʔìɴgə̰leiʔʔə̰bḭdaɴ/ (Myanmar Language Commission 1993). Les ouvrages de référence existent donc, mais ils ne sont pas nécessairement accessibles ou systématiquement consultés. On constate que, par rapport à un pays développé comme la France, où le secteur du dictionnaire peut vendre jusqu'à trois millions dictionnaires de français par an (Dutilleul 2018), le tirage d'un dictionnaire birman est souvent limité à quelques milliers exemplaires officiels (5000 pour le dictionnaire monolingue မြန်မာအဘိဓာန် /mjàɴmà?ə̰bḭdaɴ/ publié par la Myanmar Language Commission (2008) par exemple). Ces ouvrages

sont diffusés plus largement par le piratage, soit au format papier ou numérique, mais on ne peut pas mesurer avec certitude l'influence de la norme officielle. Il est possible que l'usage ait plus d'influence que les ouvrages de référence, et cela explique pourquoi certaines variantes, bien que non répertoriées officiellement semble assez répandues. A titre d'exemple, တီဗွီ ⟨tïbŭvï⟩ (*télé*) est moins répandu que တီဗီ ⟨tïbï⟩, la seule variante que nous trouvons dans les dictionnaires en ligne, mais la première représente plus d'un million de résultats d'une recherche sur Google Myanmar[45] et elle est utilisée notamment dans la presse écrite, notamment par le service de diffusion en birman de la Voix de l'Amérique[46]. Il est donc parfois difficile à déterminer s'il s'agit d'une variante graphique acceptée d'usage courant ou d'une faute de frappe, car toutes les variantes ne sont pas recensées dans les dictionnaires existants.

2.5.4.2 Classification d'erreurs

En ce qui concerne les types d'erreurs, Aye Myat Mon et Thandar Thein (2013) en identifient quatre :

1. *les erreurs typographiques*
2. *les séquences de lettres erronées*
3. *les erreurs phonologiques (phonème-graphème)*
4. *les erreurs de contexte* et nous rajoutons un cinquième type :
5. *les fautes de substitution graphique*

Les erreurs typographiques sont de simples fautes de frappe physique. L'auteur connaît la forme correcte et substitue la mauvaise lettre, ou n'appuie pas suffisamment sur les touches du clavier omettant une lettre pour produire une unité qui n'a pas le sens voulu dans le contexte ou bien qui n'existe pas. Par exemple, dans la phrase သူ**ကျောင်**သို့သွားသည်။ ⟨θù**kyĕaŋˣ**θiu¹θwă²θɲˣ.⟩ (*Il va a l'école.*) ကျောင် ⟨kyĕaŋˣ⟩ devrait s'écrire ကျောင်း ⟨kyĕaŋˣ²⟩ (*école*) avec le caractère ◌း ⟨²⟩. Il nous semble évident que l'on devrait chercher à corriger ce type d'erreur dans le corpus.

Les erreurs d'ordre de caractères (séquences erronées) peuvent survenir suite à des fautes de frappe ou de changement d'encodage. Elles concernent principalement l'ordre de signes vocaliques et les marques de tons autour des signes

45. https://www.google.com.mm/
46. https://burmese.voanews.com/

consonantiques. Par exemple ေ ⟨e⟩ + ရ ⟨ṙ⟩ au lieu de ရ ⟨ṙ⟩ + ေ ⟨e⟩ → ေရ ⟨ṙe⟩ (*water*). Comme nous avons vu précédemment, certaines polices birmanes cachent ces erreurs ; elles subsistent dans la mémoire de l'ordinateur et perturbent le traitement informatique, mais elles sont invisibles à l'œil nu du lecteur. Là aussi, nous avons essayé de corriger ce type d'erreur.

Les erreurs phonologiques surviennent quand l'utilisateur ne connaît pas l'orthographe standard et écrit selon la prononciation. Aye Myat Mon et Thandar Thein (2013) donnent l'exemple ကွန်ယက် ⟨kŭnˣykˣ⟩ à la place de ကွန်ရက် ⟨kŭnˣrkˣ⟩ /kʊ̀ɴjɛʔ/ *réseau (informatique)*), où le caractère ယ ⟨y⟩ est remplacé par ရ ⟨r⟩, car les deux ont la même prononciation. Parfois, la forme fautive ne veut rien dire, mais parfois il s'agit d'une forme qui a un sens dans un autre contexte. Ici, ယက် ⟨ykˣ⟩ /jɛʔ/ veut dire *râteau* (la forme correcte ရက် ⟨ṙkˣ⟩ /jɛʔ/ a le sens de *tisser*). Aye Myat Mon et Thandar Thein (2013) appellent ce quatrième type d'erreur une erreur de contexte. Le mot ou syllabe est bien formé dans certains contextes, mais il n'est pas approprié dans d'autres. L'usage de ယက် ⟨ykˣ⟩ /jɛʔ/ *râteau* au lieu de ရက် ⟨ṙkˣ⟩ /jɛʔ/ *tisser* est un bon exemple, car le premier étant un substantif, l'usage à la place d'un verbe serait fautif. Quand il s'agit de syllabes mal formées, que ce soit à cause d'erreurs typographiques ou de mauvais séquencements de caractères, la détection d'erreurs est plus aisée, car à l'étape de la segmentation les erreurs provoquent des erreurs de segmentation[47], mais les cas où l'erreur implique une substitution d'une syllabe valide pour une autre, la détection est plus difficile, car nous n'avons pas accès à un outil de vérification d'orthographe contextuelle. Certaines substitutions sont si répandues que l'on pourrait légitimement se demander si ces cas ne devraient être préservés en tant que variantes acceptables par l'usage au lieu de les considérés comme de simples fautes de frappe à corriger dans nos corpus.

Le texte dactylographié birman comporte souvent notre cinquième type de faute de frappe, presque indétectable à l'œil nu, que nous appellerons faute de substitution graphique. C'est-à-dire la substitution d'un caractère visuellement similaire à la forme correcte selon les normes Unicode d'encodage. Selon nos corpus, la faute de substitution graphique la plus répandue serait la substitution du chiffre zéro ○ ⟨o⟩ U+1040 MYANMAR DIGIT ZERO pour la lettre ဝ ⟨w⟩ U+101D MYANMAR LETTER WA[48]. La substitution inverse (ဝ ⟨w⟩ pour ○ ⟨o⟩) existe

47. Nous verrons cela en plus de détails au chapitre suivant sur la segmentation.

48. La substitution du chiffre zéro pour cette lettre est compréhensible, car non seulement les

également, mais dans une moindre mesure. Alors que la plupart de polices birmanes utilise un même glyphe pour les deux, certaines, comme *Myanmar3*, distinguent la lettre ဝ ⟨w⟩ du chiffre pour zéro avec le glyphe d'un cercle discontinu, ဝ ⟨o⟩. Ces substitutions posent un problème pour la reconnaissance automatique de chaînes de caractères, et doivent être corrigées avant tout autre traitement. Si certaines combinaisons de caractères qui n'existent pas par ailleurs, comme ဉ + ◌ိ (⟨ụ⟩ U+1025 suivi de ⟨ï⟩ U+102E) à la place du seul caractère ဉိ (⟨ụ̈⟩ U+1026), peuvent se corriger par simple substitution, les substitutions concernant les chiffres doivent se faire avec plus de précautions. Fort heureusement, le cas de substitution de ဝ ⟨w⟩ (U+101D) par le chiffre zéro est de loin le plus répandu et le plus facile à détecter avec une expression régulière, car le chiffre zéro apparaît presque toujours après un autre chiffre. Les autres cas sont substitués par la lettre ဝ ⟨w⟩ (U+101D). Nous avons vérifié manuellement les instances de ၇ ⟨7⟩ (U+1047) pour vérifier qu'il ne s'agit pas de la consonne ရ ⟨ṛ⟩ U+101B. Voir tableau 2.15 pour une liste des fautes de frappe les plus fréquemment rencontrées.

Dans les cas de substitutions graphiques, plusieurs raisons nous incitent à favoriser la correction. D'abord, l'auteur qui pense que ces deux caractères sont interchangeables sans conséquence ne pense qu'à leur aspect visuel identique, et n'a pas l'intention d'écrire une variante. Pour lui il s'agit de *formes* identiques et non seulement d'*unités linguistiques* équivalentes. La deuxième raison concerne le traitement statistique ultérieur que nous voulons effectuer. Nous sommes intéressée par les statistiques de fréquence d'usage d'unités linguistiques en tenant compte de variantes orthographiques acceptées par la MLC, ou répertoriées.

Nous verrons dans le chapitre qui suit que la segmentation en elle-même aide à détecter les erreurs, car elles provoquent des erreurs de segmentation.

2.5.4.3 Normalisation du caractère vocalique (U+102B)

Un cas particulier concerne les formes du caractère vocalique ◌ါ ⟨ä⟩ (U+102B MYANMAR VOWEL SIGN TALL AA), une forme de ◌ာ ⟨a⟩ (U+102C MYANMAR VOWEL SIGN AA), qui n'est autre qu'une variante graphique en birman, utilisée après certains signes consonantiques[49] parfois simplement par préférence stylis-

deux ont le même aspect, mais aussi le chiffre est plus facile de trouver sur le clavier et ne nécessite pas l'usage de la touche majuscule.

49. La forme haute est employée après ခ, ဂ, စ, ဒ, ဓ, ဟ, ou ဝ.

Forme correcte	Point de code	Appellation Unicode	Forme fautive	Forme fautive décomposée	Point(s) de code	Appellation(s) Unicode
ဝ	U+101D	MYANMAR LETTER WA	ဝ		U+1040	MYANMAR DIGIT ZERO
ရ	U+101B	MYANMAR LETTER RA	?		U+1047	MYANMAR DIGIT SEVEN
ဈ	U+1008	MYANMAR LETTER JHA	ဈ	စ +	U+1005	MYANMAR LETTER CA
				ျ	U+103B	CONSONANT SIGN MEDIAL YA
ဦ	U+1026	MYANMAR LETTER UU	ဦ	ဥ +	U+1025	MYANMAR LETTER U
				ီ	U+102E	VOWEL SIGN II
ဩ	U+1029	MYANMAR LETTER O	ဩ	သ +	U+101E	MYANMAR LETTER SA
				ြ	U+103C	CONSONANT SIGN MEDIAL RA
ဪ	U+102A	MYANMAR LETTER AU	ဪ	သ +	U+101E	MYANMAR LETTER SA
				ြ +	U+103C	CONSONANT SIGN MEDIAL RA
				ေ +	U+1031	VOWEL SIGN E
				ာ +	U+102C	VOWEL SIGN AA
				်	U+103A	SIGN ASAT
။	U+104B	MYANMAR SIGN SECTION	။	। +	U+104A	MYANMAR SIGN LITTLE SECTION
				।	U+104A	LITTLE SECTION

TAB. 2.15 : Fautes de frappe fréquentes en birman

tique, mais plus probablement pour écarter la confusion entre lettres simples et combinaison de lettres. Par exemple, après la lettre ဂ ⟨g⟩ on préfère écrire ဂါ ⟨gä⟩ que ကာ ⟨ga⟩ (ဂ + ◌ာ, ⟨g⟩ + ⟨a⟩) combinaison qui ressemble à la lettre က ⟨k⟩. Le choix en Unicode d'utiliser des points de code différents pour ces caractères au lieu règles d'affichage contextuel au niveau de polices est motivé par la nécessité de s'adapter à une autre langue de Myanmar, le karen s'gaw, qui ne possède que la forme haute. D'un point de vue informatique, ces deux sont souvent considérés comme identiques, car les deux graphies sont généralement acceptées. Par exemple ၎ါး ⟨ŋhä²⟩ (*emprunter, prêter*) est une variante de ၎ား ⟨ŋha²⟩. Il est donc prudent de normaliser la graphie, de préférence remplaçant toutes les occurrences de ◌ာ ⟨a⟩ (U+102C MYANMAR VOWEL SIGN AA) par ◌ါ ⟨ä⟩ (U+102B MYANMAR VOWEL SIGN TALL AA) après ခ ⟨kʰ⟩, ဂ ⟨g⟩, င ⟨ŋ⟩, ဒ ⟨d⟩, ဓ ⟨đ⟩, ပ ⟨p⟩, et ဝ ⟨w⟩ pour faciliter la lecture du corpus. Il est toutefois possible de simplement remplacer toutes les formes hautes ◌ါ ⟨ä⟩ par ◌ာ ⟨a⟩ sans porter atteinte à l'intégrité du corpus.

2.5.4.4 Normalisation de l'ordre des caractères *aukmyit* et *asat*

Nous avons expliqué au chapitre précédent que l'ordre de stockage des caractères doit être figé en Unicode, à une exception près, l'ordre de la marque de ton *aukmyit*, ◌့ ⟨¹⟩, (U+1037 MYANMAR SIGN DOT BELOW) et le caractère virāma visible *asat*, ◌် ⟨ˣ⟩, (U+103A MYANMAR SIGN ASAT), dont les deux ordres doivent être considérés comme équivalents. Pour que nos statistiques soient fiables, nous devons normaliser nos données pour qu'il y ait un seul ordre, soit *aukmyit* + *asat* (l'ordre canonique), soit *asat* + *aukmyit* (l'ordre non canonique).

Pour respecter le standard Unicode, nous avons opté pour le premier, mais ce n'est pas le choix le plus pratique, car *aukmyit* peut indiquer un changement de ton lié à une valeur syntaxique, que l'on peut vouloir séparer lors de la segmentation en tokens. Dans l'exemple 55, il indique la possession.

(55) ဘုရင့် ဥယျာဉ် *jardin du roi (jardin royal)*
/bu̯jɪ̯ɴ/ /ʔu̯jjaɴ/
roi.POSS jardin

Le résultat de segmentation souhaitable serait comme l'exemple 56 :

(56) ဘုရင် ◌့ ဥယျာဉ်
/bṵjìɴ/ ⟨¹⟩ /ʔṵjjaɴ/
roi POSS jardin

Nous avons résolu ce problème lors de la segmentation.

2.6 Résumé

Ce chapitre a détaillé les corpus authentiques utilisés pour notre travail, et aussi les corpus didactiques que nous utilisons pour identifier le modèle de lexique que nous devons utiliser pour segmenter nos textes authentiques en unités minimales de traitement, procédé décrit dans le chapitre suivant. Plusieurs prétraitements doivent s'effectuer avant l'étape de la segmentation, certains obligatoires, d'autres facultatifs selon l'analyse des textes. Les prétraitements les plus importants sont résumés par la figure 2.4. Il est souhaitable de corriger l'orthographe avant la segmentation, mais comme nous découvrirons au chapitre suivant, la segmentation même est un moyen de détecter les fautes, et il est plus facile de segmenter plusieurs fois en vérifiant l'orthographe que de tout corriger en amont.

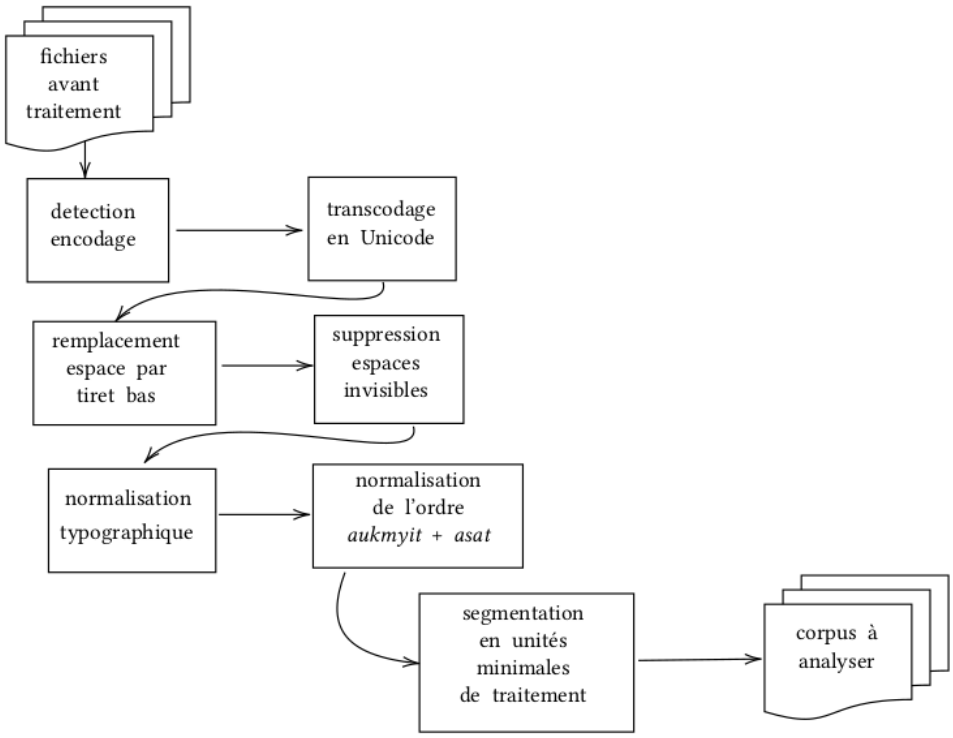


FIG. 2.4 : Schéma de prétraitements simples

Chapitre 3

La segmentation de textes birmans en tokens

> *The definition of the Burmese word … is not a matter to be taken for granted, neither can it be regarded as a luxury.*
>
> (Minn Latt 1959)

3.1 Introduction

Étape préliminaire essentielle dans le traitement automatique des langues naturelles, la segmentation en unités minimales de traitement, appelés *tokens*, se fait pour la plupart des systèmes d'écriture en découpant la chaîne de caractères qui constituent le texte informatisé avec des séparateurs tels que les espaces typographiques et certains caractères de ponctuation, c'est-à-dire selon les frontières des mots-formes ou mots graphiques. Les langues à scriptio continua, comme le japonais, le birman, le thaï ou le chinois, font appel à des outils externes de segmentation qui découpent le texte en tokens, habituellement avec la présupposition que le token équivaut au mot (sans pour autant décrire la définition du terme *mot*). Comme l'a fait remarqué Baroni (2008), l'identification de tokens et l'assimilation de tokens aux mots ne constituent pas des tâches insignifiantes, car celles-ci influencent les résultats statistiques relatifs. Les conséquences des choix de segmentation et d'identification de mots peuvent s'avérer plus ou moins sérieuses selon la tâche à accomplir. A titre d'exemple, dans le domaine de la censure automatique sur les réseaux sociaux, les erreurs peuvent paraître simplement ridicules, illustrées par les cas de censure de posts contenant des mots anodins, comme ကုလားထိုင် /kʊ̰lát^{h}àɪɴ/ ⟨kula^{2}t^{h}iuŋ^{x}⟩ *chaise* ou ကုလားပဲဟင်း /kʊ̰lápɛ́híɴ/ ⟨kula^{2}pɛhŋ^{x2}⟩ *soupe aux pois cassés* qui contiennent un segment potentiellement offensant ကုလား ⟨kula^{2}⟩ (Aung Kaung Myat 2017), mais

peuvent aussi s'avérer catastrophiques quand un système de censure automatique s'appuyant sur un outil statistique de traduction en amont (préalablement entraîné sur des corpus imparfaits) contribue à la diffusion de discours de haine (Stecklow 2018). Le niveau élémentaire optimal de segmentation n'est pas nécessairement le *mot* ; on peut choisir de segmenter un texte en caractères, en syllabes, ou en syntagmes selon les besoins d'analyse ultérieure du texte. Par exemple, Ye Kyaw Thu, Finch, Sagisaka et al. (2013) ont trouvé que la segmentation en syllabes était plus performante qu'une segmentation en mots pour la traduction automatique entre le birman et des langues de structure syntaxique similaire telles que le coréen et le japonais. Il nous semble donc important de prendre en considération le contexte du corpus dans la définition des tokens, autrement dit les spécificités de la langue du corpus et comment le corpus sera utilisé.

A la fin d'un texte didactique ou d'un ouvrage d'apprentissage de langue, on trouve généralement une liste diversement appelée *glossaire*, *vocabulaire*[1] ou *lexique*, qui contient les éléments discrets des textes. Ce sont les unités de base auxquelles le didacticien souhaite attirer l'attention de l'apprenant. Nous préférerons le terme *vocabulaire*, employé avec deux significations selon le contexte[2]. D'une part au niveau du traitement automatique pour désigner l'ensemble des *types*, c'est-à-dire des tokens différents d'un texte ou d'un corpus. D'autre part pour désigner le vocabulaire de l'apprenant, les unités discrètes de la langue présentées aux apprenants pour les aider à comprendre la langue. Dans ce chapitre, nous décrirons comment nous avons abordé la définition des tokens et de leur relation au vocabulaire de l'apprenant. Intuitivement, il nous semble que le token devrait correspondre à la *lexie*, définie ainsi :

> **Lexie**—Unité lexicale mémorisée au cours de l'apprentissage d'une langue et constituant un élément de la compétence d'un usager.
>
> *Dictionnaire de lexicologie française* (Tournier et Tournier 2017)

Nous examinerons dans quelle mesure cette notion est appropriée pour le birman et d'un point de vue pratique, comment elle peut être appliquée à son traitement automatique. Le chapitre commencera par une présentation de notre outil de segmentation dont le fonctionnement nécessite une liste de tokens de référence, en précisant quelques précautions d'emploi. La suite portera sur nos

1. Le terme équivalent *vocabulary* est le plus courant dans les ouvrages en anglais.
2. Le *vocabulaire* est un ensemble dont les membres s'appellent des *vocables*.

réflexions sur le contenu de cette liste qui détermine comment le texte est découpé : nous présenterons d'abord les aspects généraux du découpage du texte birman qui pourraient concerner tout type d'étude avant de nous pencher sur ceux qui s'appliquent spécifiquement à l'apprentissage du birman.

3.2 L'outil de segmentation automatique

3.2.1 *L'état de l'art de la segmentation automatique du birman*

Le premier aspect que nous devons prendre en compte dans la définition des tokens est le contexte de son traitement automatique, les limitations de l'outil choisi et les avantages et possibilités qu'il présente.

3.2.1.1 Segmentation en syllabes

D'un point de vue technique, la segmentation en syllabes est très commode, car assez simple à mettre en œuvre, elle a l'avantage de fournir un résultat régulier, prévisible pour l'usager d'un corpus. C'est pour cette raison que nous avons utilisé cette segmentation en syllabes pour caractériser la taille de nos corpus. Des méthodes de segmentation en syllabes à base d'automates à états finis ont été développées par Zin Maung Maung et Mikami (2008) avec une précision annoncée de 99,96% et aussi par Aye Myat Mon et Thandar Thein (2013). Plus récemment, Hlaing et Mikami (2014) ont utilisé une approche de segmentation en syllabes basée sur des transducteurs finis avec une précision de 99,93% y compris avec des mots à orthographe irrégulière comme les mots étrangers. Nous utilisons l'outil `sylbreak` de Ye Kyaw Thu décrit dans la section 2.4, basé sur l'article de Zin Maung Maung et Mikami (2008), pour segmenter en syllabes. La pertinence de la segmentation en syllabes pour notre étude du vocabulaire pour les apprenants de birman langue étrangère est discutée plus loin dans la section 3.3.1.

3.2.1.2 Segmentation en « mots »

Jusqu'à présent la plupart des études sur la segmentation en mots du birman ont concerné une approche combinant un dictionnaire et un algorithme privilégiant

la plus longue chaîne de caractères, appelée *forward maximum matching*[3], FMM (Tun Thura Thet et al. (2008) ; Win Pa Pa et Ni Lar Thein (2008)), parfois avec quelques règles supplémentaires. Hla Hla Htay et Narayana Murthy (2008) y ont adjoint un filtrage de mots vides. Ce dernier outil a été testé sur un corpus de 5000 phrases avec une précision de 99,11%, rappel de 98,81% et F-mesure de 98,95%. Pour confirmer l'efficacité de cette approche, une étude de Ye Kyaw Thu, Finch, Sumita et al. (2014) utilisant dictionnaires et un modèle bayésien non supervisé a conclu que l'approche FMM, la plus longue chaîne d'abord, était supérieure.

Ding, Ye Kyaw Thu et al. (2016) ont comparé un grand nombre d'approches de segmentation automatique sur un corpus de textes touristiques. Ils remarquent que les méthodes à base de dictionnaire sont moins performantes que les approches statistiques à base de corpus, mais ces dernières nécessitent plus de données d'entraînement, un problème pour les langues peu dotées. Ils ont mis en œuvre plusieurs méthodes à base de dictionnaire au moyen de correspondance maximale avant, arrière et bidirectionnelle. Ensuite, ils ont testé des méthodes statistiques : un modèle de langue à n-grammes de syllabes utilisant l'algorithme de Viterbi, une approche utilisant un SVM, (en anglais *support vector machine* une machine à vecteurs de support) et une autre avec un CRF, (en anglais *conditional random field* champs aléatoire conditionnel). C'est l'approche à CRF qui a donné les meilleurs résultats. Toutefois, ils notent que l'écart entre annotateurs et l'ambiguïté de segmentation dans le corpus posent problème. Ils préconisent la combinaison des étapes d'étiquetage morphosyntaxique et de la segmentation.

Ćavar et al. (2016) développent actuellement des outils de segmentation basés sur des transducteurs finis pour le birman qui utilisent des listes de mots et des expressions régulières Foma[4]. Ils expérimentent aussi avec des outils de segmentation à base de CRFs. La segmentation à base de CRFs entraînés sur corpus a permis la création d'une librairie python de segmentation pour plusieurs langues de Birmanie appelée *Pyidaungsu*[5] créé par Kaung Htet San.

3. C'est-à-dire *correspondance avec la plus longue chaîne d'abord* aussi appelé *correspondance maximale avant*.

4. https://fomafst.github.io/

5. Le 15/7/2021 à la version 0.1.4, *Pyidaungsu* est disponible dans les dépôts python via `pip install` et le code source est disponible sur GitHub à l'adresse suivante https://github.com/kaunghtetsan275/pyidaungsu, consultée le 5 septembre 2021.

3.2.2 *L'outil de segmentation* Motor

Dans la section précédente, nous avons vu que la plupart des outils de segmentation récents sont à base d'algorithmes d'apprentissage statistique, tels que les CRFs ou les SVMs. L'inconvénient de ces outils est la nécessité de disposer de corpus de grande taille préalablement segmentés à la main pour servir de modèle d'entraînement. Puisque nous ne disposons pas d'un tel corpus, nous utilisons un outil à base d'une liste de mots ou *dictionnaire*, appelé *Motor* (de Malézieux et al. 2014 ; Berment 2004) qui s'appuie sur un algorithme FMM combiné avec des heuristiques propres à la langue (Berment 2014). L'avantage de cette approche à base de dictionnaire, c'est que nous pouvons définir explicitement les tokens et nous assurer d'une régularité dans le résultat de segmentation, autrement dit, que la même chaîne de caractères sera toujours segmentée de la même façon. L'algorithme de *Motor* parcourt chaque phrase de gauche à droite et passe en revue toutes les possibilités de segmentation de la phrase avec référence au dictionnaire, avant de choisir la solution ayant le nombre minimal de tokens. Voici un court exemple pour illustrer son fonctionnement. Prenons la phrase suivante :

(57) ကျွန်တော် ကချင်ပြည်နယ်မှာ ပျော်တယ်။
⟨kǔẅnˣteaˣ kkʰǔŋˣpǰŋˣnyˣmȟa pǔeaˣtyˣ.⟩
1SG kachin-état-à aimer-NFUT
'J'aime l'état kachin.'

Leçon 16, niveau intermédiaire, SEAsite (Center for Southeast Asian Studies 2017)

Dans la figure 3.1, nous avons un exemple de dictionnaire complet et la segmentation de cette phrase résultant du traitement avec *Motor*. Nous observons que la chaîne de caractères ကချင် ⟨kkʰǔŋˣ⟩, qui représente la minorité ethnique kachin, n'est pas segmentée en deux, bien que les chaînes က ⟨k⟩ (marqueur du sujet) et ချင် ⟨kʰǔŋˣ⟩ (la minorité ethnique chin) figurent aussi dans le dictionnaire. Sans ကချင် ⟨kkʰǔŋˣ⟩ dans le dictionnaire (figure 3.2), le résultat aurait été différent, et tout à fait cohérent, *J̌'aime l'état chin*, car le marqueur က ⟨k⟩ serait bien placé après le sujet de la phrase ကျွန်တော် ⟨kǔẅnˣteaˣ⟩ /tɕʊɴɔ̀/ (*J̌e*).

Bien que Ding, Ye Kyaw Thu et al. (2016) aient démontré la supériorité des méthodes à base d'algorithmes d'apprentissage statistique quand les corpus d'entraînement et de test sont distincts (les F-mesures des méthodes SVM et CRF pouvant atteindre 0,978 et 0,979 respectivement, alors que la méthode de FMM ne

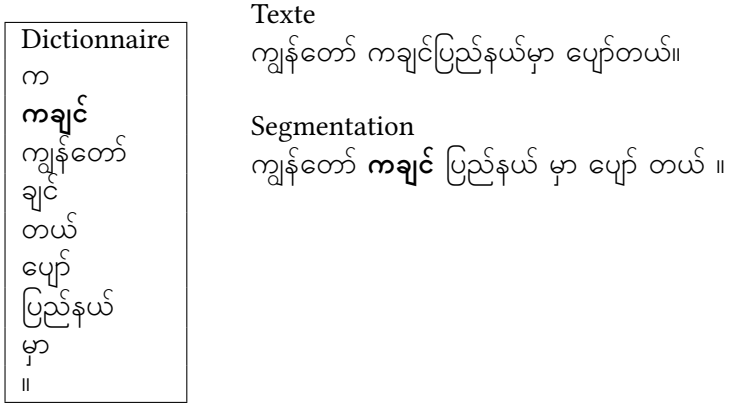


FIG. 3.1 : FMM avec dictionnaire complet. Le token ကချင် est présent dans le dictionnaire.

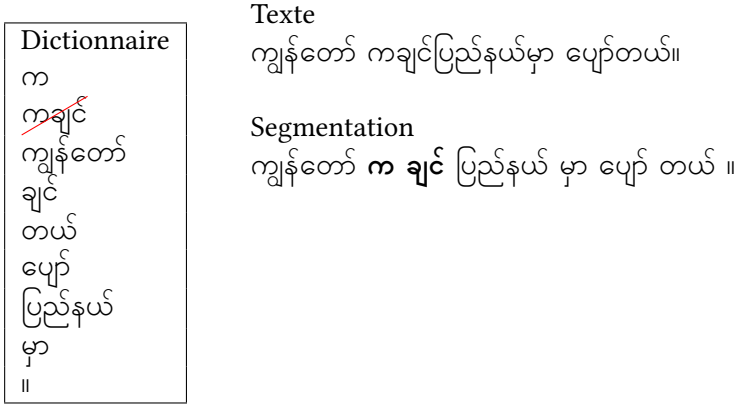


FIG. 3.2 : FMM avec dictionnaire incomplet. Le token ကချင် est absent du dictionnaire.

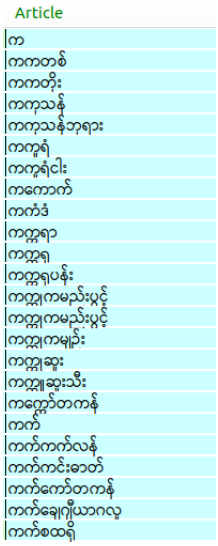


FIG. 3.3 : Début du dictionnaire birman de *Motor*

serait que de 0,924), la F-mesure obtenue avec la méthode FMM atteindrait 0,989 si le corpus de test est inclus dans le corpus d'entraînement. Ceci nous incite à croire que les performances de l'algorithme seront suffisantes pour nos besoins quand *Motor* est entraîné pour des corpus spécifiques. En effet, la qualité de la segmentation dépend en majeure partie de l'exhaustivité du dictionnaire fourni par rapport au texte à traiter. Par conséquent, ce système de segmentation nous semble plus adapté aux textes sur lesquels l'outil a été spécifiquement entraîné et moins performant pour des textes inconnus.

Un dictionnaire de *Motor* se présente sous la forme d'une base de données SQLite3 (figure 3.3). Pour une chaîne de caractères donnée, on peut interroger *Motor* par une interface en ligne[6] (figure 3.4) et ce pour plusieurs langues : le birman, le lao, le thaï, le khmer et le tibétain. Nous disposons d'une version conçue spécialement pour notre projet, permettant de charger le dictionnaire via une interface spécifique sur le web, puis d'interroger le serveur directement en `curl`.

Notre dictionnaire a d'abord été basé sur les vedettes du Wiktionnaire birman (ဝစ်ရှင်နရီ /wɪʔʃɪnnə̰jì/ my.wiktionary.org) archivé sur le site d'archivage de Wikimédia, dumps.wikimedia.org, et téléchargé fin 2013. Cette liste a été nettoyée

6. Disponible ici : lingwarium.org/motor/Segmentation/Segmentation.php

FIG. 3.4 : L'interface en ligne de *Motor*

soigneusement, puis augmentée au fur et à mesure que nous ramassions les textes de nos corpus. Les textes ont été segmentés en tokens, et examinés à l'aide de concordanciers pour relever les erreurs de segmentation survenues en raison de tokens manquants dans le dictionnaire.

3.2.3 *Précautions appliquées pour améliorer la segmentation*

Étant donné que Motor utilise un dictionnaire fini, nous avons fait certains choix pour réduire sa taille et améliorer la régularité de la segmentation résultante.

Nous avons déjà mentionné que l'inclusion des marqueurs de pluriel dans le dictionnaire permet de gagner de l'espace, car les formes plurielles n'ont pas besoin d'être incluses séparément. Cette approche est pratique, puisqu'elle effectue également une lemmatisation de base, à moins que l'utilisateur ne souhaite que la forme plurielle soit spécifiquement retenue, auquel cas elle ne serait pas incluse dans le dictionnaire. Nous avons adopté une approche similaire pour le traitement des chiffres.

La combinaison de deux caractères particuliers, appelés *aukmyit* et *asat*, nécessite des précautions particulières en matière de segmentation. Les deux combinaisons possibles semblent identiques, quel que soit l'ordre dans lequel elles sont tapées, mais parfois le caractère *aukmyit* a une fonction grammaticale et doit être séparé lors de la segmentation pour éviter qu'un même mot avec et sans *aukmyit* grammatical soit compté séparément.

3.2.3.1 La segmentation des chiffres

Pour traiter facilement et de manière uniforme les chiffres birmans de nos corpus, nous avons tout simplement ajouté tous les chiffres birmans de ၀ ⟨0⟩ à ၉ ⟨9⟩ à notre dictionnaire (voir les exemples de traitement de la figure 3.5). Ceci a pour effet de segmenter tous les chiffres, qui ne constituent plus un ensemble potentiellement infini, mais un ensemble composé de dix membres, facile à éliminer avec un prétraitement ou si nécessaire de les rassembler avec un post-traitement. Ainsi, le vocabulaire concernant les nombres en chiffres ne dépasse jamais plus de dix éléments et n'encombre pas le dictionnaire de *Motor*.

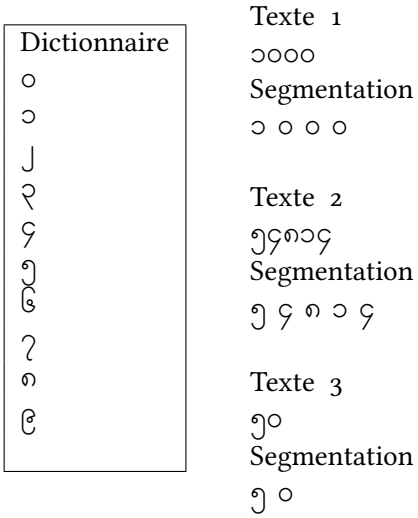


FIG. 3.5 : La segmentation des chiffres. Dans le premier exemple, ၁၀၀၀ ⟨1000⟩ est découpé en quatre tokens : un seul chiffre 1 ၁ ⟨1⟩ et trois zéros ၀ ⟨0⟩, ce qui nous donne un vocabulaire avec seulement deux éléments. Peu importe la longueur du chiffre, les dix éléments du dictionnaire suffisent pour isoler les composants en autant de tokens, que ce soit un chiffre long comme le deuxième exemple ၅၄၈၁၄ ⟨54814⟩ (cinq tokens) ou un chiffre court comme le suivant ၅၀ ⟨50⟩ (deux tokens).

3.2.3.2 Normalisation de l'ordre des caractères *aukmyit* et *asat*

La séquence des deux caractères *aukmyit* et *asat*, telle qu'elle est précisée dans la norme Unicode ne correspond pas malheureusement à l'usage, ce qui conduit à un mélange d'ordres standard et non standard. Il est donc important de veiller à normaliser l'ordre de ces deux caractères, afin de ne pas introduire d'erreurs dans le corpus lors de la segmentation et d'assurer la cohérence du traitement automatique. La normalisation s'avère délicate, en vertu du fait qu'il faut parfois séparer ces deux caractères lors de la segmentation. Nous avons trouvé qu'il est plus commode de traiter ce problème juste avant et après la segmentation.

Le caractère *aukmyit* အောက်မြစ် /ʔaʊʔmjɪʔ/[7], ့ ⟨¹⟩, (U+1037 MYANMAR SIGN DOT BELOW) est une marque de ton positionnée en dessous de la lettre finale d'une syllabe, comme l'illustrent les exemples suivants.

(58) ဒေ့ ⟨ne¹⟩ (*jour*)

(59) ပျော့ ⟨pyɔ̆a¹⟩ (*faible*, *mou*)

Le caractère virāma[8] visible *asat* အသတ် /ʔa̰θaʔ/, ် ⟨ˣ⟩, (U+103A MYANMAR SIGN ASAT), sert à annuler la voyelle inhérente d'un signe consonantique. Par exemple, le caractère န ⟨n⟩ est prononcé /na̰/, alors qu'en combinaison avec *asat* န် ⟨nˣ⟩ n'a plus de voyelle inhérente ; cette combinaison est utilisée pour indiquer un nasal /ɴ/ en position finale.

En Unicode, la norme de l'ordre de stockage des deux caractères, lorsqu'ils sont contigus, doit être *aukmyit*+*asat* ့+် ⟨¹ˣ⟩ (U+1037 + U+103A). Selon Hosken (2012), il s'agit d'une erreur dans la vérification de la norme, car effectivement, cette combinaison est habituellement écrite dans le sens inverse, aussi bien en écriture manuscrite qu'en dactylographie. Le chapitre du standard Unicode qui concerne les normes d'encodage pour les systèmes d'écriture de l'Asie du Sud-Est (Unicode Consortium 2019a) considère que les deux ordres sont équivalents et préconise que les implémentations supportent les deux ordres et les traitent comme fondamentalement identiques. En effet, aucune police n'indique que l'ordre inversé est erroné par l'affichage du rond en pointillé[9].

7. ou အောက်ကမြှစ် /ʔaʊʔka̰mjɪʔ/

8. Appelé aussi *halant* par Jacobs et al. (2018).

9. En Unicode le caractère du rond en pointillé ◌ (U+25CC DOTTED CIRCLE) est utilisé pour illustrer l'effet d'une *marque de combinaison* (combination mark) telle que les caractères diacritiques

En plus de compliquer le traitement automatique, on ne peut pas dire que toutes les occurrences de cette combinaison soient identiques, car le caractère *aukmyit* ◌့ ⟨¹⟩ indique tantôt une syllabe au ton haut immuable, comme dans မုန့် /mo̰ʊɴ/ ⟨mun¹ˣ⟩ (*collation*), tantôt un ton haut induit par une relation syntaxique, comme on peut voir dans les exemples suivants tirés de Okell (1969, p. 19) (les translittérations sont données en ordre canonique Unicode).

(60)
ကျနော်	+ ◌့	+ ကို	→	ကျနော့်ကို
⟨kyănea^{x}⟩	+ ⟨¹⟩	+ ⟨kiu⟩	→	⟨kyănea^{1x}kiu⟩
1SG	+ *asat*	+ OBJ	→	*vers moi*

(61)
အတင်	+ ◌့	+ မှာ	→	အတင့်မှာ
⟨ʔtɪŋ^{x}⟩	+ ⟨¹⟩	+ ⟨mha̍⟩	→	⟨ʔtɪŋ^{1x}mha̍⟩
Atin	+ *asat*	+ à	→	*à Atin*

(62)
နင်	+ ◌့	+ အမေ	→	နင့်အမေ
⟨nɪŋ^{x}⟩	+ ⟨¹⟩	+ ⟨ʔme⟩	→	⟨nɪŋ^{1x}ʔme⟩
2SG	+ *asat*	+ mère	→	*ta mère*

La translittération dans ces exemples donne à voir la difficulté posée par l'ordre canonique Unicode lors de la segmentation. Dans la phrase l'*aukmyit* ◌့ ⟨¹⟩ syntaxique se trouve avant *asat* ◌် ⟨ˣ⟩. Si nous voulons calculer le nombre d'occurrences d'un token comme ကျနော် ⟨kyănea^{x}⟩, y compris celles comprises dans ကျနော့်ကို ⟨kyănea^{1x}kiu⟩, par exemple, il faut prendre en compte non seulement les occurrences de ကျနော် ⟨kyănea^{x}⟩, mais aussi les occurrences de ကျနော့် ⟨kyănea^{1x}⟩. Il est donc souhaitable de séparer *aukmyit* ◌့ ⟨¹⟩ dans ces cas et par conséquent bien plus commode d'inverser l'ordre de *aukmyit+asat* ◌့+◌် ⟨¹ˣ⟩, en *asat+aukmyit* ◌်+◌့ ⟨ˣ¹⟩ en amont de la segmentation. De cette façon, lorsque *aukmyit* se trouve à la fin du token, les entrées du dictionnaire qui se terminent par *asat* sans *aukmyit* seront séparées du *aukmyit* lors de la segmentation. Cela signifie toutefois que toutes les occurrences de cette combinaison se trouveront inversées, y compris celles des tokens dont l'*aukmyit* n'a pas de valeur syntaxique, comme မုန့် /mo̰ʊɴ/ ⟨mun¹ˣ⟩ (*collation*) ou စောင့် /sa̰ʊɴ/ ⟨s̈eaŋ¹ˣ⟩ (*soleil*). Les versions de ces tokens avec ces caractères inversés devront donc figurer dans le dictionnaire de notre outil de segmentation, *Motor*.

ou les caractères vocaliques dépendants en birman, sans le caractère porteur (en birman il s'agit souvent d'un caractère consonantique). Le rond en pointillé apparaît quand une combinaison est mal formée.

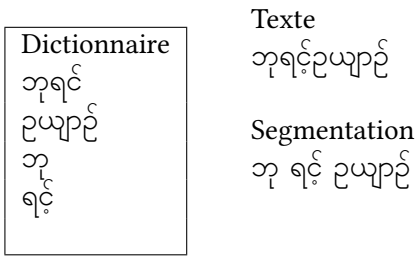


FIG. 3.6 : Processus de segmentation avec *Motor* sans l'inversion de l'ordre de *aukmyit* et *asat*

Pour illustrer le premier cas de segmentation où l'emploi de *aukmyit* ◌့ ⟨¹⟩ est syntaxique, nous reprenons l'exemple mentionné au chapitre précédent, l'exmeple 63. Ici, *aukmyit* indique la possession. Deux possibilités de stockage de caractères sont possibles, illustrées par les deux options de translittération, soit ⟨¹ˣ⟩, soit ⟨ˣ¹⟩. Dans l'exmeple 63, l'ordre de stockage indiqué est celui de l'Unicode ⟨¹ˣ⟩.

(63) ဘုရင် + ◌့ + ဥယျာဉ် → ဘုရင့်ဥယျာဉ်
⟨ɓurŋˣ⟩ + ⟨¹⟩ + ⟨ụyў̆aɲ̀ˣ⟩ → ⟨ɓurŋ¹ˣụyў̆aɲ̀ˣ⟩ ou ⟨ɓurŋˣ¹ụyў̆aɲ̀ˣ⟩
roi + POSS + jardin → *jardin du roi (jardin royal)*

Le résultat de segmentation souhaitable serait comme l'exemple 64.

(64) ဘုရင် ◌့ ဥယျာဉ်
⟨ɓurŋˣ⟩ ⟨¹⟩ ⟨ụyў̆aɲ̀ˣ⟩
roi POSS jardin

Dans le dictionnaire de *Motor* figurent ဘုရင် ⟨ɓurŋˣ⟩ *roi* et ဥယျာဉ် ⟨ụyў̆aɲ̀ˣ⟩ *jardin*. Sans inversion préalable de l'ordre d'*aukmyit* et *asat*, la segmentation est erronée (figure 3.6). Il est à noter que les tokens résultants figurent aussi dans le dictionnaire, ရင့် /jɪ̰ɴ/ ⟨rŋ¹ˣ⟩ (*mûr, foncé, impoli*) et ဘု /bṵ/ ⟨ɓu⟩ (*nodule, grosseur, intransigeant, brusque*).

Seule l'inversion de l'ordre de *aukmyit* et *asat* permet la reconnaissance de ဘုရင် ⟨ɓurŋˣ⟩ (*roi*) par *Motor*, car ဘုရင့် ⟨ɓurŋ¹ˣ⟩, ne figure pas dans le dictionnaire, puisque *aukmyit* ◌့ ⟨¹⟩ se trouve avant son dernier caractère *asat* ◌် ⟨ˣ⟩, signifiant qu'il ne s'agit pas d'un token valable. Les exemples qui suivent (65 à 67) montrent l'inversion du stockage des caractères avant la segmentation, qui insère une espace typographique (U+0020 SPACE) entre les tokens.

Etat initial :

(65) *aukmyit* + *asat*

ဘုရင့် U+1018 U+102F U+101B U+1004 U+1037 U+103A ့ + ်
roi.POSS

↓

Inversion :

(66) *asat* + *aukmyit*

*ဘုရင့် U+1018 U+102F U+101B U+1004 U+103A U+1037 ် + ့
roi.POSS

↓

Segmentation :

(67) ဘုရင် ့ U+1018 U+102F U+101B U+1004 U+103A U+0020 U+1037
roi POSS

Afin de rétablir l'ordre canonique Unicode, l'ordre des caractères *asat* + *aukmyit*, contigus dans tout le texte, est inversé de nouveau après la segmentation. Les cas tels que l'exemple 67 ne sont pas modifiés, car l'insertion de l'espace U+0020 fait que les caractères ne sont plus adjacents. Par contre, s'il s'agit d'un token qui n'a pas lieu d'être segmenté en deux, lorsque l'*aukmyit* n'a pas de fonction syntaxique, l'ordre canonique est rétabli en inversant de nouveau *asat* + *aukmyit* après la segmentation, de cette manière :

(68) Phrase à segmenter : ဒီတော့ နေပူထဲစောင့်တော့။
Ainsi, ils attendent au soleil.

Dans l'exemple 68, l'ordre est canonique, *aukmyit* + *asat* ့ + ်, l'unité စောင့် ⟨s̀eaŋ^{1x}⟩ /sa̰ʊɴ/ (*attendre*) est encodé comme l'exemple 69.

(69) *aukmyit* + *asat*

စောင့် U+1005 U+1031 U+102C U+1004 U+1037 U+103A
စ ေ ာ င ့ ် ်

Ensuite, pour prendre en compte les cas comme l'exemple 67, l'ordre de *asat* et *aukmyit* est inversé pour tout le corpus, comme ceci :

(70) *asat* + *aukmyit*

စောင့် U+1005 U+1031 U+102C U+1004 U+103A U+1037
စ ေ ာ င ် ့

Ensuite la phrase est segmentée en tokens avec notre outil *Motor* :

(71) Phrase en tokens : ဒီတော့ ၊ နေ့ ပူ ထဲ စောင့် တော့ ။

Puisque *aukmyit* ့ n'a pas de fonction syntaxique, mais fait partie intégrante de l'unité စောင့်, reconnue en tant que token par notre outil de segmentation, la chaîne de caractères စောင့် est un seul token. Nous inversons l'ordre de *aukmyit* et *asat* de nouveau :

(72) *aukmyit* + *asat*

စောင့် U+1005 U+1031 U+102C U+1004 U+1037 U+103A

စ ေ ာ င ့ ်

On peut résumer le processus ainsi :

1. Ajouter au dictionnaire de l'outil de segmentation *Motor* une deuxième version de toutes les entrées qui contiennent la combinaison *aukmyit+asat* dans l'ordre inversé *asat+aukmyit*.
2. Transformer l'ordre de toutes les occurrences dans le corpus de *aukmyit+asat* en l'ordre inversé *asat+aukmyit*.
3. Segmenter le corpus.
4. Transformer l'ordre de toutes les occurrences dans le corpus de *asat+aukmyit* contigu en ordre Unicode *aukmyit+asat*.

Un puriste préconiserait une règle d'écriture qui placerait *aukmyit* avant *asat* quand il ne s'agit pas d'un *aukmyit* grammatical, et un *aukmyit* que l'on pourrait qualifier de « libre » tout à la fin quand il s'agit d'un usage grammatical tel que les exemples 60 à 62. Mise à part l'inévitable application inégale d'une règle aussi ésotérique, certains cas posent problème, et d'un point de vue théorique et pour le traitement automatique. ကျနော့် ⟨kỳnea^{1x}⟩ (ou ⟨kỳnea^{x1}⟩) en est l'illustration. Dans l'exemple 60 il veut dire *vers moi*, qui devrait s'écrire avec *asat+aukmyit* (⟨kỳnea^{x1}⟩), mais dans d'autres cas il s'agit de l'équivalent d'un pronom personnel possessif *mon*. Devrait-on considérer qu'il s'agit d'une unité et l'écrire *aukmyit+asat* ⟨kỳnea^{1x}⟩ ou bien s'agit-il du pronom personnel *je* + marque de possession ? Le fait d'utiliser l'Unicode strict pour le pronom personnel, et l'ordre inversé pour le complément d'objet direct permettrait de distinguer entre les deux. Étant donné que les pronoms personnels possessifs constituent une classe finie, et le fait que les cas de pronom personnel possessif sont bien plus nombreux, nous avons retenu la première option, acceptant que

le cas de complément d'objet direct ne soit pas segmenté correctement en toute rigueur, tout en reconnaissant la validité théorique de l'autre approche.

Bien que nous ayons choisi de suivre strictement l'ordre Unicode, on voit à travers ces exemples qu'il est peu pratique et qu'il serait probablement préférable de suivre plutôt l'usage courant *asat+aukmyit*. Malheureusement, la norme Unicode ne peut pas être changée, la préservation de la compatibilité rétroactive étant un principe immuable de l'Unicode (Hosken 2012), il est donc souhaitable que toute documentation accompagnant un texte (corpus, base de données) en birman destiné à être partagé précise l'ordre choisi de la combinaison *aukmyit+asat*/*asat+aukmyit*.

3.2.4 *L'évaluation de la segmentation avec* Motor

On peut décrire l'évaluation d'un outil de segmentation de manière quantitative et qualitative, la première étant privilégiée dans la recherche en traitement automatique des langues. Une description quantitative de l'évaluation se base sur la similarité du résultat de segmentation d'un corpus de référence à un résultat de référence, le même corpus pré-segmenté (appelé un *gold standard* en TAL). Les mesures statistiques classiques utilisées sont la précision et le rappel, combiné dans un indice de performance comme la F-mesure (l'indice de Dice), chacun exprimé en pourcentage. La précision exprime la fraction de segments trouvés qui sont corrects et indique la validité des résultats, alors que le rappel représente la fraction de segments corrects qui ont été trouvés, c'est-à-dire, l'exhaustivité des résultats.

$$\text{précision} = \frac{|\text{réponses correctes} \cap \text{réponses}|}{|\text{réponses}|}$$

$$\text{rappel} = \frac{|\text{réponses correctes} \cap \text{réponses}|}{|\text{réponses correctes}|}$$

$$\text{F-mesure} = 2 \times \frac{\text{précision} \times \text{rappel}}{\text{précision} + \text{rappel}}$$

Nous avons utilisé un échantillon de mille phrases du corpus *myPOS* (Khin War War Htike et al. 2017) pour l'évaluation, la petite taille de l'échantillon

étant motivée par la volonté de faire aussi une évaluation qualitative, c'est-à-dire comprendre la cause des erreurs rencontrées. Le choix du corpus *myPOS* plutôt que le corpus *ALT Wikinews* (Riza et al. 2016 ; Ding, Utiyama et al. 2018 ; Ding, Aye et al. 2019) est motivé par les choix de segmentation de ce dernier, qui nous paraît sous-segmenté par rapport à *myPOS*. En ce qui concerne les fins de phrase par exemple, là ou *ALT* ne segmente pas du tout les éléments qui suivent le verbe (l'exemple 73), *myPOS* sépare les marqueurs et particules (voir l'exemple 74), ce qui correspond davantage à la façon dont la langue est enseignée.

(73) … ပြောခဲ့သည် ။ *ALT*
… ⟨pǰeak^{h}ɛ^{1}θɲ^{x}⟩ ⟨.⟩
… parler-DISPL-NFUT .
'a dit'

(74) … ပြော ခဲ့ သည် ။ *myPOS*
… ⟨pǰea⟩ ⟨k^{h}ɛ^{1}⟩ ⟨θɲ^{x}⟩ ⟨.⟩
… parler DISPL NFUT .
'a dit'

3.2.4.1 Test avec dictionnaire complet

La première évaluation concerne l'outil uniquement et les limites de l'algorithme utilisé par *Motor*, *forward maximum matching* (FMM), c'est-à-dire la correspondance avec la plus longue chaîne d'abord. Pour ce faire, nous avons décidé d'utiliser un dictionnaire complet, c'est-à-dire contenant tout le vocabulaire du corpus de référence contre lequel nous comparons la segmentation résultant du traitement de *Motor* (un sous-corpus de mille phrases du corpus *myPOS*). Cette première évaluation, où on « triche » avec un dictionnaire complet, permet aussi de mieux apprécier l'effet d'un dictionnaire de plus grande taille qui pourrait être incomplet pour ce corpus test, mais utilisable sur d'autres textes. Ce dictionnaire contient 4826 entrées. La taille du corpus de référence est de 20 270 tokens. Les espaces d'origine du corpus n'ayant pas été préservées par les créateurs de *myPOS*, nous n'avons pu qu'enlever les espaces (et les étiquettes morphosyntaxiques) de la version segmentée avant d'effectuer la segmentation avec *Motor*. *Motor* segmente le même corpus en 20 284 tokens. Les résultats (résumés dans le tableau 3.1) sont plus que satisfaisants avec un dictionnaire

	Dictionnaire du corpus	**Dictionnaire général**
Taille dictionnaire	4826 entrées	72 096
Taille référence	20 270 tokens	20 270 tokens
Vocabulaire réf	4826	4826
Taille test *Motor*	20 284 tokens	21 932 tokens
Précision	0,993	0,853
Rappel	0,994	0,962
F-mesure	0,993	0,904

TAB. 3.1 : Évaluation quantitative de segmentation par *Motor*

complet avec un F-mesure de 0,993, encore plus intéressante que lors d'une expérience similaire menée par Ding, Ye Kyaw Thu et al. (2016) ayant obtenu une F-mesure de 0,989 avec la méthode FMM et un dictionnaire complet. Il est intéressant de noter en passant l'importance de la normalisation de l'ordre de caractères (discutée dans la section ci-dessus). Sans ce prétraitement, le rappel n'est pas sensiblement affecté (0,992), mais la précision baisse à 0,927 et la F-mesure à 0,958.

Ces statistiques d'évaluation seules ne permettent pas d'évaluer les performances de l'outil, car un peu plus de la moitié des erreurs de segmentation (56 sur 102) sont en réalité causées par des incohérences dans la segmentation de la référence. Nous entendons par incohérence que les mêmes chaînes de caractères ne sont pas segmentées de la même manière partout dans le corpus de référence. Par exemple, dans le corpus de référence ကိုလိုနီခေတ် ⟨kiuliunïk^{h}et^{x}⟩ (*période coloniale*) est parfois segmenté en tant que tel, et parfois ကိုလိုနီ ခေတ် ⟨kiuliunï⟩ ⟨k^{h}et^{x}⟩ (*colonie période*). Puisque nous avons ajouté tout le vocabulaire de la référence dans le dictionnaire de *Motor*, la segmentation de *Motor* est toujours ကိုလိုနီခေတ် ⟨kiuliunïk^{h}et^{x}⟩. De la même manière, *Motor* donne toujours le résultat ပထမဆုံး ⟨pt^{h}mś^{h}uň^{2}⟩ (*le premier*) au lieu de ပထမ ဆုံး ⟨pt^{h}m⟩ ⟨ś^{h}uň^{2}⟩ (*premier fin* ou particule de superlatif), car les deux sont dans la référence. Il n'y a pas forcément de solution « correcte » quand il s'agit de mots composés, mais il y a des cas où l'on peut s'attendre à des segmentations différentes de la même chaîne de caractères dans des contextes différents. Prenons le cas de la chaîne de caractères ငှက်မျိုး ⟨ŋȟk^{x}mўiu^{2}⟩ qui est composée de ငှက် ⟨ŋȟk^{x}⟩ (*oiseau*) + မျိုး ⟨mўiu^{2}⟩ (*type, espèce, catégorie*). Dans la référence nous avons deux segmentations différentes dans deux contextes

différents. Dans le premier exemple 75, ငှက်မျိုး ⟨ŋȟkˣmўiuˀ⟩ est segmenté en deux dans la référence.

(75) အခြား ငှက် မျိုး များ
autre oiseau espèce PL
⟨ʔkʰўaˀ⟩ ⟨ŋȟkˣ⟩ ⟨mўiuˀ⟩ ⟨mўaˀ⟩
'd'autres espèces d'oiseau'

Alors que dans l'exemple suivant (76), la chaîne de caractères n'est pas découpée. Dans l'esprit de la personne qui a fait la segmentation de la référence, il est possible que compter quelque chose (le deuxième မျိုး ⟨mўiuˀ⟩ est utilisé comme un classificateur pour compter) implique que ငှက်မျိုး ⟨ŋȟkˣmўiuˀ⟩ soit considéré comme une unité, alors que le marqueur du pluriel n'a pas cet effet dans l'exemple précédent.

(76) ငှက်မျိုး တစ် မျိုး
oiseau.espèce un CL
⟨ŋȟkˣmўiuˀ⟩ ⟨tśˣ⟩ ⟨mўiuˀ⟩
'une espèce d'oiseau'

D'autres erreurs de ce type sont discutables et il est difficile de voir quel choix est valable. Doit-on segmenter အမျိုးသားဒီမိုကရေစီအဖွဲ့ချုပ်ပါတီ (*parti de la Ligue nationale pour la démocratie*) en အမျိုးသားဒီမိုကရေစီအဖွဲ့ချုပ် (*la Ligue nationale pour la démocratie*) + ပါတီ (*parti*) ? L'erreur réside plutôt dans le manque de cohérence dans la référence que dans le choix de segmentation lui-même.

Il semble qu'il y ait aussi des incohérences dans la segmentation de verbes et expressions adverbiales. Dans la référence, une expression qui veut dire *pendant longtemps* ရှည်ကြာစွာ ⟨rȟŋˣkў̆aśw̆a⟩ est tantôt segmentée en trois ရှည် ကြာ စွာ ⟨rȟŋˣ⟩ ⟨kў̆a⟩ ⟨św̆a⟩ (*long* + *durer* + particule pour former les adverbes) tantôt en deux ရှည်ကြာ စွာ ⟨rȟŋˣkў̆a⟩ ⟨św̆a⟩ (expression verbale *qui dure longtemps* + particule pour former les adverbes). Les cas d'expressions avec စွာ ⟨św̆a⟩, une particule pour former les adverbes, devraient être tous traités de la même manière. Cette syllabe peut être présentée comme une particule qui termine un syntagme adverbial (par exemple par Okell et Allott (2017)) ou un suffixe productif qui forme des adverbes (Bernot, Cardinaud et al. 2001). La dernière interprétation ne segmenterait pas du tout la chaîne de caractères.

Certaines différences entre la référence et le résultat de notre segmentation avec *Motor* proviennent de nos choix de traitement. Les chiffres par exemple

(27 erreurs sur 102), sont tous découpés par *Motor*, car nous voulons réduire le vocabulaire du résultat. Les nombres constituent un ensemble potentiellement infini, on ne peut les ajouter au dictionnaire, mais on aurait pu décider de les ignorer lors de la segmentation. Tous les chiffres étant segmentés en caractères individuels, ၂၂၄.၆၅ ⟨224.69⟩ devient ၂ ၂ ၄ . ၆ ၅ ⟨2 2 4 . 6 9⟩. Par ailleurs, les chiffres étant complètement transparents pour les étudiants, il n'est pas utile de les inclure dans nos listes de fréquence. Ce procédé rend leur élimination très facile. Il ne s'agit pas à proprement parler d'erreurs de segmentation.

Nous avons rencontré plusieurs types de différences entre la référence et notre résultat de segmentation que l'on peut qualifier d'erreurs dues à l'outil de segmentation.

Deux d'entre elles ont été résolues grâce à des modifications faites à l'outil lui-même. La première (8 erreurs sur 102) provient d'une limitation dans la longueur des entrées du dictionnaire. Au-delà d'une certaine longueur, même si un mot est dans le dictionnaire, *Motor* les segmente. Par exemple ဗမာနိုင်ငံလုံးဆိုင်ရာကျောင်းသားသမဂ္ဂ (*Union de tous les étudiants de Birmanie*) n'est pas segmenté dans la référence, *Motor* nous donne comme résultat ဗမာ နိုင်ငံ လုံး ဆိုင်ရာ ကျောင်းသားသမဂ္ဂ. Ce problème a été résolu en portant la longueur maximale à 60 caractères. Le deuxième type est potentiellement plus problématique, et concerne l'algorithme de *Motor* qui favorise les chaînes les plus longues d'abord (FMM) (11 erreurs sur 102). On peut corriger ce problème manuellement en ajoutant des espaces dans le corpus, mais ce n'est pas une solution très satisfaisante. L'autre problème que nous avons rencontré concerne la ponctuation à l'intérieur des mots, notamment les parenthèses. *Motor* a été modifié pour permettre la ponctuation à l'intérieur des entrées de son dictionnaire (voir la section sur les prétraitements simples du chapitre précédent pour plus de détails).

Le seul type de différence entre la segmentation de la référence et la segmentation de *Motor* dû à l'algorithme de *Motor* qui ne peut être résolu définitivement concerne la préférence que l'algorithme donne aux chaînes de caractères longues sur les chaînes courtes dans une phrase. Dans la référence, nous avons des instances de အလုပ်လုပ်ကိုင် ⟨ʔlupˣlupˣkiuŋˣ⟩, cette chaîne existe dans notre dictionnaire de test de *Motor*. Là où la référence segmente la chaîne de caractères အလုပ်လုပ်ကိုင် ⟨ʔlupˣlupˣkiuŋˣ⟩ en အလုပ် လုပ်ကိုင် ⟨ʔlupˣ⟩ ⟨lupˣkiuŋˣ⟩, *Motor* va segmenter en အလုပ်လုပ် ကိုင် ⟨ʔlupˣlupˣ⟩ ⟨kiuŋˣ⟩, même si အလုပ် ⟨ʔlupˣ⟩ (*travail*) et

လုပ်ကိုင် ⟨lup^{x}kiuŋ^{x}⟩ (*faire qqch pour vivre, s'engager à faire qqch*) figurent aussi dans son dictionnaire, puisque အလုပ်လုပ် ⟨ʔlup^{x}lup^{x}⟩ (*travailler*) et plus long que အလုပ် ⟨ʔlup^{x}⟩. Dans ces cas, nous avons recours à une solution peu satisfaisante, l'ajout manuel d'espaces avant la segmentation automatique pour forcer la segmentation correcte. La plupart des erreurs de ce type que nous avons rencontrées dans nos corpus concerne la particule de négation မ ⟨m⟩ qui se trouve collée au mot précédent dans des contextes où elle ne devrait pas l'être. Par exemple, သူမ ⟨θùm⟩ dans le sens *il* + NEG est systématiquement sous-segmenté သူမ ⟨θùm⟩ (*elle*) au lieu de သူ မ ⟨θù⟩ ⟨m⟩, tout comme ဆရာမ ⟨s̀^{h}ram⟩ est aussi sous-segmenté ဆရာမ *professeure* au lieu de ဆရာ မ ⟨s̀^{h}ra⟩ ⟨m⟩ *professeur* + NEG. Ce cas peut être résolu si une espace typographique est insérée le cas échéant avant မ ⟨m⟩ avant segmentation. Il faut noter que nous n'avons que onze erreurs de ce type, ce qui veut dire que ce type d'erreur ne concerne qu'une vingtaine de tokens dans un corpus de plus de vingt mille tokens, si le dictionnaire est complet.

Un autre type d'erreur difficile à corriger concerne les ambiguïtés lexicales propres à la forme écrite, erreurs qui sont souvent désambiguïsées à l'oral grâce à des pauses, l'accentuation ou l'intonation. Ce problème n'est pas du tout limité au birman[10], mais l'absence de délimiteurs lexicaux évidents et la nécessité de segmenter le texte pour le traiter imposent souvent une interprétation là où dans une autre langue une ambiguïté subsisterait. Par exemple, la chaîne de caractères birmans suivante မိန်းကလေးကစားတယ် ⟨min^{x2}kle^{2}kə̀a^{2}ty^{x}⟩ peut s'interpréter de deux manières selon la segmentation မိန်းကလေး ကစား တယ် ⟨min^{x2}kle^{2}⟩ ⟨kə̀a^{2}⟩ ⟨ty^{x}⟩ (trois tokens, *La fille joue*) où မိန်းကလေး က စား တယ် ⟨min^{x2}kle^{2}⟩ ⟨k⟩ ⟨ə̀a^{2}⟩ ⟨ty^{x}⟩ (quatre tokens, *La fille mange*). Si la chaîne de caractères s'écrit sans espaces typographiques, *Motor* va toujours choisir la première solution, qui a moins de tokens. Dans les faits, et dans ce cas précis le caractère က ⟨k⟩, quand il est employé comme marqueur du sujet, comme dans notre deuxième exemple, est toujours suivi d'une espace typographique ainsi မိန်းကလေးက စားတယ် ⟨min^{x2}kle^{2}k ə̀a^{2}ty^{x}⟩,

10. En anglais par exemple, la phrase *She gave her dog biscuits.* peut s'interpréter de plusieurs manières : soit elle a donné des biscuits à son chien, soit elle a donné des croquettes pour chien à quelqu'un (*dog biscuits* est un mot composé potentiel). Si le lecteur interprète la phrase selon la deuxième interprétation au lieu de la première, il s'agit d'un cas de sous-segmentation. Le tiret court est parfois utilisé pour signaler les mots composés *dog-biscuits*, mais son usage est limité aux noms composés en fonction épithète (on écrit *dog-biscuit sales* mais *sales of dog biscuits*). Sans le point à la fin de la phrase on pourrait même interpréter la signification comme étant *elle a donné ses propres biscuits pour chien.*

la bonne segmentation est alors assurée.

3.2.4.2 Test avec dictionnaire général

Notre deuxième évaluation concerne le dictionnaire que nous avons entraîné sur nos corpus. Le contenu du dictionnaire est assez important, plus de soixante-douze mille entrées, mais n'inclut pas spécifiquement le vocabulaire du texte de référence utilisé pour réaliser le test. Les résultats sont relativement décevants, surtout en ce qui concerne la précision, seulement 0,853% des résultats sont corrects. En revanche, le rappel de tokens corrects s'élève à 0,962%. Combiné ensemble en F-mesure, le résultat n'est pas très bien, mais meilleur que ce à quoi l'on pouvait s'attendre, 0,904. En plus des difficultés mentionnées dans la section précédente, le test avec dictionnaire général souffre particulièrement de l'absence du dictionnaire de certaines entités nommées du texte de référence et l'incohérence du traitement de celles-ci. Par exemple, la séparation de nom et prénom n'est pas systématique dans le corpus de référence. A titre d'exemple, *Steve Wozniak*[11] est un seul token dans la référence စတိဗ်ဝေါ့ဇနစ် ⟨śtib^{x}weä^{1}żnś^{x}⟩, alors que les deux éléments du nom sont des entrées séparées du dictionnaire de *Motor*, *Steve* စတိဗ် ⟨śtib^{x}⟩ et *Wozniak* ဝေါ့ဇနစ် ⟨weä^{1}żnś^{x}⟩. Choix que nous assumons qui provient du souhait de réduire la taille du dictionnaire de segmentation là où c'est possible. Ceci concerne surtout les toponymes, dans la mesure où nous séparons le toponyme du nom qu'il décrit. Nous séparons donc ကချင်ပြည်နယ် ⟨kk^{h}ǔŋ^{x}pǰɲ^{x}ny^{x}⟩ *l'état kachin* en deux entrées ကချင် ⟨kk^{h}ǔŋ^{x}⟩ *kachin* et ပြည်နယ် ⟨pǰɲ^{x}ny^{x}⟩ *l'état*. Ce dernier va servir à segmenter d'autres noms d'état, tout comme l'entrée ကချင် ⟨kk^{h}ǔŋ^{x}⟩ *kachin* va servir à segmenter non seulement *l'état kachin*, mais aussi ကချင်ဘာသာစကား ⟨kk^{h}ǔŋ^{x}ba̰θaśka^{2}⟩ *la langue kachin*, ကချင်လူမျိုး *l'ethnie kachin* ⟨kk^{h}ǔŋ^{x}lùmǔiu^{2}⟩ et ainsi de suite. Lors du test, မဟာရဋ္ဌပြည်နယ် ⟨mhar̰ɪ̈ǰɲ^{x}ny^{x}⟩ *l'état du Maharashtra* a donc été scindé en deux, même si မဟာရဋ္ဌ ⟨mhar̰ɪ̈⟩ *Maharashtra* figure bien dans le dictionnaire de segmentation.

Ces tests démontrent l'importance de la cohérence de la segmentation lors de la création de corpus de référence, non seulement pour les cas spécifiques, mais aussi pour les groupes de tokens de même nature, noms propres de personnes et toponymes, mais aussi les chaînes de caractères auxquelles on peut attribuer une même fonction syntaxique. La segmentation au moyen d'un outil à dictionnaire

11. Un des trois cofondateurs de la société *Apple*

a l'avantage de donner un résultat uniforme, tant que le dictionnaire est compilé en tenant compte de ces aspects. Le gros inconvénient de cette approche est que l'outil repose sur une liste finie, alors que le vocabulaire d'une langue est potentiellement infini. Ceci est particulièrement problématique quand il s'agit de segmenter des textes comme les articles de presse contenant un grand nombre de noms propres dont la fréquence d'apparition varie avec le temps. Dans ces cas la correction de corpus et l'ajout de ces noms deviennent très importants pour s'assurer d'une segmentation acceptable.

3.2.5 *La correction de corpus à l'aide de la segmentation*

Un autre avantage de l'utilisation d'un outil de segmentation à dictionnaire réside dans son usage pour identifier les erreurs systématiques à corriger dans le corpus même. Cette correction est primordiale pour notre projet, car les résultats préliminaires de nos calculs de distribution de types sur un ensemble composé de nos propres corpus authentiques (BBC Burmese, Lotaya et TED Talks) démontrent l'importance de l'effet des erreurs de segmentation provoquées par les fautes dans les corpus.

L'erreur de substitution graphique discutée dans le chapitre précédent, la substitution du chiffre zéro ဝ ⟨o⟩ U+1040 MYANMAR DIGIT ZERO pour la lettre ဝ ⟨w⟩ U+101D MYANMAR LETTER WA, illustre bien comment les erreurs orthographiques et typographiques deviennent apparentes après la segmentation. Puisque ces tokens erronés ne sont pas inclus dans le dictionnaire de *Motor* (il serait très ardu voir impossible de prévoir toutes les fautes de frappe possible dans une langue), elles provoquent des erreurs de segmentation. A titre d'exemple, même si l'entrée ဝန်ကြီးချုပ် ⟨wn^{x}kǰï^{2}k^{h}y̆up^{x}⟩ (*premier ministre*) est inclus au dictionnaire de *Motor*, l'usage de ဝ ⟨o⟩ au lieu de la lettre ဝ ⟨w⟩ aura pour résultat la segmentation erronée ဝ နဲ့ ကြီး ချုပ် ⟨o n ^{x} kǰï^{2} k^{h}y̆up^{x}⟩.

Certains caractères qui n'apparaissent jamais seuls dans une syllabe birmane bien formée ont non seulement une fréquence élevée, mais aussi une distribution importante. A titre d'exemple, le caractère ် ⟨^{x}⟩ est en 96^{e} position de la liste des 25 000 types de ce corpus de textes authentiques après calcul de la fréquence globale moyenne réduite et normalisée[12].

12. Il s'agit de la méthode de Čermák et Křen (2011) qui prend en compte la fréquence, la dispersion et les tailles relatives de chaque corpus, détaillée dans le chapitre suivant.

Si certaines erreurs comme les fautes de frappe identifiées dans la section 2.5.4 peuvent se corriger simplement par substitution, d'autres erreurs de segmentation doivent être regardées de plus près, car ce qui paraît être une seule erreur indique souvent une unité polysyllabique mal segmentée.

La première méthode d'identification d'erreurs se fait à l'aide d'un logiciel d'analyse de données textuelles (ou textométrie) compatible avec le système d'encodage informatique birman comme TXM (Heiden et al. 2010)[13] ou Lexico5 (Lebart et Salem 1994)[14]. Il suffit de générer le lexique d'un corpus pour constater les erreurs les plus évidentes de segmentation. La liste est constituée de tous les tokens du corpus, chacun accompagné de leur fréquence brute du nombre d'occurrences du token dans le corpus. L'ordre alphabétique regroupe beaucoup d'erreurs de syllabes malformées à la fin. Voici (figure 3.7) un extrait du lexique de nos corpus authentiques réunis ensemble après une segmentation préliminaire (3 242 634 tokens) :

Le premier constat est que certains de ces tokens sont rares et témoignent d'une erreur peu fréquente qui est probablement peu significative d'un point de vue statistique, alors que d'autres tokens sont très fréquents et potentiellement auront une incidence sur la fiabilité de nos calculs statistiques ultérieurs. Quand il n'y que quelques occurrences, il s'agit souvent d'une seule erreur. Par exemple, le token ◌္◌ (◌ ⟨ ˭ ⟩ U+1039 MYANMAR SIGN VIRAMA et ◌ီ ⟨ï⟩ U+102E MYANMAR VOWEL SIGN II), qui n'apparaît que cinq fois, n'apparaît qu'à l'intérieur d'un seul mot, အင်္ကျီ ⟨ʔŋ˭kўï⟩ (*chemise*), écrit အင်္ကြီ ⟨ʔkўŋ˭ï⟩ dans le corpus et segmenté အင်္ကျင် ◌္◌ီ ⟨ʔkўŋˣ⟩ ⟨ ˭ ï⟩. Il s'agit soit d'une faute de frappe, soit d'une erreur de transcodage, dans les deux cas à corriger dans le texte d'origine.

Les tokens fréquents qui constituent des syllabes mal formées, comme ◌် ⟨ˣ⟩ qui apparaît 2750 fois, signalent plusieurs types d'erreurs différents. Pour examiner le détail des erreurs, nous générons le concordancier du token, ici, nous prenons le cas de ◌် ⟨ˣ⟩ comme exemple[15]. Le premier type d'erreur, comme pour l'exemple précédent, est la faute de frappe évidente à corriger dans le texte d'origine. Dans la figure 3.8, les segments sursegmentés sont en gras. Dans l'exemple de la première ligne, လျှောက် ⟨lўňeakˣ⟩ (*se promener*) manque le caractère ◌ေ ⟨e⟩. Dans

13. http://textometrie.ens-lyon.fr/?lang=fr
14. http://lexi-co.com/index.html
15. Il s'agit du seul caractère U+103A MYANMAR SIGN ASAT qui ne peut paraître seul dans une syllabe bien formée.

lemme	Fréquence
◌	773
◌စ	1
◌စင်း	3
◌စိန်ခေါ်မှု	1
◌ဆောင်ရွက်	1
◌တ	2
◌ထ	1
◌ဓ	1
◌ပုံနှိပ်	1
◌မ	1
◌ဝယ်	1
◌ာ	2
◌ား	2
◌ီ	5
◌်	2750
◌်ကူညီ	1
◌်တစ်	1
◌်တပ်	1

FIG. 3.7 : Extrait du lexique d'un corpus après segmentation préliminaire

l'exemple de la deuxième ligne ◌ိုင်ံ ⟨niuŋˣŋǩ⟩ (*nation, pays*), où les caractères ◌ ⟨ŋ⟩ et ◌် ⟨ˣ⟩ ont malencontreusement été insérés au milieu du mot qui se trouve scindé en deux. Nous ne voulons pas inclure des erreurs dans le dictionnaire de notre outil de segmentation (un chercheur qui voudrait étudier les erreurs aurait peut-être choisi de les inclure), il faut donc corriger les textes d'origine, puis refaire la segmentation.

text_id	Contexte gauche	Pivot	Contexte droit
See_your_stories…	မ်း **လျှာ ကာ**	◌်	နေ တုန်း ဒါ ကို
Rebuilding_Myanm…	အမေရိကန် _ **နိုင်** ◌	◌်	**ံ** မှာ _ ရရှိ

FIG. 3.8 : Extrait de concordancier d'erreurs de segmentation dues à une faute de frappe

Un autre type de faute de frappe qu'il faut corriger dans les textes d'origine avant de resegmenter concerne l'usage du chiffre zéro birman, ◌ ⟨o⟩, à la place

du caractère consonantique ဝ ⟨w⟩, désigné U+101D, MYANMAR LETTER WA en Unicode. Dans l'exemple de la figure 3.9, cette substitution a provoqué une sursegmentation du verbe ဝင် ⟨oŋˣ⟩ écrit avec le chiffre (verbe polysémique, qui a le sens primaire d'*entrer, participer*), car seul ဝင် ⟨wŋˣ⟩ avec le caractère ဝ ⟨w⟩ figure dans le dictionnaire de *Motor*.

text_id	Contexte gauche	Pivot	Contexte droit
1988	စားသောက်ဆိုင် ထဲ **o c**	်	ပြီး စ်ခတ် ခဲ့ တဲ့

FIG. 3.9 : Extrait de concordancier d'erreurs de segmentation dues à la substitution de WA par zéro

La correction de cette erreur se fait en tenant compte de la différence entre l'environnement de la lettre ဝ ⟨w⟩ et l'environnement du chiffre ဝ ⟨o⟩ dans les textes. Là aussi, un concordancier avec le chiffre zéro birman comme mot pôle est très utile. Après avoir examiné le chiffre zéro birman à l'aide d'un concordancier, nous n'avons trouvé aucun exemple du caractère du chiffre zéro seul, il est toujours suivi ou précédé par un autre caractère de chiffre, la virgule (comme ici ၁ ၈, ဝ ဝ ဝ ⟨1 8 , o o o⟩), par un caractère de ponctuation utilisé pour noter l'heure (၀ ၂ : ဝ ဝ ⟨o 2 : o o⟩) ou bien un point comme dans les pourcentages (၀. ဝ ၇ % ⟨o. o 7 %⟩). Tous les autres caractères sont remplacés par la lettre ဝ ⟨w⟩ par notre script de nettoyage.

La majorité des erreurs de segmentation est provoquée par des lacunes du dictionnaire. Si un mot n'y figure pas, il est segmenté en plusieurs tokens au lieu d'un. La figure 3.10 montre la segmentation d'une ligne de texte quand le nom propre ဟိုင်နန် ⟨hiuŋˣnnˣ⟩, *Hainan* n'est pas dans le dictionnaire — il est segmenté en quatre tokens au lieu d'un seul.

text_id	Contexte gauche	Pivot	Contexte droit
The_future_of …	နောက်ဆုံး _ **ဟို c**	်	**နန်** ကျွန်း က _

FIG. 3.10 : Extrait de concordancier d'erreurs de segmentation dues au dictionnaire incomplet

L'erreur disparaît après l'ajout au dictionnaire et resegmentation. Si dans le cas du token ် ⟨ˣ⟩, il s'agit clairement d'une erreur, car la syllabe n'est pas bien

formée, il n'en est moins sûr pour les trois autres tokens င ⟨ŋ⟩, ဟိုဲ့ ⟨hiu⟩ et ၦၣ် ⟨nnˣ⟩, car il s'agit de syllabes bien formées que l'on ne peut détecter automatiquement. Les deux syllabes င ⟨ŋ⟩ et ၦၣ် ⟨nnˣ⟩ sont relativement rares et souvent font partie d'autres mots (U Po Hla 2008 ; Myanmar Language Commission 2008), la syllabe ဟိုဲ့ ⟨hiu⟩ est utilisée comme un pronom ou déterminant démonstratif dans la langue parlée (*ça, ce*). Cet exemple illustre pourquoi il est difficile de créer des règles de correction. En effet, l'extraction à base de règles ou par la méthode de concordancier est assez fastidieuse et chronophage (résumé par la figure 3.11).

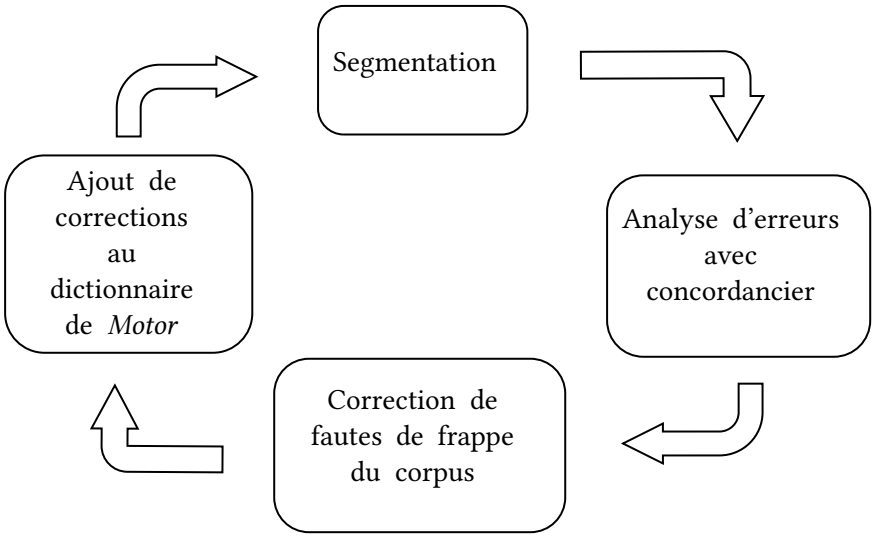


FIG. 3.11 : Processus d'identification d'erreurs à l'aide de la segmentation

3.2.6 *Les contraintes de l'outil de segmentation*

Nous avons déjà mentionné la régularité de segmentation comme principal avantage de la méthode de segmentation à base de dictionnaire, mais ceci repose sur l'existence préalable du token dans le dictionnaire, une liste finie. Or, le lexique d'une langue est potentiellement infini et en constante mutation. Si le noyau central du lexique est stable, une partie importante de ses éléments moins courants, néologismes, mots empruntés et noms propres, est en constante évolution. Le pourcentage de ces éléments dans un corpus donné a donc une influence sur la fiabilité de la segmentation.

Puisque le dictionnaire interne sur lequel *Motor* s'appuie ne peut pas être infini, il faudrait faire des choix qui prennent cet aspect en compte tout en réduisant la taille du dictionnaire. Ces choix sont présentés dans les sections qui suivent.

3.3 La définition de tokens

Compte tenu de la paucité des études basées sur corpus en birman, il n'existe pas encore de règles explicites, ou même de recommandations pour segmenter du texte birman en tokens pour le traitement automatique. À ceci s'ajoute la difficulté que les Birmans ne semblent pas avoir une notion intuitive de *mot* (Ding, Ye Kyaw Thu et al. 2016), ce qui rend le désaccord entre locuteurs natifs sur l'évaluation de la segmentation inévitable. Nomoto et al. (2018) remarquent que les principes utilisés dans la construction du corpus birman du *Asian Language Treebank*, qui sont basés sur les étiquettes morphosyntaxiques du corpus, laissent trop de latitude aux jugements personnels des annotateurs dans la définition de tokens. Ceci s'explique en partie par le fait que le lexique du birman contient un grand nombre d'éléments que l'on peut considérer comme des mots composés analysables à partir de leurs composants, comme par exemple လူကြီး /lùtɕí/ (*personne* + *grand* = *adulte*) ou မိဘ /mḭbə̰/ (*mère* + *père* = *parents*). Même si la difficulté de cerner exactement les *mots* en birman peut paraître gênante tant pour l'étudiant que pour le taliste, en fait ceci a l'avantage de laisser une certaine latitude dans la définition des tokens. Nous sommes en effet libre de définir les tokens selon nos besoins, prenant en compte les limitations de notre outil de segmentation *Motor*.

Nous avançons l'idée qu'il n'y a pas de segmentation en tokens idéale qui conviendrait à tous les usages de corpus possibles. La segmentation devrait fournir des tokens cohérents entre eux qui conviennent à l'étude en question et les analyses faites dessus doivent prendre en compte les spécificités de la segmentation. Dans le cas précis, cela veut dire que les tokens qui résultent de la segmentation doivent avoir une utilité pour les apprenants du birman langue étrangère et être cohérent par rapport aux pratiques d'explication de la langue par les formateurs en birman langue étrangère.

Pour ce faire, nous ne cherchons pas à avancer des hypothèses nouvelles sur la langue birmane — autant que possible nous nous contentons de nous inspirer

des travaux antérieurs, en lexicologie du birman, en création de corpus birman et en didactique du birman langue étrangère, pour nous aider à définir les tokens. Nous effectuons aussi des analyses statistiques sur nos corpus pour dégager des régularités. Concrètement, nos observations résultant d'analyses de ces travaux informent les choix d'entrées dans le dictionnaire de *Motor*, qui sert à segmenter nos corpus. Nous analysons ensuite le résultat de cette segmentation pour repérer des erreurs et des incohérences dans la segmentation afin d'en trouver les causes et y remédier.

En définissant nos tokens, nous gardons à l'esprit quelques observations et objectifs :

1. Le dictionnaire de *Motor* est une liste finie.
2. Plus la chaîne de caractères du token est longue, plus la segmentation est précise.
3. Le vocabulaire résultant doit être utile pour les apprenants.
4. Le vocabulaire résultant doit être le plus cohérent possible.
5. Le vocabulaire résultant doit correspondre le plus possible avec les outils pédagogiques et dictionnaires existants.

D'emblée, quelques lignes directrices se dégagent. Les éléments des catégories morphosyntaxiques qui constituent des listes finies, telles que les pronoms et les éléments qui indiquent les relations syntaxiques (les marques de fin de phrase, les auxiliaires et les marques verbales, les marqueurs syntaxiques et les subordonnées[16]) sont faciles à répertorier et ajouter au dictionnaire de *Motor*, alors que les éléments de catégories avec un nombre important et imprévisible d'éléments posent problème.

Nous avons déjà remarqué qu'il est plus commode de segmenter les nombres en chiffres simples pour écarter le besoin de le rajouter au dictionnaire. De la même manière, pour éviter de rajouter les formes et au singulier et au pluriel de tous les membres de classes de vocabulaire très étendues telles que les verbes et les noms, nous rajoutons les marqueurs du pluriel au dictionnaire. Ceci nous permet non seulement de réduire la taille du dictionnaire, mais a pour l'effet d'effectuer en même temps une lemmatisation qui assimile les occurrences de la forme du pluriel au singulier.

16. Nous utilisons où possible la nomenclature de Bernot, Cardinaud et al. (2001).

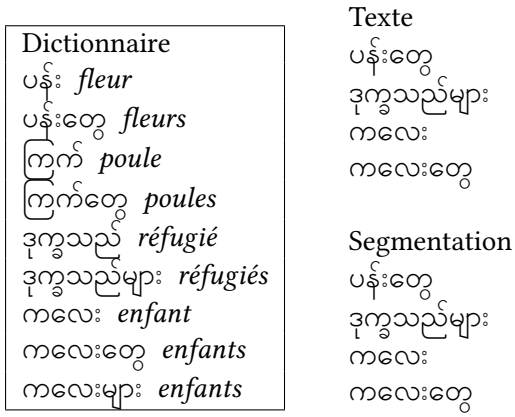


FIG. 3.12 : Dictionnaire de *Motor* avec formes au pluriel

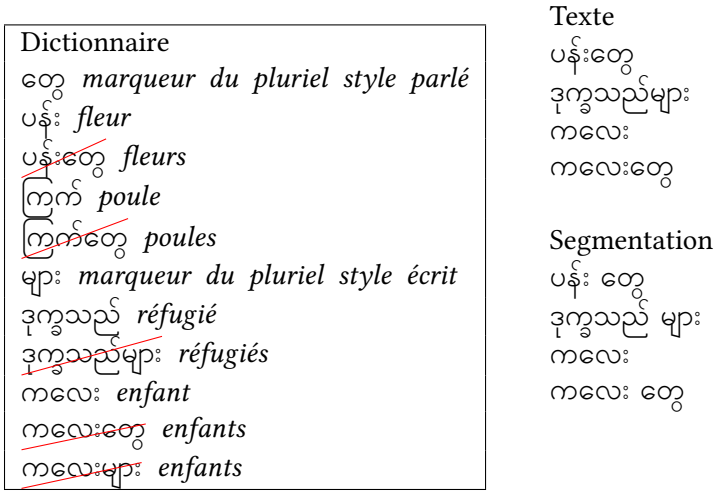


FIG. 3.13 : Dictionnaire de *Motor* sans formes au pluriel

Ce choix est particulièrement commode en ce qui concerne la segmentation de textes comportant des énumérations, car deux noms peuvent partager le même marqueur de pluriel, de la manière suivante :

(77) ကျွဲ ၊ နွားတွေ
/kjwɛ́ ၊ nwá-twè/
buffle , taureau-PL
'buffles, taureaux'

Le système nous permet toutefois d'inclure certaines formes au pluriel, ce qui serait souhaitable pour les formes plurielles irrégulières qui doivent être apprises séparément par un apprenant, comme မိဘ /mḭbə̰/ (*parents*), ou bien les formes plurielles nettement plus fréquentes que la forme au singulier. C'est le cas de လက်ခုပ်သံများ ⟨lkˣkʰupˣθǎmǐya²⟩ /lɛʔkʰoʊʔθàɴmjá/ (*applaudissements*), များ ⟨mǐya²⟩ /mjá/ indiquant le pluriel.

Le traitement d'éléments appartenant à d'autres catégories morphosyntaxiques s'avère plus délicat. D'autant plus que la définition de catégories morphosyntaxiques en birman ne fait pas l'unanimité, et que la définition des catégories a une incidence sur la segmentation.

3.3.1 *La syllabe comme unité de base*

En raison du manque de frontières lexicales explicites et de la nature alphasyllabique de son système d'écriture, le locuteur natif du birman est particulièrement attentif aux frontières syllabiques. Ceci est renforcé par la pratique de la lexicographie birmane qui trie les entrées de dictionnaires par syllabe et non par lettre (Tin Htay Hlaing et Mikami 2011). Par exemple, dans le *Myanmar-English Dictionary* မြန်မာ–အင်္ဂလိပ် အဘိဓာန်, l'entrée ဝရန်တာ ⟨wɹ̀nˣta⟩ (prononcé /wə̰jàɴtà/, *balcon*), qui a trois syllabes ဝ ⟨w⟩, ရန် ⟨ɹ̀nˣ⟩ et တာ ⟨ta⟩, est rangé avant ဝက် ⟨wkˣ⟩ (*cochon, moitié*), bien que la lettre က ⟨k⟩ précède la lettre ရ ⟨ɹ̀⟩ dans l'alphasyllabaire birman (voir tableau A.1).[17] Cette influence pourrait s'estomper

17. Les dictionnaires qui rangent leurs entrées autrement que par ordre alphabétique peuvent ralentir la recherche de vocabulaire inconnu et s'avérer frustrants pour les apprenants de langues étrangères, surtout quand il s'agit de systèmes d'écriture complexes et difficiles à maîtriser. Pour palier aux inconvénients des dictionnaires traditionnels chinois organisés par clé de caractère, le dictionnaire de DeFrancis (1996) range les entrées, des lexèmes au lieu de caractères, par sa prononciation notée en hànyǔ pīnyīn, un agencement particulièrement utile pour la compréhension de l'oral. Dans la même veine, mais avec un système moins standardisé et très souple, le dictionnaire de Oka (2002), destiné aux apprenants japonais du thaï, range les entrées phonétiquement en

avec le recours croissant aux dictionnaires en ligne, consultés par requête (ce qui d'une manière générale diminue inéluctablement l'importance de la connaissance de l'ordre alphabétique), car ceux-ci affichent leurs résultats en ordre de point de code Unicode, c'est-à-dire par lettre. En Unicode, ဝက် ⟨wkˣ⟩ (*cochon, moitié*) est rangé avant ဝရန်တာ ⟨wɹ̀nˣta⟩ (*balcon*), car le point de code de la lettre က ⟨k⟩ U+1000 MYANMAR LETTER KA précède celui de ရ ⟨ɹ̀⟩ U+101B MYANMAR LETTER RA.

La syllabe est de fait la seule unité linguistique en birman qui se rapproche du mot-forme, bien identifiable par le lecteur en vertu du système d'écriture. La première question que nous nous sommes posée est de savoir s'il est approprié de segmenter tout simplement en syllabes.

Le premier problème d'une telle approche est, bien sûr, le traitement des noms étrangers et du vocabulaire emprunté qui ne sont évidemment pas tous monosyllabiques. Si l'on prend par exemple le corpus biblique parallèle de Christodouloupoulos et Steedman (2014), le nombre de types en anglais est de 14 565, mais le texte équivalent en birman (qui, segmenté en syllabes, a la taille de 1 300 711 tokens) ne compte que 2 284 types, ce qui semble extraordinairement peu en comparaison. Pris ensemble, nos trois corpus authentiques nous donnent un échantillon de langue de plus de quatre millions de syllabes, mais avec seulement 5 619 syllabes différentes.

Plus important encore, nous voulons aussi faire attention à ne pas simplifier inutilement la représentation du lexique birman. La segmentation en tokens revient à supprimer le contexte qui l'entoure et donc à simplifier ou plus exactement à généraliser son sens. Une segmentation trop brutale risque également d'exagérer la tendance des mots plus fréquents à être plus polysémiques, un phénomène démontré par Zipf (1945). Autrement dit, une segmentation en syllabes pourrait créer une simplification dans la présentation du lexique, dans la mesure où nous aurions moins de types (tokens différents) et moins d'information concernant les relations sémantiques entre eux. Prenons un exemple simple. Pour dire *musée* en birman, on utilise ပြတိုက် ⟨pǰtiukˣ⟩, qui est composé de deux syllabes ပြ ⟨pj⟩ *montrer* et တိုက် ⟨tiukˣ⟩ *bâtiment*. Si nous segmentons en syllabes, ပြတိုက် → ပြ တိုက်, le token ပြတိုက် ⟨pǰtiukˣ⟩ *musée* disparaît et nous avons une occurrence de plus de

prenant en compte non seulement les allophones que l'on observe dans la langue thaïe parlée, mais aussi les allophones japonais. Par exemple, les entrées qui commencent avec ห ⟨h⟩ sont rangé sous *f*, car un apprenant japonais peut confondre les sons /h/ et /f/.

chacun de ပြ ⟨pj⟩ *montrer* et တိုက် ⟨tiuk^{x}⟩ *bâtiment*. De plus, ces deux syllabes n'ont ces significations que dans certains contextes. Quand ces deux syllabes sont contiguës, le sens de *montrer* est activé, car la syllabe se trouve à côté de la syllabe တိုက် ⟨tiuk^{x}⟩, qui inversement a le sens *bâtiment* parce qu'elle se trouve après la syllabe ပြ ⟨pj⟩. Si nous regardons les entrées pertinentes dans le *Myanmar-English Dictionary* မြန်မာ–အင်္ဂလိပ် အဘိဓာန်, MED, (voir la figure 3.14), nous voyons que les définitions pour ces deux syllabes sont nombreuses et ne sont pas toutes liées sémantiquement. Nous prenons donc le risque de confondre des homographes inutilement.

3.3.2 *Le vocabulaire des manuels de birman*

Les avantages techniques et la tradition lexicographique ne signifient donc pas pour autant que la segmentation en syllabes soit appropriée pour une étude sur le vocabulaire à destination des apprenants de birman langue étrangère. Au lieu d'aborder le problème d'un point de vue théorique, nous avons d'abord observé le vocabulaire tel qu'il est présenté aux apprenants usagers des ressources didactiques détaillées dans le chapitre précédent. La longueur en syllabes des unités linguistiques présentées par chaque ressource est calculée utilisant l'outil `sylbreak` de Ye Kyaw Thu décrit dans la section 2.4.

Nous dressons d'abord une liste du vocabulaire de chaque ressource, chaque vocable différent n'est donc inclus qu'une fois. Pour illustrer la présentation des données, prenons l'exemple de l'analyse de *Burmese by Ear : or Essential Myanmar* (Okell 2014) (dorénavant BBE), illustrée dans par le tableau 3.15. 23% des vocables n'ont qu'une syllabe, 39,1% en ont deux, 23% en ont trois et ainsi de suite.

Les mêmes données sont représentées sous forme de diagramme à barres (voir figure 3.16), pour mieux visualiser les différences entre les vocables de longueurs différentes. On constate que l'auteur de l'ouvrage, destiné toutefois aux apprenants débutants, ne privilégie pas le vocabulaire monosyllabique pour présenter le vocabulaire (23% seulement) et que ce sont les vocables à deux syllabes qui dominent (39,1% du vocabulaire). Une comparaison entre ressources nous donne des pistes de réflexion pour comprendre comment le vocabulaire est présenté aux apprenants du birman langue étrangère et les raisons pour lesquelles les élaborateurs de ces ressources auraient pu faire ces choix.

ပြ
1 v show; display; exhibit; indicate (as in အလှ- ၊ဗန်း-).
2 v teach; direct; instruct; demonstrate (as in စာ- ၊နည်း-).
3 v expose sth to (as in နေ- ၊မိး-).
4 *SAME AS* ပြအိုး n.
5 n (a) distance between two adjacent turrets on the palace walls; (b) [Mandalay] unit of distance measure based on 2(a) (approximately one furlong).
6 PART particle suffixed to a verb to convey the sense of demonstrating or relating sth (as in လုပ်- ၊စား- ၊ပြော-).

တိုက်
1 v dash against; bump against/ into; crash; collide.
2 v blow (as in လေ-); sail (as in ရွက်-).
3 v attack; fight (as in -ခိုက်).
4 v rob (as in ဓားပြ-).
5 v incite; set on (as in ရန်-).
6 v brush (as in သွား-); polish, shine (as in ဖိနပ်-); scrub (as in ကြမ်း-); scour (as in အိုး-).
7 v (a) coincide (as in အစည်းအဝေးနဲ့-နေလို့မလာဖြစ်ဘူး။); synchronize, check (as in ဒီနာရီမနက်ကပ်-ထားတာ။); (b) check against; verify (as in ကျမ်းကိုး-) .
8 v rehearse (as in ဇာတ်-).
9 v give, offer or entertain(with liquid refreshments).

တိုက်
1 v cart or transport (things) in vehicles (as in လှည်းဖြင့်-).

တိုက်
1 n brick building.
2 n cell (as in ကြိုး-); chamber (as in မြေ-).
3 n building or place where some specialized activity is carried out, such as business concerns, depository, works, store, etc (as in ဆေး-ပိုးထည်-လက်နက်-စာကြည့်-ကုန်-).
4 n continent (as in အာရှ-).
5 n administrative unit; circle (as in သူကြီး-).
6 n group of buildings belonging to a monastery or nunnery.

FIG. 3.14 : Entrées pour ပြ et တိုက် du MED

Nombre de syllabes	Pourcentage
1	23
2	39.1
3	23
4	8.9
5	4.5
6	1.4
7	0.1
8	0

FIG. 3.15 : Syllabes du vocabulaire de *Burmese by Ear*, BBE

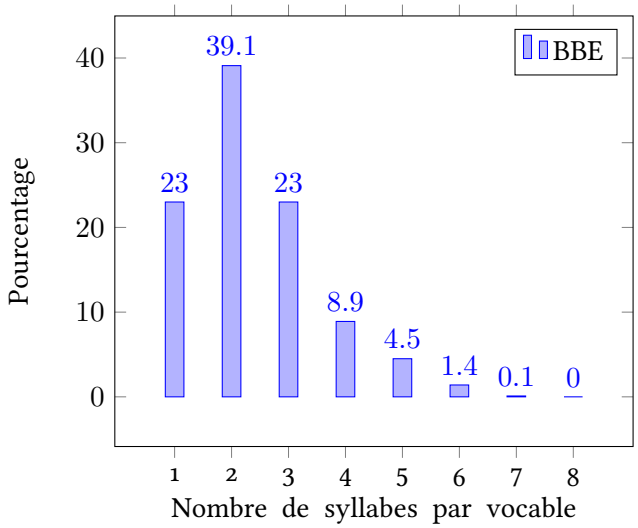


FIG. 3.16 : Longueur en syllabes du vocabulaire de *Burmese by Ear*, BBE

Les données de huit ressources différentes sont représentées par la figure 3.17. Comme pour la figure précédente, les axes horizontaux représentent le nombre de syllabes par vocable, et les axes verticaux le pourcentage du vocabulaire total. Ainsi, la première barre verticale de chaque diagramme représente le pourcentage de vocables à une syllabe, la deuxième barre verticale, la proportion de vocables à deux syllabes et ainsi de suite. Nous signalons en passant que dans le septième diagramme, qui représente la longueur du vocabulaire de *Myanmar Newspaper Reader* (dorénavant MNR), la dernière barre indique les unités de neuf syllabes et plus.

D'abord, nous constatons que les diagrammes de BBE, *Colloquial Burmese* (CB)

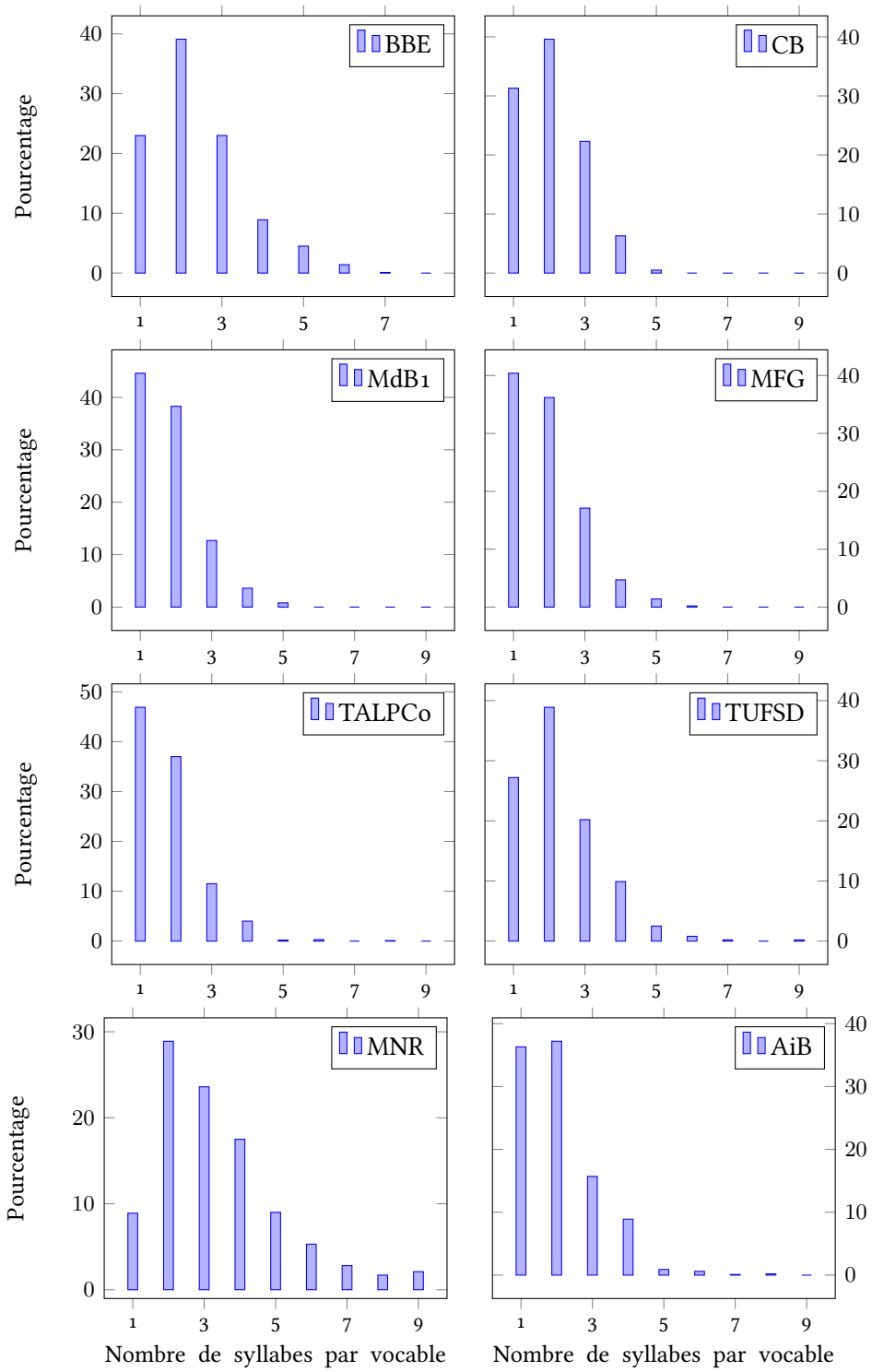


FIG. 3.17 : Comparaison de longueur en syllabes du vocabulaire de huit ressources différentes : *Burmese by Ear*, *Colloquial Burmese*, *Manuel de birman 1*, *Myanmar Flower Grammar*, *TALPCo-TUFS Vocabulary*, *TUFS Dialogues*, *Myanmar Newspaper Reader*, *Advancing in Burmese*

et *TUFS Dialogues* (TUFSD) se ressemblent, dans la mesure où la proportion de vocables à deux syllabes est la plus grande. Pour les deux dernières ressources, les vocables à une syllabe représentent environ un tiers du vocabulaire. Ces trois ressources sont destinées aux apprenants débutants, avec une approche plus communicative basée sur des dialogues qui peuvent servir dans la vie réelle.

Les trois ressources dont la proportion de vocabulaire à une syllabe est la plus grande sont *Manual de birman 1* MdB1, *Myanmar Flower Grammar* MFG et *TALPCo-TUFS Vocabulary* TALPCo. Il s'agit également de ressources pour les apprenants débutants, dont les phrases sont délibérément composées de vocables simples, car l'objectif principal des leçons semble être plus grammatical que communicatif. Nous avons remarqué que les ressources de ce dernier groupe utilisent moins de mots d'emprunts et de noms propres, qui eux sont souvent polysyllabiques.

Les deux ressources, MNR et *Advancing in Burmese* (AiB), sont destinées aux apprenants de niveaux intermédiaire ou avancé. Il est intéressant de noter que le vocabulaire à une et à deux syllabes de la ressource AiB est presque en proportions égales, 36,3% et 37,2% respectivement, ce qui conforte notre théorie que les manuels axés plus sur les explications grammaticales que les compétences communicatives utilisent davantage de vocabulaire simple et moins de noms propres ou de mots d'emprunt polysyllabiques. La ressource MNR par contre, constituée d'articles de presse sans simplification, contient beaucoup de noms propres polysyllabiques, birmans (par exemple နိုင်ငံတော်ငြိမ်ဝပ်ပိပြားမှုတည်ဆောက်ရေးအဖွဲ့ /nàɪɴɲàɴtɔ̀ɲèɪɴwaʔpḭpjám̰u̱tesʰaʊ̯ʔjéʔa̱pʰwɛ̰/ *Conseil d'État pour la restauration de la loi et de l'ordre*, treize syllabes) et étrangers (par exemple ဘရိုင်ရင်ဝီလီယံ /bə̱jàɪɴjɪ̀ɴwìlìjàɴ/ *Brian Williams*, sept syllabes). Ceci n'est pas la seule explication de ce profil de vocabulaire. L'auteur de MNR, Luzoe (1996), a glosé très méticuleusement les textes, et même les formes au pluriel sont parfois incluses dans les listes[18], ce qui augmente légèrement les longueurs constatées, mais pas plus d'un ou deux pour cent. Nous y avons constaté beaucoup plus de composés que dans les autres ressources, comme အကူအညီ /ʔə̱kùʔə̱ɲì/ (*assistance*, quatre syllabes), ဒေသဖွံ့ဖြိုးရေး /dèθə̱pʰwə̱ɴpʰjójé/ (*développement régional*, cinq syllabes) ou ယဉ်ကျေးမှုအမွေအနှစ် /jɪ̀ɴtɕém̰u̱ʔə̱mwèʔə̱ɲɪʔ/ (*patrimoine culturel*, sept syllabes). Nous supposons que

18. Les chiffres et les dates, qui sont également glosées, ne sont pas inclus dans notre liste.

ces différences relèvent du fait qu'il s'agit de textes semi-officiels au style écrit, plus soutenu, d'un genre spécifique (la presse) et que la taille importante du vocabulaire joue également un rôle dans les proportions respectives des longueurs des vocables.

On pourrait émettre l'hypothèse que la langue des apprenants auxquels la ressource est destinée joue un rôle dans le profil du vocabulaire présenté, qu'un vocable dans la langue des apprenants soit plus aisément représenté par plusieurs vocables en birman par exemple, mais nous n'avons pas constaté ce phénomène à la lecture de nos ressources. Les données sur la longueur en syllabes présentées ne semblent pas soutenir cette hypothèse. Par exemple, les ressources destinées aux apprenants japonais, TUFSD et TALPCo ne présentent pas de similitudes, alors que MdB1, MFG et TALPCo, qui sont destinées aux francophones, apprenants thaïs et japonais respectivement, ont des profils de vocabulaire assez similaires.

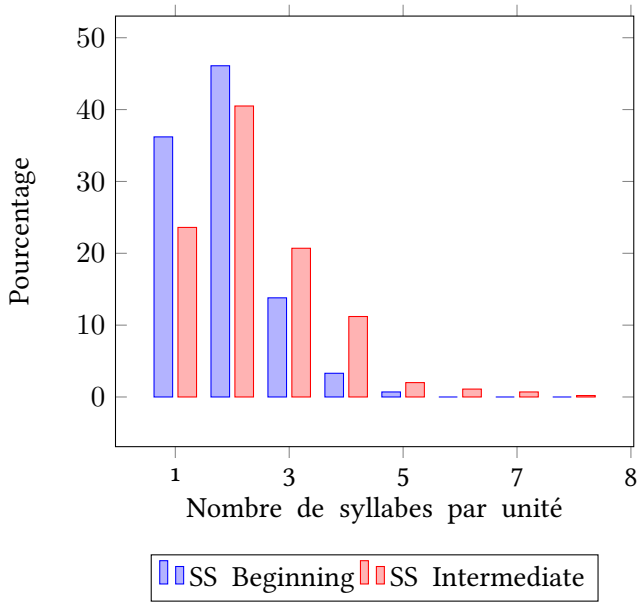


FIG. 3.18 : Comparaison de la longueur en syllabes du vocabulaire des deux niveaux du site *Seasite*, SS

Afin d'examiner la relation entre niveau de difficulté et longueur du vocabulaire, nous avons ensuite regardé les ressources qui comportent plusieurs niveaux de difficulté. Les lexiques du site didactique Seasite sont regroupés en deux niveaux, *Beginning* (débutant) et *Intermediate* (intermédiaire), ce dernier répète peu les

vocables du niveau débutant : seuls dix vocables sont communs aux lexiques des deux niveaux. Le niveau intermédiaire contient très majoritairement du vocabulaire qui est considéré comme plus difficile, et non pas une représentation de la totalité du vocabulaire utilisé dans les textes des leçons. La figure 3.18 démontre clairement une diminution de l'importance du vocabulaire à une syllabe entre le niveau débutant et le niveau intermédiaire avec une importance plus grande donnée aux unités polysyllabiques au niveau plus élevé. Cette tendance est encore plus marquante pour la ressource *Myanmar Language Test*, figure 3.19, qui est divisée en cinq niveaux de difficulté (MB est le niveau débutant). Rappelons qu'il s'agit de listes de vocabulaire recommandées pour des apprenants souhaitant passer un test de compétence en langue birmane, chaque liste étant composée de vocabulaire que les auteurs estiment qu'il faut connaître à un certain niveau de compétence.

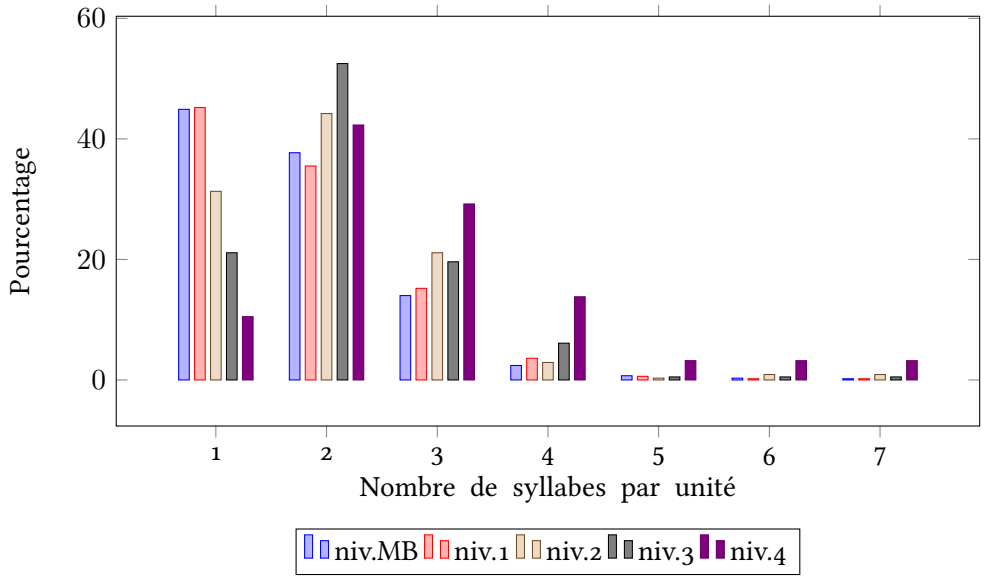


FIG. 3.19 : Longueur en syllabes du vocabulaire du *Myanmar Language Test*, MLT

Nous observons de nouveau que l'importance de vocables monosyllabiques diminue avec l'augmentation de la difficulté et à partir du niveau deux, le vocabulaire dissyllabique gagne en importance. Là encore, les listes ne sont pas cumulatives, les auteurs de la ressource s'attendent à ce que l'apprenant maîtrise tous les niveaux en dessous, le vocabulaire des listes de niveaux inférieurs

n'étant pas répété. S'agit-il vraiment d'une corrélation entre niveaux de difficulté et l'importance relative de la longueur de vocabulaire en syllabes, ou l'effet de la taille de l'échantillon ? Les données présentées en tableaux de la figure suivante 3.20 (les mêmes que le diagramme à barres précédent, figure 3.19) ont été utilisées pour calculer le coefficient de corrélation (Pearson *r*). L'importance relative (représentée en pourcentages) de chaque longueur de vocabulaire est comparée à la taille de vocabulaire à l'intérieur de chaque niveau. Nous constatons qu'il existe bien une corrélation entre tailles de vocabulaire et l'importance relative de la longueur du vocabulaire, sauf dans le cas de vocables à une et à deux syllabes. La corrélation négative entre vocabulaire d'une syllabe et difficulté montre que plus le niveau est difficile, plus la proportion de vocabulaire à une syllabe diminue. Ceci s'explique en partie par l'importance du vocabulaire des fonctions syntaxiques, qui est souvent à une syllabe, mais aussi par le fait que les formateurs de langue pourraient considérer que le vocabulaire court est plus simple. La corrélation faible entre niveau de difficulté et vocabulaire à deux syllabes illustre l'importance générale du vocabulaire à deux syllabes dans la langue.

Niveau	**Taille**
MLTMB	292
MLT1	330
MLT2	616
MLT3	1321
MLT4	3265

Syllabes	MLTMB	MLT1	MLT2	MLT3	MLT4	*r*
1	44.9	45.2	31.3	21.1	10.5	−0.91
2	37.7	35.5	44.2	52.5	42.3	0.31
3	14	15.2	21.1	19.6	29.2	0.93
4	2.4	3.6	2.9	6.1	13.8	0.99
5	0.7	0.6	0.3	0.5	3.2	0.92
6	0.3	0	0.2	0	0.9	0.97
7	0	0	0	0	0.2	0

FIG. 3.20 : Corrélation (coefficient de Pearson *r*) entre taille de vocabulaire et l'importance relative de longueur d'unités de vocabulaire, *Myanmar Language Test*, MLT

Nous comparons ce profil avec celui des 28 400 vedettes birmanes du dictionnaire birman-anglais de référence, le *Myanmar-English Dictionary* မြန်မာ-အင်္ဂလိပ် အဘိဓာန် ou MED (Myanmar Language Commission 1993), figure 3.21.

Il existe une corrélation étonnement parfaite entre le plus haut niveau du MLT, MLT4 et le profil du MED (le coefficient de corrélation *r* de Pearson est à 1), ce qui suggère que ces niveaux demandent une connaissance du vocabulaire qui est similaire au profil du vocabulaire du dictionnaire qui lui-même se veut

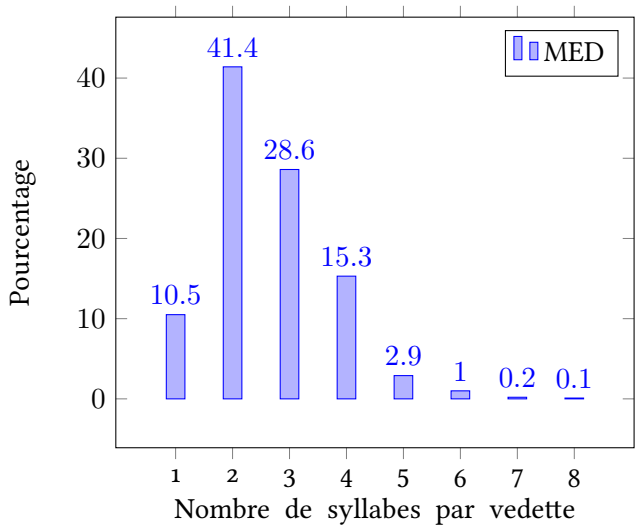


FIG. 3.21 : Longueur en syllabes des vedettes du *Myanmar-English Dictionary*, MED

une représentation plus générale de la langue. Nous voyons ici aussi l'importance du vocabulaire à deux syllabes (plus de quarante pour cent du dictionnaire). Ce constat est confirmé quand on considère le vocabulaire d'un autre corpus, le corpus *myPOS* (Khin War War Htike et al. 2017), qui est représenté dans la figure 3.22. Les résultats pour la segmentation (fournie par les auteurs du corpus) en *words* en bleu rappelle ceux du MED (figure 3.21) et le niveau MLT4 du MLT (figure 3.19).

Nous voyons donc que le vocabulaire de la langue n'est jamais présenté uniquement comme des monosyllabes par les manuels de birman, et ceci ne concerne pas seulement le vocabulaire d'emprunt et les noms propres. Nous devons donc tenter de découvrir ce qui définit un vocable en birman et comment l'identifier afin de définir les entrées du dictionnaire de *Motor*.

3.3.3 *L'épineuse question du « mot »*

Bien qu'il n'y ait pas eu d'études sur la segmentation en birman en tant que telle, nous pouvons avoir une idée de ce que les chercheurs considèrent comme des unités de segmentation appropriées en regardant comment ils abordent le concept de mot en birman et les catégories de parties du discours qui leur sont attribuées.

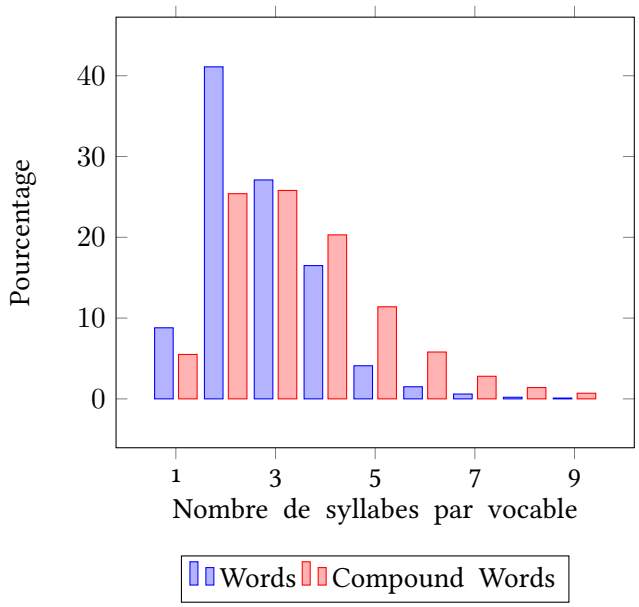


FIG. 3.22 : La longueur en syllabes du vocabulaire (les types) du corpus *myPOS*. En bleu la segmentation (fournie par les auteurs du corpus Khin War War Htike et al. (2017)) en *words*, en rouge la segmentation en *compound words*.

Certains mettent en doute l'universalité d'un concept de *mot* (Dixon et Aikhenvald 2002), parce que certaines langues n'auraient pas un terme qui veut dire *mot*[19], mais d'autres considèrent que le concept de *mot* est une réalité universelle (Goddard 2011 ; Wierzbicka 1996). Cela ne veut pas dire que le concept de mot peut être défini d'une manière universelle. Même si le locuteur d'une langue possède un sens inné de ce que c'est qu'un *mot* dans sa langue, ce terme est ambigu (Müller 1977), car non seulement le locuteur confond sa notion abstraite du mot et le mot concret sur la page, mais aussi *mot* peut se référer à des notions différentes selon son contexte d'utilisation. Le nombre de mots dans une phrase peut vouloir dire le nombre de mots uniques, les *types*, ou le nombre total de mots (le décompte de toutes les répétitions de tous les mots-formes différents), les tokens. On peut considérer que les formes au singulier et au pluriel sont un même mot, aussi bien que les formes fléchies d'un verbe. Dans ces cas, on modifie les tokens par un prétraitement pour refléter cette notion du mot (remplacer les formes au pluriel par le singulier, ou les formes fléchies

19. Dixon et Aikhenvald (2002) constatent que la grande majorité de locuteurs de petites langues non écrites ont un terme pour *nom propre*, mais n'ont pas un terme pour *mot*.

par l'infinitif par exemple).

D'une manière générale, le concept de mot peut être abordé sous de nombreux angles différents. Le plus évident pour le lecteur est en effet le mot graphique, appelé *mot-forme* ou *mot orthographique*, qui, dans un texte à écriture alphabétique est séparé de ses semblables par les espaces ou la ponctuation. Le traitement automatique s'appuie souvent sur le mot-forme pour découper le texte en tokens, supposant que ces tokens représenteront une unité linguistique significative dont l'analyse fournira des informations sur la langue étudiée. L'objectif n'est pas d'analyser le token, mais ce que le token représente, appelé *mot* dans la littérature scientifique concernant le traitement automatique du birman[20].

En linguistique théorique, la littérature sur la notion de mot est très abondante. Selon Haspelmath (2017), généralement les linguistes considèrent que le *mot morphosyntaxique* intervient à deux niveaux d'organisation grammaticale : la structure des mots est expliquée en termes de morphologie et la structure des phrases par la syntaxe. Face à des désaccords sur la segmentation parmi les locuteurs natifs, il énumère les propriétés utilisées par des linguistes pour définir les mots. Ces propriétés sont divisées en catégories : sémantiques, orthographiques, phonologiques et morphosyntaxiques. La première catégorie, sémantique, concerne la non-compositionnalité. Cette propriété est celle utilisée par la définition du terme စကားလုံး /sa̰kálóʊɴ/ (*mot* en birman) dans le *Myanmar Dictionary* မြန်မာအဘိဓာန် (Myanmar Language Commission 2008), qui nous semble un peu circulaire : **စကားလုံး** အနက်အဓိပ္ပာယ်ရှိသော စကားတစ်ခု /**sa̰kálóʊɴ** ʔa̰nɛʔʔa̰deɪʔpaʔʃḭθɔ́ sa̰kátɪʔkʰṵ/ *Un mot qui a une signification.* Haspelmath (2017) souligne que les critères sémantiques ne permettent pas de différencier entre les expressions idiomatiques et les mots, car les expressions idiomatiques doivent être apprises et mémorisées en entier tout comme les mots. Il explique que l'orthographe, n'est pas non plus une catégorie utile, non seulement parce qu'il existe des systèmes d'écriture sans délimitation explicite entre les mots, mais à cause de l'incohérence des règles d'orthographe dans beaucoup de langues. On peut ajouter les critiques de Grefenstette (2010) sur l'influence de l'orthographe via les dictionnaires traditionnels sur le traitement automatique, qui mettent l'accent sur les mots orthographiques individuels au lieu des concepts composés de mots multiples. Les deux catégories restantes, phonologiques et morphosyntaxiques (grammati-

20. Sans exception, l'usage du terme *word* concerne toutes les études en traitement automatique du birman citées.

cales) sont les plus citées par les linguistes pour délimiter les mots. Les critères d'identification de mot de ces deux catégories n'identifient pas systématiquement les mêmes frontières de mots, car la structure prosodique ne correspond pas toujours à la structure syntaxique (Hildebrandt 2015).

Le *mot prosodique* birman est décrit par Green (1995) et Green (2005). Il explique qu'en termes de prosodie, la plupart des mots prosodiques sont soit monosyllabiques (ex. ကျား /tɕá/ tigre), soit composés d'une syllabe mineure suivie d'une syllabe principale (ex. ထမင်း /tʰəmíɴ/ *riz*). Le mot peut être considéré dans ces cas comme étant constitué d'un mot prosodique et d'un pied. Green décrit deux types de mot composé, qu'il appelle les composés non réducteurs et les composés réducteurs. Les composés non réducteurs consistent simplement de deux (voir plus) mots prosodiques, qui ensemble peuvent s'analyser comme un seul mot prosodique (ex. ရောင်း /jáʊɴ/ *vendre* + ဝယ် /wɛ̀/ *acheter* > ရောင်းဝယ် /jáʊɴwɛ̀/ *commerce*). Dans un composé réducteur, l'avant-dernière syllabe devient une syllabe mineur qui perd son ton, et sa voyelle devient un schwa /ə/ (ex. ငါး /ŋá/ *poisson* + ပိ /pḭ/ *être pressé* > ငါးပိ /ŋəpḭ/ *pâte de poisson*). Green fait mention aussi de mots qu'il qualifie de *superlong*, qui ne sont pas des mots composés, la plupart selon lui, des mots d'emprunt, indécomposable en birman, à l'instar de ချောကလက် /tɕʰɔ́kə̰lɛʔ/ (*chocolat* < anglais) et အင်းဆက် /ʔíɴsʰɛʔ/ (*insect* < anglais). Puisque le traitement automatique repose sur la forme écrite du birman, il n'est pas possible de distinguer entre ces différents types de mot automatiquement.

Les derniers types de critères linguistiques d'identification de mot selon Haspelmath (2017) sont morphosyntaxiques. Il en identifie une dizaine. Les critères les plus cités (notamment par Trask (2004)) sont l'ininterruptibilité et la mobilité (voir aussi l'*indissociabilité* et l'*autonomie faible* de Kahane (2008)). Sans vouloir reprendre en détail l'intégralité de son article[21], Haspelmath (2017) démontre qu'aucun critère, soit considéré seul, soit en combinaison, ne peut être appliqué pour définir la notion de mot pour toutes les langues et ainsi fournir un moyen fiable de distinguer la morphologie de la syntaxe.

21. Haspelmath (2017) passe en revue les critères suivants : les pauses dans la chaîne parlée (*Potential pauses*), l'occurrence libre (*Free occurrence*), la mobilité externe et la fixité interne (*External mobility & internal fixedness*), l'ininterruptibilité (*Uninterruptibility*), la non-sélectivité (*Non-selectivity*), la non-coordinatabilité (*Non-coordinatability*), l'insularité anaphorique (*Anaphoric islandhood*), la non-extractibilité (*Nonextractability*), les idiosyncrasies morphophonologiques (*Morphophonological idiosyncrasies*) et les divergences par rapport à la biunicité (*Deviations from biuniqueness*).

Une alternative serait de proposer une notion de mot morphosyntaxique qui soit spécifique à la langue, ou bien, comme nous proposons, non seulement spécifique à une langue, mais aussi spécifique à une tâche. C'est ce qui est proposé par Aroonmanakun (2007) pour la segmentation du texte en thaï. Selon lui, un système de traduction automatique segmenterait en concepts (qu'il appelle *lexemes*) qui correspondent aux entrées de dictionnaires classiques, mais d'autres usages pourraient plus facilement segmenter le texte en tokens plus simples, qui correspondraient aux morphèmes et aux mots composés véritables dont le sens diffère sensiblement de la signification de ses composants, comme แม่น้ำ /mɛ̂ːnáːm/ (*rivière, fleuve*) < แม่น้ำ /mɛ̂ː/ (*mère*) + น้ำ /náːm/ (*eau*). Il s'agit donc de critères sémantiques. Aroonmanakun (2007) introduit la notion de *degré de liaison sémantique*. Pour lui, un système de segmentation devrait être minimaliste, séparant les composants de mots composés qui ne sont pas *étroitement liés* en mots multiples. Il soutient que la segmentation minimaliste évite la plupart de désaccords sur la segmentation, mais il faut garder à l'esprit que ce système présuppose qu'un traitement après la segmentation minimaliste réunira les expressions à mots multiples en mots composés, que le locuteur natif considérait habituellement non segmentables, comme หมอฟัน /mɔ̌ːfan/ (*dentiste*) préalablement segmenté en หมอ /mɔ̌ː/ (*spécialiste*) + ฟัน /fan/ (*dent*).

En linguistique du birman, de nombreux chercheurs qualifient le birman de langue principalement monosyllabique (Bernot 1980 ; Hnin Tun 2013 ; Myint Soe 1999) et décrivent le *mot* ou *mot simple* comme correspondant au morphème (Okell 1969 ; Vittrant 2019). Hopple (2003) parle de *mot simple*, qui « correspond approximativement à un seul morphème faisant preuve d'une autonomie lexicale, sémantique et grammaticale. »[22] Cette définition présuppose que le mot lexical (celui qui est mémorisé par le locuteur), le mot morphosyntaxique et le mot sémantique coïncident, ce qui n'est pas une évidence. Quand deux, voire plus de ces *mots* sont juxtaposés, ils sont appelés *mots composés* (Okell 1969), *mots polysyllabiques* (Hnin Tun 2013) ou *mots complexes* (Hopple 2003). Okell (1969, p. 2) distingue les *mots composés* (qui contiennent plus d'un mot simple, comme မြစ်ရေ /mjɪʔjè/ *l'eau de rivière* < မြစ် /mjɪʔ/ *rivière* + ရေ /jè/ *eau*) des *mots dérivés* qui sont créés à partir de *mots* et ce qu'il nomme *formatives*, des affixes dérivationnels (comme le préfixe အ /ʔə̰/, dans အပူ /ʔə̰pù/ *la chaleur* <

22. *The simple word corresponds roughly to a single morpheme that demonstrates lexical, semantic, and grammatical autonomy.*

ပူ /pù/ *être chaud*) ou la répétition (par exemple မြန်မြန် /mjànmjàɴ/ *vite* < မြန် /mjàɴ/ *être rapide*). Hopple distingue deux types de mots complexes : d'abord le *mot composé simple* qui est composé de deux noms simples (N + N) (ထောင်သား /tʰàʊɴθá/ *prisonnier* < ထောင် /tʰàʊɴ/*prison* + သား /θá/*fils*) ou deux verbes simples (V + V) (တိုးတက် /tóteʔ/ *faire des progrès* < တိုး /tó/ *avancer* + တက် /teʔ/ *monter*). Ensuite, elle désigne une deuxième type de mot complexe, le *composé étendu*, qui serait un mot composé dont l'une de ses parties est un nom complexe qui est structurellement un syntagme ou une proposition. Dans ces cas, il est difficile de déterminer s'il s'agit d'un mot composé étendu ou d'un syntagme.

3.3.3.1 Parties du discours pour le birman

Nous devons déterminer dans quelles situations il est approprié de segmenter ou non les mots composés. Les travaux sur les catégories morphosyntaxiques nous informent un peu davantage à ce propos.

En linguistique, les travaux en anglais ont d'abord été résumés par Minn Latt (1959), et cet état de l'art a été complété par la suite par Hopple (2003). Elle constate que le nombre de catégories retenu par les chercheurs est très variable, allant de dix-neuf (Minn Latt 1959) à trois, le *nom* et le *verbe* et la *particule* grammaticale (Myint Soe 1999). Hopple note que la position officielle du Myanmar Language Commission sur les parties du discours suit les catégories de parties du discours traditionnelles indo-aryennes, pâli, peut-être en raison du prestige du pâli dans la culture birmane et de l'influence de l'éducation européenne. Elle résume son état de l'art en disant que l'approche minimaliste gagne du terrain parmi les linguistes dont l'orientation est plus théorique, par opposition à ceux qui étaient par le passé plus intéressés par l'enseignement de la langue.

Les constats de Bernot (1971) sont plus nuancés. Selon elle, le birman est bien une langue « sans catégorie adjectivale », mais elle note que parfois des adjectifs d'emprunt peuvent fonctionner comme des adjectifs ou des substantifs. Dans Bernot (1983) et Bernot, Cardinaud et al. (2001), il est remarqué qu'il y a très peu d'adverbes simples comme ယခု /jə̰kʰṵ/ (*maintenant*) et သိပ် /θeiʔ/ (*très*), mais il y a une grande quantité « de composés et de périphrases de longueurs et de formes variables… [qui peuvent] correspondre à des catégories différentes en français : nom, adjectif et adverbe ». Depuis, la grammaire de Jenny et Hnin Tun (2017) a démontré la complexité à l'intérieur de ces parties

de discours (voir le tableau 3.2).

Nominals	Nouns	
	Pronouns	
	Measure words	
	Classifiers	
Verbs	Main verbs	
	Auxiliaries	
	Types of verbs	Intransitive verbs
		Transitive verbs
		Property verbs (adjectives)
Adverbs	Phrasal adverbs	
	Clausal adverbs	
Grammatical markers	Phrasal markers	
	Clausal markers	
Pragmatic particles	Phrasal particles	
	Clausal particles	

TAB. 3.2 : Les parties du discours décrites dans *Burmese : A Comprehensive Grammar* (Jenny et Hnin Tun 2017)

3.3.3.2 Étiquettes morphosyntaxiques en traitement automatique du birman

En traitement automatique de langues, actuellement aucun jeu d'étiquettes morphosyntaxiques ne fait unanimité. Le corpus *myPOS* (décrit dans la section 2.2.2) (Khin War War Htike et al. 2017) emploie un système de seize étiquettes différentes, inspiré des dix parties de discours qui apparaissent dans le MED, le *Myanmar-English Dictionary* မြန်မာ-အင်္ဂလိပ် အဘိဓာန် (Myanmar Language Commission 1993). Kyaw Htet Minn et Khin Mar Soe (2019) utilise quatorze étiquettes, là aussi ce jeu d'étiquettes est inspiré du travail de la Myanmar Language Commission (2005), mais cette fois de leur ouvrage de grammaire. Dim Lam Cing et Khin Mar Soe (2020) utilisent douze étiquettes, basées sur les mêmes sources. Le corpus ALT (décrit dans la section 2.2.1), qui est aussi basé sur les parties de discours du MED, n'a que douze étiquettes, alors que sa version NOVA (Ding, Hnin Thu Zar Aye et al. 2016) n'en emploie que sept. Ce dernier jeu d'étiquettes se rapproche le plus de l'approche minimaliste proposée par la linguistique birmane à trois catégories, puisque seules les quatre premières peuvent réellement être considérées comme des catégories syntaxiques : **n** (nom),

v (verbe), **a** (adjectif), **o** (autre modificateur), **1** (chiffre), . (ponctuation) et **+** (token avec rôle syntaxique indéfini). Selon cette approche, les tokens sont les morphèmes rassemblés en syntagmes. Une chercheuse adopte une approche complètement opposée, proposant des jeux d'étiquettes très étendus allant de 88 Myint (2011a) à 109 Myint (2011b) étiquettes morphosyntaxiques pour rendre compte de toutes les parties du discours en birman qui selon elle, permet une segmentation avec une précision moyenne de 84,71%.

Le recours aux catégories traditionnelles de la Myanmar Language Commission et l'utilisation de la catégorie adjectif en particulier semblent particulièrement problématique pour le traitement automatique. La catégorie « adjectif » est attribuée par les chercheurs à des tokens qui semblent avoir des comportements syntaxiques très divers : très souvent pour les verbes d'état comme ကြီး /tɕí/ (*être grand*), mais aussi pour des nombres ordinaux (par Ding, Aye et al. (2019)), l'équivalent des adjectifs démonstratifs français (par (Khin War War Htike et al. 2017)) et même pour des parties de mots composés qui seuls seraient considérés comme des substantifs (par Myint (2011b)). Ces choix ont des répercussions sur ce qui est considéré comme une segmentation « correcte ». Par exemple, Soe Lai Phyue (2020) définit les adjectifs comme un mot doté d'un suffixe ဝေါ /θɔ́/, သည့် /θi̱/ ou မည့် /mi̱/ qui a le sens « être X » quand X est un adjectif en anglais. Parfois ces suffixes sont traités comme partie intégrante du token « adjectival », pour les distinguer de verbes, qui ont quasi le même comportement syntaxique. L'utilisation d'une traduction d'une autre langue pour catégoriser les tokens semble particulièrement peu pratique. En termes de catégorie grammaticale, Jenny et Hnin Tun (2017) classent ces formes parmi les verbes, appelés *property verbs* ou *verbes d'état et de qualité* par Bernot, Cardinaud et al. (2001). Jenny et Hnin Tun (2017) remarquent qu'il y a peu de différence dans le comportement syntaxique entre ces verbes et les autres verbes.

En ce qui concerne les ressources d'apprentissage de la langue, seuls Niyomtham et al. (2017), Okano et al. (2016) et dans une certaine mesure Yadana Aung (2020) font usage de la catégorie adjectif. Niyomtham et al. (2017) groupe les adverbes et les verbes d'état dans une même catégorie comme en thaï. Les catégories du vocabulaire du Myanmar Language Test de Okano et al. (2016) semblent identiques à celles du MED. Yadana Aung (2020) semble en faire un usage plus restreint, mais parfois assez incohérent, par exemple ရှိုင်း /jáɪɴ/ (*être grossier*, *être*

sauvage) est un adjectif alors que ပျင်း /pjíɴ/ (*s'ennuyer*) est un verbe (adjectif dans le MED), et ရှေ့ /ʃḛ/ (*devant*, *avant*) est aussi considéré un adjectif, alors que son usage syntaxique dans le manuel n'est pas toujours le même.

Nous avons remarqué que même lorsque les chercheurs en traitement automatique s'accordent sur les noms des catégories morphosyntaxiques, ils ne les emploient pas de la même manière, car il n'existe pas de guide d'annotation uniforme. Par exemple, la syllabe စွာ ‹śwa› /swà/, définie dans le *Myanmar-English Dictionary* မြန်မာ–အင်္ဂလိပ် အဘိဓာန် comme une particule formant un adverbe, est parfois étiquetée *particle* (voir Khin War War Htike et al. (2017)) et fait parfois partie d'un token étiqueté *adverb* (voir Dim Lam Cing et Khin Mar Soe (2020)).

Une réévaluation des catégories morphosyntaxiques dépasse largement le cadre de notre projet de thèse, mais nous voulions voir par nous-mêmes quelles seraient les conséquences de l'utilisation de ces catégories en entraînant nous-même un outil d'étiquetage morphosyntaxique.

3.3.3.3 Étiquetage morphosyntaxique simple avec TreeTagger

TreeTagger est un outil qui associe des étiquettes morphosyntaxiques à des tokens selon un modèle probabiliste basé sur des arbres de décision, dont le fonctionnement est décrit dans (Schmid 1994)[23]. Un outil d'entraînement permet d'en créer un outil spécifique à une nouvelle langue à l'aide d'un corpus d'entraînement, un jeu d'étiquettes et un lexique. *TreeTagger* est rapide, simple à entraîner, et peut s'intégrer à d'autres logiciels facilement avec une prise en charge intégrale d'Unicode.

Nous avons démarré notre entraînement de *TreeTagger* pour le birman en nous servant du corpus étiqueté *myPOS* (Khin War War Htike et al. 2017), divisé en deux : 80% corpus d'entraînement et 20% en corpus test avec validation croisée, dont les étiquettes morphosyntaxiques sont basées sur le MED (voir tableau 3.3). C'est ce même dictionnaire que nous utilisons comme lexique pour l'entraînement de l'outil (avec toutefois l'ajout de tous les types en lettres latines du corpus *myPOS*, qui sont étiquetés *fw*). Le dictionnaire compte 26 504 entrées au total, dont 13 599 homographes. Dans ce dictionnaire, seules 1577 vedettes ont autant d'étiquettes que de définitions et sont ainsi potentiellement désambiguïsables par

23. Il existe des modèles d'étiquetage pour de nombreuses langues. Voir la page web de TreeTagger pour des modèles spécifiques https://www.cis.lmu.de/~schmid/tools/TreeTagger/

un outil d'étiquetage morphosyntaxique. Toutefois, ceci ne prend pas en compte la fréquence d'apparition de ces entrées dans les textes.

myPOS	**MED**	Exemples
abb : abbreviation	–	အထက ဂျီဒီပီ အယ်လ်အယ်လ်ဒီစီ
adj	adjective	လှပ ထို ဒုတိယ ကောင်း
ppm	post-positional marker	နှင့် သည် ကို က တွင် ၏ တယ်
adv	adverb	စောစော မြန်မြန် အလွန် တစ်ခါတစ်ရံ
conjunction	conj	၍ ပြီး သို့မဟုတ် ထိုအခါ
fw (foreign script)	–	search keywords SunMicrosystems
int	interjection	စိတ်မရှိပါနဲ့ ကျေးဇူးပြုပြီး
n	noun	ဆရာကြီး အိန္ဒိယ စူဇူကီကေအိဂျီ
num	–	၉၆၅၈.၉၆
tn : text number	–	တစ်ထောင်
part	particle	သော တို့ ကြီး ခဲ့ မှု ပေး
part_neg : negative particle	–	–
pron	pronoun	ကျွန်တော် သူ ၎င်း ဘယ်တော့
punc	punctuation	။ ၊ ( ) _
sb : symbol	–	:? & #
v	verb	ခေါ် ဖြစ် ပါဝင် ရှိ နားလည်

TAB. 3.3 : Système d'étiquettes morphosyntaxiques du corpus myPOS et parties de discours du dictionnaire MED. Exemples tirés du corpus myPOS. On note que l'étiquette part_neg apparaît dans l'article de Khin War War Htike et al. (2017), mais ne figure pas dans le corpus.

Pour notre essai, nous ne tenons pas compte des tokens composés du corpus. Ce système obtient une précision de 94,0% (pourcentage d'étiquettes correctes). Notre premier constat est que 24,8% des erreurs comprennent la catégorie *adj* (3 539 sur 14 253 erreurs au total, dont 1 125 faux positifs (étiquetés *adj* incorrectement) et 2 414 des tokens du corpus étiquetés *adj* dans le corpus *mpPOS* ne sont pas reconnues lors des tests comme des adjectifs. Comme nous avons remarqué dans les articles scientifiques qui décrivent les expériences d'étiquetage morphosyntaxique du birman, le flou autour de la définition d'une catégorie adjectivale est responsable d'un bon nombre d'incohérences et de confusions entre « adjectifs » et verbes, noms, et même adverbes.

En général, Ding, Aye et al. (2019) attribuent les taux d'erreur élevés de l'étiquetage morphosyntaxique à une mauvaise qualité des corpus d'entraînement, en raison de désaccords entre annotateurs, mais ne semblent jamais remettre en question les systèmes d'étiquetage en eux-mêmes. Dans son livre sur les méthodes informatiques d'annotation et d'analyse de corpus, Lu (2014) fait une remarque intéressante. Il dit qu'un jeu d'étiquettes morphosyntaxiques doit être motivé par des considérations linguistiques et pratiques. Trop d'étiquettes réduit la précision de l'étiquetage. Être trop spécifique, ou pas suffisamment spécifique doit être évité. Il nous semble que jusqu'alors, les considérations linguistiques n'ont pas eu une influence suffisante pour rendre l'étiquetage du birman pratique et utile.

3.3.4 *Le « mot » et l'apprentissage du vocabulaire*

Comment réconcilier les observations sur l'enseignement de la langue avec les descriptions du birman comme langue « monosyllabique » et l'apparente confusion autour des étiquettes morphosyntaxiques en traitement automatique du birman ?

Tout d'abord, quand un locuteur de birman segmente en tokens un *texte* en birman, à la main, il est davantage dominé par les tokens d'une syllabe que ne semble le montrer la liste de son *vocabulaire*. Jusqu'ici nous avons examiné la longueur du vocabulaire des textes, les types, c'est-à-dire chaque token compté qu'une fois, mais si nous voulons avoir une idée de l'importance relative contextuelle de la longueur du vocabulaire dans un texte en tenant compte de la fréquence, nous comptons toutes les occurrences des tokens en contexte dans un corpus. La différence entre les deux façons de se rendre compte de l'importance du contenu d'un même corpus se voit quand on compare les fréquences des *types* de longueurs différentes du corpus *myPOS*, représentées de la figure 3.22, avec la fréquence de *tokens* de longueurs différentes dans le corpus *myPOS* de la figure 3.23. On voit clairement que dans un texte, ce sont les tokens d'une syllabe qui dominent le texte (environ 60%). L'importance relative entre les tokens de longueurs différentes change peu entre les deux types de segmentation. On voit principalement que les tokens longs dans la segmentation *compound words* contiennent souvent du vocabulaire à deux syllabes, comme အင်တာနက်ဝန်ဆောင်မှု (*internet service*) composé de အင်တာနက် (*internet*, trois syllabes) ဝန်ဆောင် (*assumer*

une responsabilité, deux syllabes) et မှု (particule substantivant) ou bien အုပ်ချုပ်ရေး (*administration*) composé de အုပ်ချုပ် (*administrer*, deux syllabes) et ရေး (particule qui désigne un état, une action ou une compétence).

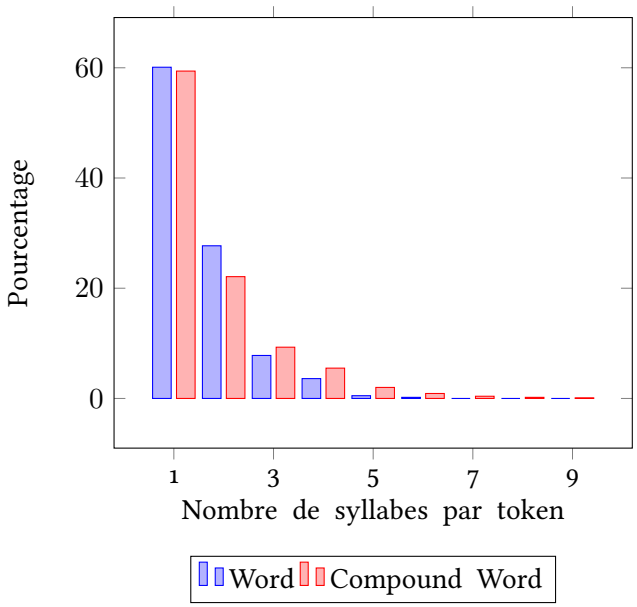


FIG. 3.23 : La longueur en syllabes des tokens du corpus *myPOS*. En bleu une segmentation par *word*, en rouge par *compound word*.

La deuxième hypothèse pour expliquer cette qualification de « monosyllabique » est que l'approche théorique minimaliste qui analyse les composants de ce que les enseignements de langue considèrent des mots composés comme des éléments de syntaxe permet une plus grande cohérence dans l'analyse syntaxique de la langue. Le birman est doté de procédés de dérivation lexicale productifs, comme la suffixation à des verbes par particules substantivantes telles que ချက် /tɕʰɛʔ/ et ခြင်း /tɕʰíɴ/, ou la suffixation de la particule စွာ /swà/ qui permet de créer un adverbe ou une expression adverbiale. Selon cette approche, ces particules sont des éléments d'un syntagme nominal ou adverbial au lieu d'être des suffixes de dérivation à l'intérieur d'un mot (voir l'exemple 78).

(78)

မျှော်	လင့်	ချက်
/mjɔ̰/	/l̰ɪɴ/	/tɕʰɛʔ/
espérer-V	attendre-V	NOMR-PTC
‘espoir’		

Une étude quantitative du degré de productivité de chacun de ces procédés reste à faire, mais nous avons constaté qu'un même procédé peut se trouver aussi bien dans un mot composé fréquent que dans un mot qui n'apparaît qu'une fois. L'approche minimaliste qui analyse en morphèmes permet de prendre en compte tous les usages d'un procédé de dérivation de la même manière. Le découpage en syllabes étant souvent proche du découpage en morphèmes et la syllabe étant ce qui est le plus proche du mot orthographique en birman, cela semble être une façon plus nette et plus birmane d'analyser la langue, surtout quand le sens des morphèmes ensemble est transparent, rassembler les éléments semble inutile. Il y a de bonnes raisons pratiques pour adopter une approche qui segmente le plus possible, car cela veut dire aussi que le dictionnaire de notre outil de segmentation a besoin de moins d'entrées. Prenons par exemple နားလည် /nálɛ̀/ (*comprendre*) qui est composé de deux syllabes *oreille* + *tourner*. Le substantif *compréhension* se forme simplement en ajoutant la particule မှု /m̥ṵ/, နားလည်မှု *nálɛ̀m̥ṵ*. Toutefois, pour la négation de *comprendre*, il faut placer la particule de négation မ /mə̰/ après *oreille* et avant *tourner*. Donc, *incompréhension* doit être formé de *oreille* + NEG + *tourner* + PTC : နားမလည်မှု /námə̰lɛ̀m̥ṵ/. Rajouter toutes les instances possibles de processus productifs de dérivation du lexique semble une tâche insurmontable[24]

Notre analyse des manuels de langue suggère cependant qu'une approche aussi cohérente de la syntaxe birmane ne reflète pas nécessairement la façon dont la langue est traitée et stockée dans le cerveau, qui est moins cohérente que ne le suggère la ligne de démarcation traditionnelle entre grammaire et vocabulaire utilisée pour enseigner les langues. Le lecteur birman est peut-être plus conscient des morphèmes, en raison du système d'écriture, mais des unités plus grandes que le morphème semblent être lexicalisées, car de telles unités sont répertoriées dans les dictionnaires et présentées dans les manuels d'apprentissage de la langue. Une approche de découpage maximal du texte cohérente d'un point de vue syntaxique semble ne pas prendre en compte la lexicalisation mentale de certains mots composés, surtout ceux qui sont lexicalisés, mais formés par des processus productifs de dérivation de mots. Par exemple,

24. Cette considération concerne spécifiquement l'outil que nous avons choisi d'utiliser, car il repose sur une liste finie. Bien sûr, certains procédés de dérivation même infiniment productifs pourraient être pris en compte de manière automatique d'une autre manière. Par exemple, le procédé très productif qui permet de créer les substantifs à partir de verbes en rajoutant le préfixe အ /ʔə̰/, အ + V → N comme အ + ဆို /sʰò/ (*déclarer*) → အဆို *déclaration*.

la particule substantivant မှု /m̥ṵ/ semble très productive. Il y a 687 types de mots composés dans le corpus *myPOS* qui terminent en မှု /m̥ṵ/. On pourrait considérer tous ces exemples comme des syntagmes à plusieurs tokens, ainsi : လူသေမှု /lùθèm̥ṵ/ (*personne* + *mort* + မှု → *homicide involontaire*) et de la même manière လူသတ်မှု /lùθaʔm̥ṵ/ (*personne* + *tuer* + မှု → *meurtre*). Toutefois, si certains (comme ces deux exemples) sont considérés dignes de figurer comme entrée dans le *Myanmar-English Dictionary* မြန်မာ–အင်္ဂလိပ် အဘိဓာန်, très peu ayant la même structure de dérivation sont inclus dans le *Myanmar-English Dictionary* မြန်မာ–အင်္ဂလိပ် အဘိဓာန်, même lorsqu'il s'agit du vocabulaire rencontré régulièrement dans nos corpus, par exemple ကူညီမှု /kùɲìm̥ṵ/ (*l'aide*) ou ငြိမ်းချမ်းမှု /ɲéɪntɕʰáɴm̥ṵ/ (*la paix*).

Il est intéressant de noter qu'une étude de Winskel et al. (2011) examinant l'effet de l'espacement sur la lecture en thaï, une langue avec un système d'écriture similaire, semble soutenir l'idée d'unité conceptuelle plus grande. Des expériences mesurant les mouvements oculaires de lecteurs comparant un texte non espacé et un texte artificiellement segmenté en mots ont montré que les lecteurs thaïlandais lisent par mots et non par morphèmes, car l'espacement artificiel du texte thaï en mots améliore le temps de lecture, mais n'a pas d'effet sur l'identification des mots. Dans une étude similaire mesurant les mouvements oculaires, Li et al. (2014) ont démontré aussi que les lecteurs de chinois lisent par mot et non pas par idéogramme (qui sont, comme la syllabe birmane, souvent aussi des morphèmes), d'une manière très analogue aux lecteurs de langues à écriture alphabétique espacée.

Les études en sciences cognitives utilisant l'imagerie neuronale semblent aussi étayer cette idée. En effet, les mécanismes qui sous-tendent la connaissance et l'acquisition du langage (dont certains sont également utilisés pour d'autres fonctions cérébrales), tels que les structures arborescentes imbriquées, les schémas algébriques et le chunking, ne sont pas clairement divisés en fonction de leur utilisation exclusive dans la formation des mots ou dans la syntaxe (Dehaene et al. 2015). Nous pouvons trouver des structures arborescentes aussi bien à l'intérieur d'un mot que dans la syntaxe par exemple. Le chunking, ou groupement en morceau de segments de langage fréquents, semble particulièrement important, comme un moyen efficace de comprimer statistiquement le langage pour faciliter la mémorisation et l'accès au vocabulaire. L'une des raisons de ce recours au

chunking pourrait être la capacité limitée d'un locuteur à traiter des syntagmes, peut-être limités à huit ou dix mots par syntagme (Pawley et Hodgetts Syder 1983). Wray (2014) remarque que l'apprentissage d'une langue semble se faire par unités plus larges qui sont ensuite décomposées et analysées si nécessaire. Cela n'exclut pas l'utilisation et la prise de conscience des relations morphologiques, qui, selon Nagy et al. (1993), se développent tôt dans l'apprentissage de la langue par un locuteur natif. Le simple fait d'être conscient des morphèmes n'est donc pas un bon argument pour l'inexistence d'une unité lexicale cognitive entre le morphème et la syntaxe.

Pour que le vocabulaire de nos listes de fréquence soit utile aux apprenants, les tokens qui résultent de la segmentation devraient se rapprocher le plus possible de cette unité cognitive, qu'est la *lexie*. Selon Tournier et Tournier (2017), le terme *lexie* regroupe la lexie affixée (« lexie construite composée d'une base, ou radical et d'un ou plusieurs affixes »), mais elle peut aussi être complexe, composée ou construite.

Même si nous acceptons le morphème comme une unité cognitive minimale, les unités constituées de groupes de morphèmes semblent également être des unités cognitives importantes, cruciales dans l'apprentissage des langues. En fait, selon Schmitt (2010), il est de plus en plus évident que les formules linguistiques sont tout aussi importantes dans l'apprentissage du vocabulaire que les mots individuels. L'importance de l'acquisition des phrases aussi bien que les éléments isolés de vocabulaire est aussi soulignée par Leech (2011), et il soutient que c'est la fréquence d'exposition du vocabulaire qui favorise son enracinement dans le cerveau.

Haspelmath (2017) suggère que le concept de *mot* serait un concept flou, que les frontières entre affixes, clitiques et mots indépendants seraient indéterminées sur un continuum et ainsi la distinction la plus importante ne serait pas entre morphologie et syntaxe, mais ailleurs, toujours autant indéterminée. Cette notion d'espace gris entre mot et syntaxe nous a poussé à définir nos tokens d'une manière plus souple, prenant en considération les statistiques concernant la fréquence d'occurrence de syllabes ensemble comme un moyen de déterminer ce qui constitue le vocabulaire utile pour l'apprentissage. Nous avons pris en compte d'abord d'autres critères qui indiqueraient les lexies en birman, à commencer par la typographie et les traditions de romanisation. Nous avons également accordé

une attention particulière aux entités nommées, dont la segmentation nécessite la prise en compte de certaines particularités linguistiques et orthographiques.

3.3.5 *Indications de frontières existantes*

Bien qu'il n'existe pas de délimitation explicite des unités lexicales mono ou polysyllabiques en birman, certains signes typographiques peuvent indiquer le début ou la fin d'une unité significative. Certains caractères de ponctuation de l'alphabet latin (comme % ou $) sont utilisés de la même manière (ou presque) qu'en français, d'autres (comme l'espace typographique et évidemment les caractères de ponctuation birmans) sont utilisés de manière spécifique au birman. Ces usages nous donnent parfois une indication de frontières conceptuelles des éléments de vocabulaire.

3.3.5.1 Le rôle de l'espace typographique en birman

Une méconnaissance sur le rôle de l'espace typographique en birman peuvent conduire à des résultats surprenants, tels que ceux rapportés par Christodouloupoulos et Steedman (2014) sur son corpus massif de traductions de la Bible. Selon eux, le birman (et le thaï d'ailleurs) présenterait une richesse lexicale extraordinaire par rapport à d'autres langues, avec un rapport type-token standardisé[25] (STTR) au-dessus de 90%, alors que, selon eux, même les langues agglutinantes telles que le turc et le coréen se situent vers 65%[26]. Il est clair que dans ce cas les tokens définis uniquement par les séparateurs comme l'espace et la ponctuation contiennent plusieurs unités lexicales. Mais quel est le rôle de l'espace ? Selon Ye Kyaw Thu, Finch, Sagisaka et al. (2013), les espaces typographiques ne sont pas nécessaires, ne sont pas employées dans des phrases courtes et sont utilisées simplement pour faciliter la lecture. Il n'existerait pas de règles, libre à l'auteur d'insérer ou non des espaces entre mots, entre syntagmes, voire entre racine et affixes ou postpositions (Hla Hla Htay et Narayana Murthy 2008). Dans la même veine, Hopple (2007) rapporte que des professeurs de langue birmane, des spécialistes de la littérature et des personnes ordinaires qualifient l'espacement comme « sans importance », que l'on est « libre de mettre de l'espace là où

25. C'est le pourcentage moyen de types par token pour chaque millier de tokens dans un texte.

26. Les auteurs admettent que leur méthode de segmentation utilisant les espaces signifie que les résultats sont trompeurs pour certaines langues.

cela semble le mieux du point de vue esthétique ». Toutefois, il semblerait que la position de l'espace, bien qu'en théorie facultative, ne soit pas totalement libre. Hnin Tun (2013, p. 85) fait remarquer que l'espace marque généralement la fin d'une phrase ou d'un syntagme. D'après Hopple (2007), l'espace délimite la fin d'une unité conceptuelle plus grande qu'un mot. Cette unité conceptuelle, qu'elle qualifie de *segment*, se terminerait habituellement par une particule post-positionnelle ou un substantif suivis de l'espace typographique, ou bien l'un des deux signes de ponctuation spécifique à l'écriture birmane (voir plus loin). Malgré la petite taille de son corpus (deux textes composés de 75 et 165 mots chacun), les conclusions faites par Hopple (2007), qui trouvent une corrélation positive entre les particules grammaticales et l'espace, sont confirmées par l'examen de nos corpus plus conséquents. Selon Hopple, le segment ainsi délimité organise les unités pour créer une unité psychologiquement « réelle » dans la grammaire birmane, et non pas juste esthétique. Le bon usage de l'espace serait donc un usage logique pour le locuteur, l'aidant à organiser le contenu du texte (en production et en réception).

Nous avons examiné des cooccurrents contigus de l'espace typographique à l'aide de TXM (Heiden et al. 2010) dans l'ensemble de nos corpus authentiques segmentés par *Motor* (plus de trois millions de syllabes). Le caractère espace (U+0020 SPACE) est remplacé par le tiret bas _ (U+005F LOW LINE[27] avant la segmentation, car ce traitement utilise SPACE pour séparer les tokens. De cette façon, nous pouvons facilement analyser l'espace comme n'importe quel autre token.

La commande Cooccurrences de TXM permet de spécifier la distance voulue entre le motif (pour nous il s'agit de _) et son cooccurrent, et en même temps de filtrer les résultats. Nous nous intéressons uniquement aux cooccurrents situés immédiatement à gauche de l'espace (c'est-à-dire les cooccurrences avec une distance moyenne de .0 entre le cooccurrent et le motif). La commande génère un tableau qui liste les cooccurrents, triés par défaut par l'indice de cooccurrence (appelé *score*), leur fréquence individuelle, et leur co-fréquence avec le motif. L'indice de cooccurrence est un indicateur de probabilité de rencontre (Lafon 1980) ; plus il est élevé, plus il est probable que le cooccurrent apparaisse avec le motif. Puisque nous nous intéressons à d'éventuelles règles d'usage de

27. Autrefois appelé SPACING UNDERSCORE.

l'espace en conjonction avec les tokens les plus fréquents, nous réordonnons le tableau par ordre de fréquence des cooccurrents (figure 3.24). D'abord, on constate que les scores sont extrêmement élevés pour les cooccurrents les plus fréquents, mais qu'il y a un écart considérable entre fréquence du cooccurrent et fréquence d'apparition avec le motif (co-fréquence). Nous pouvons en déduire qu'il y a une nette préférence d'écrire une espace après ces tokens et que l'espace est beaucoup plus fréquente après certains tokens que d'autres. On ne peut pas parler donc de règle absolue, mais de tendances fortes.

Cooccurrent	Fréquence	Co-fréquence	Score	Distance moyenne
။	111 236	38 952	127	.0
ကို့	74 297	52 156	127	.0
တွေ့	58 089	15 039	180	.0
၁	54 644	39 997	127	.0
တဲ့	44 393	27 835	127	.0
၊	40 199	22 382	127	.0
မှာ	38 022	22 694	127	.0
နဲ့	26 229	17 147	127	.0
ဟာ	23 298	16 989	127	.0
လို့	20 758	12 686	127	.0
ကြီး	20 178	11 793	127	.0
ရဲ့	17 678	12 783	127	.0
တော့	17 538	9 281	127	.0
ဖို့	12 936	7 089	127	.0
၎	12 750	4 643	127	.0
လည်း	12 482	10 591	127	.0

FIG. 3.24 : Début du tableau de cooccurrents immédiats situés à gauche de l'espace typographique, trié par ordre de fréquence

Si nous voulions formuler des règles ou recommandations de l'utilisation de l'espace à partir de l'usage, il est plus intéressant de regarder dans le sens inverse. C'est-à-dire, quoi d'autre à part l'espace peut suivre un token donné, et dans quelles circonstances le caractère de l'espace doit-il être inséré ? Nous prenons comme exemple le token ဘဲ /bɛ́/ (*sans*, *canard* selon Myanmar Language Commission (1993)) qui a une fréquence de 2170, une co-fréquence avec l'espace de 1275 et un score de cooccurrence avec l'espace de 320. Quand on génère

le tableau de cooccurrents de ဘဲ /bɛ́/ (figure 3.25), on voit la régularité de l'emploi de l'espace directement après ce token. On peut encore mieux cerner son usage en regardant des exemples d'usage de chaque élément en contexte dans le concordancier. Par exemple, l'usage de ဘဲ /bɛ́/ suivi de l'espace le plus fréquent est dans la structure **_ မ + V + ဘဲ _** pour exprimer une opposition telle que *ce n'est pas parce que* _ မ ဟုတ် ဘဲ _ /_ mą̀ hoʊʔ bɛ́ _ /.

Quand il est suivi de နဲ့ /nɛ̰/ (ici, *avec*) ou ကိုး /kó/ (particule emphatique), ces deux sont eux-mêmes suivis de l'espace. Un autre constat est que le sens de *canard* est surtout activé quand ဘဲ /bɛ́/ est suivi de ဥ /ʔṵ/ *œuf*. Compte tenu du score de cooccurrence de ces deux tokens, on peut considérer qu'il s'agit en l'occurrence d'un mot composé, ဘဲဥ /bɛ́ʔṵ/ (*œuf de canard*).

Cooccurrent	Fréquence	Co-fréquence	Score	Distance moyenne
_	679 812	1 275	319	.0
နဲ့	26 229	121	59	.0
ဥ	204	27	52	.0
၊	40 199	94	23	.0
။	111 236	171	22	.0
ကိုး	552	16	20	.0

FIG. 3.25 : Début du tableau de cooccurrents immédiats situés à droite de ဘဲ, trié par ordre de score de cooccurrence

Il n'est pas étonnant de trouver un bon nombre d'exemples de fin de segment suivi de l'espace parmi la catégorie de conjonctions : ဒါမှမဟုတ်_ /dàm̥ą̀mą̀hoʊʔ/ (score 313) (*ou bien*), ရန်_ /jàɴ/ (score 255) (*afin de*), ကတည်းက_ /kə̰tíɛ́ɛ́kə̰/ (score 235) (*depuis*), သော်လည်း_ /θɔ̀líɛ́ɛ́/ (score 207) (*bien que*) entre autres.

En plus de la corrélation positive entre les particules grammaticales et l'espace, nous avons trouvé d'autres cooccurrents gauches signifiants de l'espace typographique dans nos corpus. La liste est trop longue pour tout détailler, mais certaines tendances qui se démarquent pourraient constituer la base de recommandations de bon usage.

Hopple mentionne aussi que l'un segment peut être constitué uniquement d'un substantif. Il n'est pas donc étonnant de constater un score signifiant de cooccurrence entre la particule de pluriel တွေ /twè/ directement suivie de l'espace

(score 180).

L'espace est très souvent utilisée de part et d'autre de titres de respect, surtout ceux utilisés pour les noms personnels étrangers, tels que celui calqué sur l'anglais *Mister* _မစ္စတာ_ ⟨mɪ̰̀sta⟩ /mɪʔsʊ̰tà/ (*Monsieur*) (qui a un score de co-fréquence avec l'espace de 167) ou _သမ္မတ_ /θàɴmə̰tə̰/ (*Président*) (score de co-fréquence avec l'espace 267). Les titres de respect birmans, eux, sont par contre souvent accolés aux noms personnels. Autrement dit, on trouve systématiquement l'espace avant l'équivalent birman de *Madame* _ဒေါ် /dɔ̀/, mais presque jamais après. La délimitation des deux côtés de prénoms étrangers par des espaces est aussi frappante (comme _အိုဘားမား_ ⟨ʔiuba²ma²⟩ *Obama*, par exemple), ou parfois le nom entier (prénom + nom) forme un segment à l'instar des noms birmans, comme _ကိုဖီအာနန်_ ⟨kiupʰïʔannˣ⟩ *Kofi Annan*. bien que moins évidente d'un point de vue statistique en vertu de la variété de noms propres présente dans les corpus.

Les chiffres, y compris les années ou l'heure, sont souvent entourés d'espaces pour plus de clarté. Ci-dessous, nous avons le jour de la semaine စနေနေ့ /sə̰nènḛ/ *samedi*, le mot pour *matin* မနက် /mə̰nɛʔ/, les chiffres de l'heure (10 h 25), les mots *heure* နာရီ /nàjì/ et *minute* မိနစ် /mḭnɪʔ/ séparés par des espaces : _ စနေနေ့ _ မနက် _ ၁ ဝ _ နာရီ _ ၂ ၅ _ မိနစ် အချိန် မှာ.

En ce qui concerne les pronoms personnels, dans les textes littéraires, il y a une tendance forte de mettre une espace après ကျွန်မ /tɕʊ̀ɴmə̰/ (*je (f)*, score 138) et ကျွန်မတို့ /tɕʊ̀ɴmə̰tṵ/ (*nous (f)*, score 129), surtout en position sujet quand le pronom n'est pas ouvertement marqué en tant que tel par က /ka/, que nous ne trouvons pas pour d'autres pronoms et rarement dans les autres types de textes. Bien qu'intéressant, il semblerait que cela soit simplement dû au fait que les écrivaines qui écrivent à la première personne soient surreprésentées dans notre corpus littéraire.

Nous avons remarqué une autre classe syntaxique mise en exergue par la délimitation d'espaces, la catégorie adverbiale, qu'on les qualifie d'adverbes tout courts ou d'expressions adverbiales. La particule တော် /swà/ (*particule formant un adverbe* selon le Myanmar Language Commission (1993)), qui possède une fréquence de 667 dans nos corpus authentiques et en tant que cooccurrent gauche de l'espace une co-fréquence de 477 et un score de 171, termine souvent un segment adverbial qui commence par une espace. Dans notre concordance de တော်

/swà/ nous avons trouvé de nombreux exemples différents, tels que _လူမဆန်စွာ_ /lùmə̀sʰànswà/ *de façon inhumaine*, _ပျူငှာစွာ_ /pjùŋàswà/ *gentiment*, _ပွင့်လင်းစွာ_ /pwɪ̃línswà/ *ouvertement*, _ငြိမ်းချမ်းစွာ_ /ŋéɪntɕʰánswà/ *paisiblement* (action), _အေးချမ်းစွာ_ /ʔétɕʰánswà/ *paisiblement* (état d'esprit), _လျင်မြင်စွာ_ /ljìnmjìnswà/ *rapidement* et _နီးကပ်စွာ_ /níkaʔswà/ *proche, de façon rapprochée*, pour n'en citer que quelques-uns.

L'examen du contexte gauche de l'espace est moins révélateur, car la marque de fin de phrase est souvent suivie d'une espace. Les éléments qui se situent souvent en début de phrase apparaissent donc dans la liste de cooccurrents droits de l'espace avec un score de cooccurrence élevé. A par les noms propres, titres, et chiffres, nous avons remarqué que la particule de négation မ /mə̀/ figure en bonne place parmi les cooccurrents droits avec un score de 127. Cela peut s'expliquer par le fait que la négation intervient souvent au début d'un segment. C'est en examinant le contexte gauche que nous avons vu aussi que les chiffres en toutes lettres sont précédés de l'espace, comme တစ် /tɪʔ/ (*un*).

Même si, la plupart du temps, l'espace typographique en birman est utilisée pour mettre en évidence un segment conceptuel plus long qu'une lexie, les deux extrémités du segment donnent une certaine idée sur les limites gauche ou droite de certaines lexies dans l'esprit de l'auteur du texte.

3.3.5.2 Autres séparateurs typographiques

Les deux signes de ponctuation du bloc Myanmar de l'Unicode sont ။ MYANMAR SIGN SECTION (U+104B) et ၊ MYANMAR SIGN LITTLE SECTION (U+104A).

Le premier, ။, translittéré ⟨.⟩ et appelé ပုဒ်ကြီး /poʊʔtɕí/, ပုဒ်မ /poʊʔmə̀/, ou နှစ်ချောင်းပုဒ် /n̥ɪʔtɕʰáʊnpoʊʔ/, fait office de marque de fin de phrase, mais s'utilise aussi pour séparer les numéros des éléments d'une liste numérotée et comme les guillemets itératifs[28] pour indiquer une répétition dans une liste verticale ou un tableau (Myanmar Language Commission 2005). Un signe de ponctuation constitué de deux de ces signes, habituellement séparés par une tabulation comme ceci ။ ။, appelé ပုဒ်မကြီး /poʊʔmə̀tɕí/, est employé comme un introducteur ou présentatif de ce qui suit : une note, une référence ou après le sujet d'un paragraphe[29]. L'équivalent en français serait le deux-points.

28. On utilise aussi le mot *idem* en français.
29. Comité des manuels sur le contenu des programmes d'études 2016-2017 (2015)

Le deuxième signe, ၊ (translittéré ‹,›) que l'on compare souvent à la virgule, s'appelle diversement ပုဒ်ကလေး /poʊ̯ʔkə̰lé/, ပုဒ်ဖြတ် /poʊ̯ʔpʰjaʔ/, ပုဒ်ထီး /poʊ̯ʔtʰí/, or တစ်ချောင်းပုဒ် /tɪʔtɕʰáʊ̯npoʊ̯ʔ/ en birman. Selon le *Myanmar-English Dictionary* မြန်မာ–အင်္ဂလိပ် အဘိဓာန် (Myanmar Language Commission 1993), ce signe de ponctuation constitué d'un trait vertical indique une courte pause. Selon la grammaire de la Myanmar Language Commission (2005), il est utilisé pour séparer plus de deux exemples dans une liste, pour séparer les éléments longs, pour séparer les composants de la date, de l'heure, d'une adresse et entre le nom d'une personne et sa profession. Nous avons remarqué en effet qu'il est souvent employé de la même façon que la virgule pour séparer les éléments d'une liste, comme on peut voir dans l'exemple 79 qui utilise plusieurs mots anglais écrits en lettres latines, mais il peut aussi être comparé au trait d'union « suspendu » en anglais, dans la mesure où il peut indiquer une séparation entre éléments qui sont normalement très proches, mais séparés dans la phrase[30]. Dans cet exemple, ဘာသာစကား /bàθàsə̰ká/ (*langue*), normalement accolée au nom de langue, n'est pas répété pour chaque élément. C'est l'usage du signe petite section qui indique que tous les éléments de la liste sont des langues.

(79)

Dutch	၊	**French**	၊	**English**	၊	**Irish**	၊	**Danish**	၊	**Turkish**	၊
/dʌʧ/	,	/frɛnʧ/	,	/ˈɪŋglɪʃ/	,	/ˈaɪərɪʃ/	,	/ˈdeɪnɪʃ/	,	/ˈtɜːkɪʃ/	,
néerlandais	,	français	,	anglais	,	irlandais	,	danois	,	turc	,

Macedonian	**နှင့်**	**Cubano**	**စတဲ့**	_	**ဘာသာစကား**	**တွေ**
/ˌmæsɪˈdəʊ̯njən/	/n̰ɪ̃N/	/kjuːˈbɑːnəʊ̯/	/sə̰tɛ̰/	_	/bàθàsə̰ká/	/twè/
macédonien	et	cubain	etc	_	langue	PL

‘les langues néerlandaise, française, anglaise, irlandaise, danoise, turque, macédonienne et cubaine et ainsi de suite’

Ce phénomène est illustré plus simplement par l'exemple 80. En birman le nom d'un pays est toujours suivi de နိုင်ငံ /nàɪɴŋàɴ/ (*nation*) ou ပြည် /pjì/ (*pays*). Là où en Français on désigne le pays simplement *la France*, en birman on utilise systématiquement *France-nation*, ပြင်သစ်နိုင်ငံ /pjìɴθɪʔnàɪɴŋàɴ/. Dans cet exemple, la petite section indique que la fin du mot composé commune à tous les éléments de la liste, နိုင်ငံ /nàɪɴŋàɴ/ (*nation*), ne sera pas répétée, mais partagée avec le dernier élément de la liste. Celui-ci est accompagné d'une indication du pluriel, တွေ /twè/, pour indiquer qu'il n'appartient pas uniquement au dernier élément

30. Par exemple *full- and part-time employees*

Niger.

(80) နိုင်ဂျီးရီးယား ၊ _ ချဒ် ၊ _ ကင်မရွန်း _ နဲ့ _ နိုဂျဲ _ နိုင်ငံ _
/nàɪɴdzɨ́jíjá/ , _ /tɕʰaʔ/ , _ /kɪ̀ɴmə̰jʊ̀ɴ/ _ /nɛ̰/ _ /nìdzɨ́ɛ́/ _ /nàɪɴŋàɴ/ _
Nigéria , _ Chad , _ Cameroun _ et _ Niger _ nation _
တွေ
/twè/
PL
'Le Nigeria, le Tchad, le Cameroun et le Niger'

C'est un peu paradoxal, mais il s'agit à la fois d'une indication de séparation, et en même temps d'union avec un élément qui suit. Nous avons bien des mots composés dont les éléments ont l'habitude d'apparaître ensemble, mais qui sont dans certaines circonstances séparables comme s'il s'agissait de collocations. Comme le dit Aroonmanakun (2007) à propos du thaï, l'attraction mutuelle entre les éléments du composé est plutôt une question de degré et de contexte qu'un choix binaire entre mot simple et mot composé.

3.3.5.3 La segmentation du birman romanisé

L'adaptation d'une langue dont le système d'écriture native n'utilise pas ou peu d'espacement à un autre système requérant la segmentation en unités plus petites peut aussi donner quelques pistes pour la segmentation. Bien qu'il n'existe pas de systèmes officiels (UNGEGN 2013), plusieurs systèmes de transcription et de translittération du birman coexistent, utilisés dans des domaines différents. La transcription précise de la prononciation (parfois en alphabet phonétique internationale) est réservée pour les études linguistiques ou l'apprentissage de la langue. Les systèmes de translittération précise sont plutôt destinés aux articles scientifiques. Ces systèmes ne mentionnent pas tous la séparation d'éléments de romanisation. Par exemple, le système MLCTS du MLC (*Myanmar Language Commission Transcription System*), utilisé dans les publications de la MLC et assez répandu sur le Wikipédia birman, ne semble pas préciser la façon dont les éléments doivent être séparés. Les manuels de langue utilisent souvent la transcription pour la prononciation d'éléments de vocabulaire isolés. Ceux de Okell, U Saw Tun et al. (2010c) (et d'autres manuels du même auteur) divise les phrases suivant les espaces typographiques dans l'écriture birmane, et sépare les syllabes birmanes avec des tirets, laissant le vocabulaire d'origine étrangère

intact.

Un document officiel, une notification du gouvernement de l'Union de Myanmar à une conférence des Nations Unies sur la normalisation des toponymes en 2002, consiste en une simple liste de noms de lieu (les principaux états, villes et fleuves) avec leurs équivalents en anglais, date de 1989 (Government of the Union of Myanmar 2002). L'orthographe en anglais est conçue pour reproduire la prononciation du birman. Puisque les équivalents anglais sont tous en majuscules et qu'il n'y a pas de mention de séparation d'unités, on ne peut en tirer des informations explicites sur la segmentation, mais on note toutefois que l'on n'inclut pas la nature du toponyme dans la transcription. Par exemple, သံလွင် /θàɴlwɪ̃/, le nom du fleuve *Salouen*, s'écrit simplement THANLWIN, alors qu'en birman on suffixe souvent *fleuve* au nom : သံလွင်မြစ် /θàɴlwɪ̃mjɪʔ/ (*Salouen + fleuve*). On aurait pu s'attendre à « THANLWINMYIT » ou « THANLWIN MYIT ». Ceci nous amène à considérer que la première partie comme un nom propre, et donc séparable du reste du texte. L'accord sur les toponymes entre la Grande-Bretagne et les Etats-Unis (BGN/PCGN 1970) basé sur des tableaux de translittération datant de 1907 (Office of the Superintendent 1908) ne donne pas non plus d'indications d'espacement, mais indique simplement sans précision que les majuscules et minuscules doivent être employées le cas échéant.

La monographie de Okell (1971) passe en revue tous les systèmes de romanisation du birman en vigueur dans la littérature scientifique avant de recommander des modifications. Selon lui, la plupart des systèmes soit séparent les syllabes par des espaces typographiques, système qui certes enlève toute ambiguïté de frontières de syllabes, mais rend la lecture plus difficile, soit séparent les groupes de sens. Ce dernier est privilégié par le système de translittération de Duroiselle (1913) qui sépare les « mots » ou les groupes de sens, plutôt que les syllabes. En transcription, Yêkháun (1966) utilise l'espace typographique pour indiquer le manque de voisement de la consonne initiale d'une syllabe, système qui, selon Okell, n'est guère commode pour les emprunts polysyllabiques sans voisement interne, par exemple လောက 〈leak〉 /lɔ́kə̰/ (*le monde*) serait scindé en deux syllabes. Yêkháun (1966) précise tout de même que les mots étrangers restent entiers dans sa transcription. L'espacement est aussi requis entre expressions, entre un mot et ses suffixes grammaticaux et « à la discrétion » de l'auteur. Okell (1971) souligne que ce système souffre de la difficulté à délimiter les mots en birman.

Le système de translittération ALA-LC (American Library Association, Library of Congress 2011), conçu pour les catalogues de bibliothèques, stipule que pour les mots birmans, les séparations se font entre les syllabes, mais pour les mots d'emprunt, même birmanisés, la séparation suit celle de la langue d'origine. Par exemple, l'emprunt à l'anglais ကော်မီတီ /kɔ̀mìtì/ (*comité, commission*) en romanisation ALA-LC est ⟨koʻmītī⟩ sans coupure. Par contre, un nom personnel birman polysyllabique est écrit en syllabes séparées, chacune prenant une majuscule : စောစိုင်မောင် /sɔ́sàɪnmàʊɴ/ en ALA-LC s'écrit ⟨Co Cuiṅʻ Moṅʻ⟩ à l'exception des éléments d'emprunt, မြသီတာ /mjə̰θìtà/ devient ⟨Mra Sītā⟩ et non pas ⟨Mra Sī Tā⟩. Deux systèmes (un standard et l'autre simplifié) similaires à l'ALA-LC, mais plus précis, par Lammerts et Griffiths (2016) utilise le même système d'espacement, avec quelques nuances : la syllabe birmane အ /ʔə̰/ ne se translittère pas séparée de la syllabe qui la suit, et les syllabes superposées ne sont jamais séparées. Par exemple အမတ် /ʔə̰maʔ/ est rendu ⟨ʿamat·⟩ (au lieu de ⟨ʿa mat·⟩) et le mot ကျောင် /kaʔlàʊɴ/ (vieux birman pour ကျောင်း /tɕáʊɴ/ *monastère, école*) devient ⟨kloṅ·⟩ (au lieu de ⟨ka loṅ·⟩).

A part la séparation entre nom propre toponyme et caractéristique géographique, la romanisation du birman nous fournit donc peu d'indices pour la segmentation, car la romanisation traditionnelle accorde une grande importance à l'étymologie sans prendre en compte le sens des éléments à segmenter. Ceci est dû au fait que ces systèmes sont destinés principalement aux chercheurs et aux bibliothécaires au lieu du grand public, comme les systèmes de transcription en vigueur pour le japonais (Hepburn) et le chinois (hànyǔ pīnyīn). Ce dernier en particulier possède un guide de romanisation Yin et al. (2012) qui explicite la segmentation pour la forme romanisée de toutes les parties du discours y compris les noms propres. Il est intéressant de noter que celui-ci explicite le *mot* comme unité d'orthographe et non pas la syllabe qui est toujours représenté par un caractère chinois, bien que le caractère a un statut important dans le système d'écriture et dans les dictionnaires de cette langue.

3.3.6 *Le traitement des entités nommées*

3.3.6.1 Noms personnels birmans

Le nom personnel birman est généralement composé d'un à quatre éléments, et il est souvent précédé par un titre de respect[31].

Dans le choix du nom, les parents sont influencés par nom seulement par leurs préférences personnelles, mais aussi par des considérations astrologiques, les tendances sociologiques, la position de l'enfant dans sa fratrie, ou la tradition familiale (Brac de La Perrière 1999). Ces éléments sont pris parmi du vocabulaire birman connoté positivement (par exemple ရွှေ Shwe *l'or*, စိန် Sein, *diamant*, သက် Thet *souffle, vie*), mais aussi du pâli (comme စံပယ် Sabai *jasmin*, ရတနာ Yadanar *joyaux*, ယုဇန Yuzana *oranger jasmin*) et parfois de l'anglais (ချယ်ရီ Cherry, မေရီ Mary). Le nom personnel d'un Birman ne comporte pas systématiquement une partie partagée avec les autres membres de sa famille, que ce soit les parents ou l'époux(se) (Mi Mi Khaing 1958), mais cela est possible. Par exemple, le frère de အောင်ဆန်းစုကြည် Aung San Suu Kyi s'appelle အောင်ဆန်းဦး Aung San Oo et leurs parents, အောင်ဆန်း Aung San et ခင်ကြည် Khin Kyi. Il n'y a pas toutefois de coupure entre prénom et nom de famille ; bien que les éléments birmans du nom sont puisés dans le stock de vocabulaire ordinaire et ont pour la plupart un sens transparent, il s'agit bien d'un nom propre unitaire qui fait référence à une seule personne. Les Birmans de la diaspora sont souvent obligés pour des raisons administratives d'identifier un prénom et un nom dans leur nom personnel, scindant ainsi leur nom en deux, mais cela est souvent réservé pour la forme en lettres latines, pas celle en écriture birmane. En lettres latines les noms birmans sont translittérés élément par élément, chaque élément séparé par des espaces avec la première lettre en majuscules, comme ceci မိုးသက်ဟန် → Moe Thet Han. Plus difficile que la séparation en éléments, le traitement du nom birman entier comme un seul token a l'avantage d'éviter une surreprésentation surprenante de ces éléments dans le corpus.

La séparation du terme de respect du nom lors de la segmentation peut paraître une évidence, mais ce n'est toujours souhaitable. Il n'est pas toujours aisé de séparer le terme de respect du nom, principalement parce que certains termes

31. N B. : Afin de faciliter la lecture des noms et titres birmans utilisés en anglais, nous utilisons les formes en lettres latines d'usage en anglais, sauf quand il s'agit de traitement informatique de caractères.

de parenté utilisés en tant que termes de respect sont aussi utilisés comme des éléments du nom, les plus fréquents étant ဦး U (parfois écrit Oo à l'intérieur du nom, *oncle, Monsieur*), မောင် Maung (*frère cadet*) et ကို Ko (*frère aîné*). Par exemple la syllabe ဦး apparaît deux fois dans **ဦး**အောင်ဆန်း**ဦး** **U** Aung San **Oo**, au début en tant que terme de respect et à la fin du nom. Parfois l'appellation est ambiguë comme မောင်မောင်ထိုက်အောင် – s'agit-il de *Maung Maung Htike Aung* ou *frère cadet Maung Htike Aung* ? Dans certains corpus, certains termes de respect sont peu fréquents, et le choix de séparer ou non le terme de respect a peu d'incidence sur la fréquence relative des tokens. Un corpus de presse, par exemple, mentionne davantage les personnes adultes, avec peu d'occurrences des termes de respect မောင် Maung et ကို Ko. Dans ce contexte du nom, il s'agit plutôt d'éléments de noms personnels, mais avec une réserve, la syllabe ကို /kò/, est plus fréquente comme marqueur du complément d'objet direct, il est donc préférable de coller le terme de respect au nom à condition d'être certain qu'il s'agit bien d'un nom.

L'argument majeur en faveur de la séparation en deux tokens du titre de respect et du nom est le fait que les titres de respect peuvent changer selon le contexte, l'âge ou la profession. Dans un contexte, la même personne peut être appelée *professeur*, **ဆရာ**လှဘေ **Saya** Hla Pe, ou *docteur*, **ဒေါက်တာ**လှဘေ **Doctor** Hla Pe. L'autre raison pratique de séparer les titres des noms concerne les limitations de notre outil de segmentation qui étant doté d'une liste finie, il est souhaitable de ne pas répéter le même nom de personne plusieurs fois avec des titres de respect différents, mais d'inclure tous les titres de respect une seule fois avec une liste des noms personnels. Dans ce cas, nous aurons trois tokens လှဘေ ⟨lȟ̆be⟩, ဆရာ ⟨s̓ʰra⟩ et ဒေါက်တာ ⟨deäkˣta⟩ pour segmenter ဒေါက်တာ လှဘေ ⟨deäkˣta lȟ̆be⟩ et ဆရာ လှဘေ ⟨s̓ʰra lȟ̆be⟩, et les titres de respect seront réutilisables. Nous n'avons pas donc besoin de rajouter ဒေါက်တာ ⟨deäkˣta⟩ à l'outil de segmentation pour segmenter ဒေါက်တာမောင်မောင် Doctor Maung Maung en ဒေါက်တာ မောင်မောင် ⟨deäkˣta⟩ ⟨meaŋˣmeaŋˣ⟩, juste မောင်မောင် ⟨meaŋˣmeaŋˣ⟩. Dans le cas d'usage de combinaisons de titres de respect, la séparation de ces titres est plus prudente, car la référence à une même personne peut faire usage de la combinaison ou des éléments seuls, comme dans le cas de **သမ္မတဦး**ထင်ကျော် **Président Monsieur** Htin Kyaw, qui peut aussi être désigné **သမ္မတ**ထင်ကျော် **Président** Htin Kyaw ou bien **ဦး**ထင်ကျော် **Monsieur** Htin Kyaw suivant le contexte. Malgré tout ceci,

nous ne pensons pas qu'une application stricte de ce principe soit nécessaire ou souhaitable. Pour certaines personnes l'usage systématique du terme de respect laisse croire qu'il fait partie du nom, car la séparation entre le nom et le titre est moins nette que dans d'autres cultures. Cela a également été observé par Dupertuis (2022), qui note que les titres de respect sont souvent enregistrés comme des éléments du nom de l'auteur dans les catalogues des bibliothèques et sont également utilisés lorsque les locuteurs se désignent eux-mêmes. L'utilisation de titres de respect comme éléments du nom est particulièrement fréquente pour les personnages historiques ou très célèbres. Cette assertion est corroborée non seulement par nos corpus, mais aussi par une observation des pratiques dans le Wikipédia birman. Wikipédia a fait le choix (en principe) de ne pas considérer les termes de respect comme faisant partie du nom personnel dans les titres de ses entrées, mais le principe n'est pas toujours respecté quand il s'agit d'un titre courant. L'entrée de အောင်ဆန်းစုကြည် Aung San Suu Kyi y figure sans ဒေါ် Daw (*Madame*), bien que dans nos corpus les deux se trouvent ensemble systématiquement. L'entrée pour U Nu est နု၊ ဦး Nu, U, pour indiquer le nom, suivi du titre de respect. Par contre, l'entrée pour U Thant est bien U Thant ဦးသန့် bien que သန့် Thant soit son seul nom personnel. Sachant que la performance de l'outil de segmentation est meilleure avec les segments plus longs, il est donc préférable d'inclure le titre de respect dans le nom pour les personnages connus s'il s'agit d'un nom monosyllabique et le terme de respect s'utilise systématiquement pour se référer à cette personne.

Le tableau 3.4 liste les termes les plus fréquents qui précèdent les noms personnels dans nos corpus authentiques.

Afin d'améliorer la segmentation des noms personnels birmans les plus fréquents, nous avons ajouté au dictionnaire interne de notre outil de segmentation *Motor* les données du site *Burmese Names project*[32] sur les 143 éléments les plus fréquemment trouvés dans l'annuaire téléphonique de Birmanie. Les noms étant transcrits en lettres latines, nous avons retranscrit ces 5176 noms en écriture birmane (par le biais d'un script en Perl) et le résultat a été ensuite combiné avec la liste de 2335 noms birmans (Ding, Win Pa Pa et al. 2018)[33] que nous avons modifiée légèrement pour ne pas inclure les titres de respect.

32. https://burmesenames.wordpress.com/ consulté le 7 février 2017
33. http://www.nlpresearch-ucsy.edu.mm/NLP_UCSY/name-db.html consulté le 2 novembre 2016

Terme	Traduction	Exemples d'usage
ဗိုလ်	Lieutenant	**ဗိုလ်**အောင်ကျော် **Bo** Aung Kyaw
ဗိုလ်ချုပ်	Général	**ဗိုလ်ချုပ်**အောင်ဆန်း **Bogyoke** Aung San
ဒေါ်	Madame, tante	**ဒေါ်**အောင်ဆန်းစုကြည် **Daw** Aung San Suu Kyi
ကို	frère aîné	**ကို**ကျောက်စီ **Ko** Kyauk Si
မ	sœur, Madame	မကြည်ကြည်မိုး **Ma** Kyi Kyi Moe
မောင်	frère cadet	မောင်သူရထွန်း **Maung** Thura Htoon
ဆရာ	professeur	**ဆရာ**လှဘေ **Saya** Hla Pe
ဆရာမ	professeure	**ဆရာမ**စံစံနှင်းထွန် **Sayama** San San Hnin Tun
ဦး	Monsieur, oncle	**ဦး**ထင်ကျော် **Mr** Htin Kyaw
မစ္စစ်	Madame	**မစ္စစ်** ကလင်တန် **Mrs** Clinton
မစ္စက်	Madame	**မစ္စက်** မေ **Mrs** May
မစ္စတာ	Monsieur	**မစ္စတာ** ကယ်ရီ **Mr** Kerry
ဒေါက်တာ	Docteur	**ဒေါက်တာ**လှဘေ **Doctor** Hla Pe
သမ္မတ	Président	**သမ္မတ**အိုဘားမား **Président** Obama
ဝန်ကြီးချုပ်	Premier Ministre	**ဝန်ကြီးချုပ်**ထရီဆာမေ **Premier Ministre** Theresa May

TAB. 3.4 : Exemples de termes préfixés aux noms personnels

3.3.6.2 Noms propres étrangers en birman

La seule recommandation pour l'orthographe de noms propres étrangers que nous avons trouvée provient de l'institut စာပေဗိမာန် Sarpay Beikman, l'institut de la promotion de l'écriture et de la littérature birmanes, y compris de la traduction. Selon l'ouvrage de Doctor Tin Aung et al. (p. d.), il est recommandé d'écrire l'emprunt en birman suivant la prononciation en l'anglais. Cette recommandation est problématique à plus d'un titre, donnant lieu à une variété d'orthographes pour les noms propres en birman.

Tout d'abord, il n'y a pas de système standard de transcription de l'anglais en birman. En principe, le seul standard de prononciation enseigné est le standard de l'anglais britannique[34], mais en réalité l'anglais américain a aussi une influence, non seulement en raison de l'influence culturelle des États-Unis, mais aussi en vertu de la formation américaine d'un nombre de professeurs d'anglais même avant l'indépendance de la Birmanie (Khin Khin Aye 2020). Deuxièmement, le choix d'un auteur dépendant non seulement de sa compétence en anglais, mais comment il juge qu'un locuteur de l'anglais pourrait prononcer un nom

34. Ce standard est désigné tantôt *Received Pronunciation* (RP) tantôt *BBC English*.

propre étranger d'une autre langue. Les possibilités de variation orthographiques sont multipliées aussi par le fait que le niveau de compétence des apprenants birmans en anglais est très variable et généralement considéré comme très faible (EF Education First 2021)[35]. La prononciation des apprenants de niveau faible est souvent influencée par la forme écrite en anglais (Than Than Win 1998), qui ne correspond pas systématiquement à la prononciation juste, ce qui rajoute d'avantage de possibilités de translittération en birman. De surcroît, selon Wheatley (2018) l'attribution des tons est en grande partie imprévisible. Il faut noter aussi que cette recommandation ne veut pas forcément dire faire référence aux emprunts anglais existants en birman qui sont adaptés à la phonologie du birman (Chang 2009 ; Wheatley 2018), mais il est possible que l'orthographe des emprunts à l'anglais ait aussi une influence dans l'orthographe de noms propres étrangers. Tous ces différents facteurs se combinent pour créer une certaine imprévisibilité dans l'orthographe de noms propres.

Esche (2005) note une tendance à la diminution de l'influence de l'anglais dans l'orthographe birmane de noms propres de pays non anglophones. Selon elle, *Paris* serait moins souvent orthographié ပဲရစ် ⟨pɛ́rs̀ˣ⟩ que ပါရီ ⟨pärï⟩, mais nous avons trouvé une grande variété dans l'orthographe de noms propres étrangers, que ce soit les noms propres en provenance de pays anglophone ou d'autres pays (parfois même à l'intérieur d'un seul texte). A titre d'exemple, nous avons trouvé 18 façons d'écrire *Emmanuel Macron* en birman attestées dans nos corpus, et étant donné toutes les combinaisons possibles des différentes orthographes de son nom et prénom, d'autres sont encore possibles (voir tableau 3.5).

Au lieu de créer des listes de variants orthographiques de noms propres étrangers qui seraient vite périmées, nous avons décidé d'ajouter ces variants orthographiques au Wikipédia birman en tant que liens de redirection. Ce sont des liens qui renvoient automatiquement sur une autre page, la page d'entrée de la personne en question. On peut visualiser ces variants en consultant les informations sur une page donnée, mais plus important encore, tous les liens de redirection d'un Wikipédia (comme le Wikipédia birman) peuvent être extraits ultérieurement des copies de sauvegarde de Wikipédia[36] pour constituer des listes

35. Il s'agit d'un rapport basé sur les résultats de tests effectués sur deux millions d'adultes dans 112 pays, qui classe les pays en cinq catégories de compétence. Le Myanmar arrive en 93^{e} position.

36. https://dumps.wikimedia.org/

Emmanuel	**Macron**
အီမန်နျူယဲလ် ⟨ʔïmnˣnўùyɛlˣ⟩	မက်ကွန်း ⟨mkˣkŭnˣ²⟩
အီမန်ဝဲလ် ⟨ʔïmnˣwɛlˣ⟩	မက်ခရောန်း ⟨mkˣkʰreanˣ²⟩
အီမန်နျူရယ် ⟨ʔïmnˣnўùryˣ⟩	မာခွန်း ⟨makʰŭnˣ²⟩
အီမန်ခွဲ ⟨ʔïmnˣnŭɛ⟩	မက်ခရွန် ⟨mkˣkʰrŭnˣ⟩
အမ်မန်ခွဲ ⟨ʔmˣmnˣnŭɛ⟩	မက်ခရွန်း ⟨mkˣkʰrŭnˣ²⟩
အေမန်ခွဲ ⟨ʔemnˣnŭɛ⟩	မက်ကရွန်း ⟨mkˣkrŭnˣ²⟩

TAB. 3.5 : Exemple de variants orthographiques d'un nom propre étranger

de variants orthographiques du birman.

Les autres types d'entités nommés tels que les noms de société ou d'organisation sont aussi écrits en écriture birmane, mais on trouve des noms de sociétés souvent orthographiés en lettres latines. Par exemple, on trouve plus souvent le nom de la société pétrolière *Petrobras* écrit **ပက်ထရိုဘရတ်စ်ရေနံကုမ္ပဏီ** ⟨pkˣtʰriuɓrtˣśˣrenñkuꭓmpñï⟩ (*Petrobras* + *pétrole* + *société*), mais on trouve aussi aussi **Petrobras ရေနံကုမ္ပဏီ** ⟨Petrobras renñkuꭓmpñï⟩. Les noms des organisations sont pour la plupart traduits en birman (ဥရောပသမဂ္ဂ /ʔu̯jɔ́pa̯θa̯maʔga̯/ *Europe* + *Union* = *Union européenne*), mais le sigle écrit en birman est le sigle anglais (*EU* → အီးယူ /ʔíjù/). L'influence de l'anglais sur l'usage des sigles est assez particulière. Les lettres latines sont représentées en écriture birmanes par une transcription de la prononciation du nom de la lettre en anglais. Ainsi, si le sigle est prononcé lettre par lettre, c'est la prononciation des lettres de l'alphabet anglais qui est employée. Si le sigle est prononcé à la manière d'un mot, la prononciation anglaise inspire l'orthographe en birman, comme pour la NASA, orthographiée နာဆာ ⟨naśʰa⟩, prononcée /nàsʰà/ ou ASEAN အာဆီယံ ⟨ʔaśʰïyň⟩, prononcée /ʔàsʰìjàɴ/. Il en va de même pour les initiales de noms propres, comme *George H. W. Bush* ဂျော့ အိပ်.ဒဗျူ. ဘွတ်ရှ် /dʑɔ̰ ʔeɪʔ da̯bjù bʊʔʃ/. Le tableau 3.6 reprend les occurrences de lettres anglaises transcrites en birman que nous avons rencontrées dans nos corpus[37].

Les noms personnels étrangers sont parfois écrits en lettres latines à l'intérieur du texte birman, mais le plus souvent on les trouve en écriture birmane (dans

37. Nous avons copié nos données sur Wikipedia https://en.wikipedia.org/wiki/Burmese_respelling_of_the_English_alphabet, car nous avons remarqué la présence d'une liste similaire pour le chinois.

Lettre	Prononciation en anglais britannique	Birman	Exemples
A	/ˈeɪ/	အေ	အမ်အင်အေ MNA, အေဒီ AD
B	/ˈbiː/	ဘီ	ဘီဘီစီ BBC, ဂျေစီဘီ JCB
C	/ˈsiː/	စီ	စီဒီအမ်အေ CDMA, စီအင်အင် CNN
D	/ˈdiː/	ဒီ	ဂျီဒီပီ GDP, ဒီကေဘီအေ DKBA
E	/ˈiː/	အီး, အီ	စီအီးအို CEO
F	/ˈɛf/	အက်ဖ်	အက်ဖ်–၂၂ က်ပ်တာ F22-Raptor
G	/ˈdʒiː/	ဂျီ	ဂျီဒီပီ GDP
H	/ˈeɪtʃ/	အိပ်, အိတ်ချ်, အိတ်	ဂျော့ အိပ်.ဒဗ္ဗျူ. ဘွတ်ရှ် George H. W. Bush
I	/ˈaɪ/	အိုင်	အိုင်အင်န်စီ INC
J	/ˈdʒeɪ/	ဂျေ	ဘီဂျေပီ BJP, ဂျေအေအယ်လ် JAL, ဂျေ ကေ ရိုးလင်း J.K. Rowling
K	/ˈkeɪ/	ကေ	ကေအန်ယူ KNU, ဒီကေဘီအေ DKBA
L	/ˈɛl/	အယ်လ်, အယ်	အင်အယ်လ်ဒီ NLD, အန်အယ်ဒီ NLD
M	/ˈɛm/	အမ်	စီဒီအမ်အေ CDMA
N	/ˈɛn/	အင်, အင်န်, အန်	ဂျီအင်ပီ GNP, အန်အယ်လ်ပီ NLP
O	/ˈoʊ/	အို	စီအီးအို CEO, အင်ဂျီအို NGO
P	/ˈpiː/	ပီ	ဂျီဒီပီ GDP, ပီပီပီ GDP
Q	/ˈkjuː/	ကျူ	ကိုအင်ဇိုင်း ကျူ–၁၀ Coenzyme Q10
R	/ˈɑːr/	အာ, အာရ်	ဂျေ အာ အာ တော်ကီးန် J.R.R. Tolkien အက်စ် အာရ် နာသန် S.R. Nathan
S	/ˈɛs/	အက်စ်, အက်	ဂျီအိုအက်စ် GOS
T	/ˈtiː/	တီ	တီဘီ TB[38]
U	/ˈjuː/	ယူ	ယူအီးအက်ဖ်အေ UEFA
V	/ˈviː/	ဗီ	ဗီလိဂ် ဗွီအိုအေ V-league VOA
W	/ˈdʌbəl.juː/	ဒဗ္ဗျူ, ဒဗလူ	ဂျော့ အိပ်.ဒဗ္ဗျူ. ဘွတ်ရှ် George H. W. Bush
X	/ˈɛks/	အက်စ်	အိတ်ဇ်ရေး X-ray
Y	/ˈwaɪ/	ဝိုင်	
Z	/ˈzɛd/	ဇက်ဒ်	

TAB. 3.6 : Lettres latines transcrites à l'anglaise en birman

de rares cas suivis de l'original entre parenthèses), le nom et le prénom souvent séparés à la même manière qu'en anglais ou en français (exemple 81), mais pas systématiquement (exemple 82).

(81) ဘီလ် ကလင်တန်
⟨b̈il^{x}⟩ ⟨klŋ^{x}tn^{x}⟩
'Bill Clinton'

(82) ဘီလ်ကလင်တန်
⟨b̈il^{x}klŋ^{x}tn^{x}⟩
'Bill Clinton'

(83) ဘီ လ် က လင် တန်
⟨b̈i⟩ ⟨l^{x}⟩ ⟨k⟩ ⟨lŋ^{x}⟩ ⟨tn^{x}⟩
'Bill Clinton'

Si nous ne rajoutons que la forme sans espace (exemple 82) dans l'outil de segmentation, les occurrences dans le corpus qui séparent le nom et le prénom par l'espace (exemple 81) risquent d'être mal segmentées (exemple 83). Il est donc plus prudent d'inclure le nom et le prénom séparément dans l'outil de segmentation. Ainsi, il n'est pas nécessaire d'inclure aussi la combinaison prénom-nom en entier comme celle de l'exemple 82 qui sera de toute façon segmentée comme l'exemple 81. Comme pour les titres de respect, la séparation d'éléments susceptibles d'être réutilisés permet de réduire le nombre d'éléments dans le dictionnaire de l'outil de segmentation, et comme nous voyons dans la figure 3.26, ajouter ဘီလ် ⟨b̈il^{x}⟩ (Bill) et ကလင်တန် ⟨klŋ^{x}tn^{x}⟩ (Clinton) contribue à la segmentation correcte d'autres noms personnels qui contiennent ces éléments et aussi a prévoir les cas du pluriel *Les Clinton* ကလင်တန်တို့ ⟨klŋ^{x}tn^{x}tiu^{1}⟩.

Il permet aussi de prendre en compte les exemples où le prénom ou le nom est utilisé seul, ou le prénom est séparé du nom (exemple 84 segmenté en 85).

(84) ဘီလ်နှင့်မယ်လင်ဒါဂိတ်စ်
⟨b̈il^{x}nn̆ŋ^{1x}my^{x}lŋ^{x}dägit^{x}s̈^{x}⟩
'Bill et Melinda Gates'

(85) ဘီလ် နှင့် မယ်လင်ဒါ ဂိတ်စ်
⟨b̈il^{x}⟩ ⟨nn̆ŋ^{1x}⟩ ⟨my^{x}lŋ^{x}dä⟩ ⟨git^{x}s̈^{x}⟩
Bill et Melinda Gates
'Bill et Melinda Gates'

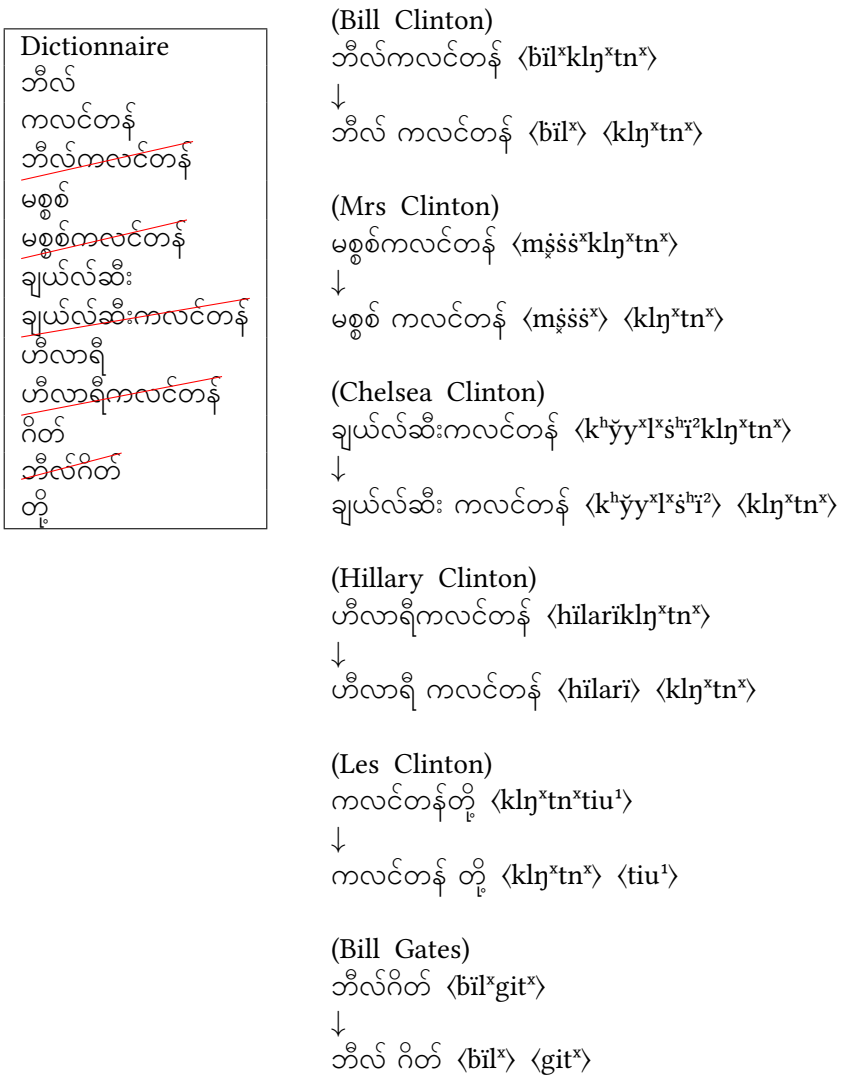


FIG. 3.26 : Segmentation de noms propres personnels

Nous avons trouvé que les titres de respect birmans sont (au moins dans la presse) très peu utilisés quand il s'agit de personnages étrangers. On préfère les titres calqués sur l'anglais comme မစ္စစ် ⟨mɪ̰sɪ̀sɪ̀ˣ⟩ *Madame* (de l'anglais *Mrs*, parfois écrit မစ္စက် ⟨mɪ̰sɪ̀kˣ⟩) et မစ္စတာ ⟨mɪ̰sɪ̀ta⟩ *Monsieur* (de l'anglais *Mister*).

3.3.6.3 Toponymes

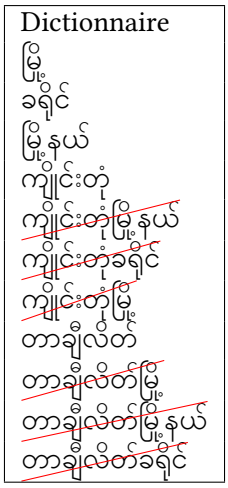
Les toponymes, noms de lieux, villes, fleuves, montagnes, villages et pays, sont souvent composés d'un nom propre et un terme générique qui désigne sa nature. Ainsi, on trouve rarement ပါရီ ⟨pärï⟩ *Paris* tout seul, mais **ပါရီမြို့** ⟨**pärï**mjŭiu¹⟩ ***Paris**-ville*. Un apprenant du birman trouve même que les toponymes sont merveilleusement étiquetés et il doit acquérir le réflexe de se référer aux toponymes de cette manière comme un mot composé. On peut dire ရန်ကုန် ⟨rnˣkunˣ⟩ pour *Yangon*, mais **ရန်ကုန်မြို့** ⟨**rnˣkunˣ**mjŭiu¹⟩ (***Yangon**-ville*) est le terme usuel à l'écrit. Il y a toutefois plusieurs raisons de séparer ces termes de l'entité nommée proprement dite. D'abord, il y a une transparence sémantique du composé pour l'apprenant. Le premier terme est identifié comme un nom propre par le deuxième, nous savons que Paris est un nom propre parce qu'il est désigné en tant que nom de ville par le deuxième terme *ville*, son hyperonyme. La deuxième raison est, comme pour les titres de respect pour les noms de personnes, un seul nom propre peut s'utiliser dans plusieurs toponymes différents. Dans le tableau 3.7, on voit que l'Ayeyarwady peut être le nom d'un fleuve et le nom d'une région.

Toponyme	Segmentation	Traduction
ချင်းတွင်းမြစ်	ချင်းတွင်း + မြစ်	Chindwin + fleuve
ဧရာဝတီမြစ်	ဧရာဝတီ + မြစ်	Ayeyarwady + fleuve
ဧရာဝတီတိုင်းဒေသကြီး	ဧရာဝတီတိုင်း + ဒေသကြီး	Ayeyarwady + région

TAB. 3.7 : Segmentation de toponymes

Séparer le nom propre du nom simple permet de réduire considérablement la liste dans le dictionnaire de l'outil de segmentation. Dans la figure 3.27 qui montre une partie du dictionnaire de *Motor*, on voit qu'il suffit d'ajouter les termes génériques une fois et ensuite le nom propre pour s'assurer d'une segmentation satisfaisante et régulière. Dans ce petit exemple, ce système a réduit

le nombre d'entrées du dictionnaire de six à cinq éléments, mais on imagine l'avantage d'ajouter qu'un seul nom propre, au lieu de trois pour les milliers de noms propres existants.



(Tachileik ville)
တာချီလိတ်မြို့ ⟨takʰy̆ilitˣmj̆iu¹⟩
↓
တာချီလိတ် မြို့ ⟨takʰy̆ilitˣ⟩ ⟨mj̆iu¹⟩

(Tachileik commune)
တာချီလိတ်မြို့နယ် ⟨takʰy̆ilitˣmj̆iu¹nyˣ⟩
↓
တာချီလိတ် မြို့နယ် ⟨takʰy̆ilitˣ⟩ ⟨mj̆iu¹nyˣ⟩

(Tachileik municipalité)
တာချီလိတ်ခရိုင် ⟨takʰy̆ilitˣkʰriuŋˣ⟩
↓
တာချီလိတ် ခရိုင် ⟨takʰy̆ilitˣ⟩ ⟨kʰriuŋˣ⟩

(Kengtung ville)
ကျိုင်းတုံမြို့ ⟨kў̆iuŋˣ²tuňmj̆iu¹⟩
↓
ကျိုင်းတုံ မြို့ ⟨kў̆iuŋˣ²tuň⟩ ⟨mj̆iu¹⟩

(Kengtung commune)
ကျိုင်းတုံမြို့နယ် ⟨kў̆iuŋˣ²tuňmj̆iu¹nyˣ⟩
↓
ကျိုင်းတုံ မြို့နယ် ⟨kў̆iuŋˣ²tuň⟩ ⟨mj̆iu¹nyˣ⟩

(Kengtung municipalité)
ကျိုင်းတုံခရိုင် ⟨kў̆iuŋˣ²tuň kʰriuŋˣ⟩
↓
ကျိုင်းတုံ ခရိုင် ⟨kў̆iuŋˣ²tuň⟩ ⟨kʰriuŋˣ⟩

FIG. 3.27 : Segmentation de toponymes

La troisième raison de séparer les noms propres de leur terme générique est aussi pratique, et concerne la probabilité d'une bonne segmentation. La liste du dictionnaire ne sera jamais complète, car le nombre de mots potentiels dans une langue est infini, et un même nom de lieu peut avoir des variations imprévisibles.

La pagode **Shwedagon**, par exemple, peut apparaître comme **ရွှေတိဂုံ**စေတီ ⟨rwĕheti-guň śetï⟩, **ရွှေတိဂုံ**စေတီတော် ⟨rwĕheti**g**uň śetïtea^{x}⟩, **ရွှေတိဂုံ**ဘုရား ⟨rwĕheti**g**uň b̀ura^{2}⟩ ou **ရွှေတိဂုံ**စေတီတော်ကြီး ⟨rwĕheti**g**uň śetïtea^{x}kjï^{2}⟩, selon que l'on choisit d'apposer le mot pour stupa (စေတီ ⟨śetï⟩ /zèidì/) ou pagode (ဘုရား ⟨b̀ura^{2}⟩ /p^{h}ayá/) avec ou sans l'affixe de révérence တော် ⟨tea^{x}⟩ /tɔ̀/. Si au moins les termes génériques sont dans le dictionnaire, seule une partie du nom propre serait sur-segmentée.

La dernière observation qui justifie une séparation entre noms propres et termes génériques des toponymes est qu'il semble que certains de ces composés sont en effet séparables quand ils apparaissent dans des listes. Les exemples ci-dessous (86 et 87) montrent des cas où le terme နိုင်ငံ ⟨niuŋ^{x}ŋň⟩ *nation* n'est pas répété après chaque nom de pays, mais lui-même suivi d'une marque de pluriel.

(86) တောင်ကိုရီးယား ၊ ဂျပန် နဲ့ ရုရှ ၊ တရုတ် နိုင်ငံ တွေ
⟨teaŋ^{x}kiurï^{2}ya^{2}⟩ , ⟨gy̆pn^{x}⟩ ⟨nɛ^{1}⟩ ⟨rurȟ⟩ , ⟨trut^{x}⟩ ⟨niuŋ^{x}ŋň⟩ ⟨tw̆e⟩
Corée du Sud , Japon et Russie , Chine nation PL
'Les nations de la Corée du Sud, le Japon et la Russie, la Chine'

(87) မလေးရှား ၊ ဖိလစ်ပိုင် ၊ ဗီယက်နမ် နိုင်ငံ တို့
⟨mle^{2}rȟa^{2}⟩ , ⟨p^{h}ilś^{x}piuŋ^{x}⟩ , ⟨bïyk^{x}nm^{x}⟩ ⟨niuŋ^{x}ŋň⟩ ⟨tiu^{1}⟩
Malaisie , Philippines , Vietnam nation PL
'Les nations de la Malaisie, les Philippines, le Vietnam'

3.3.6.4 Identification d'entités nommées

Afin d'identifier les entités nommées pour les rajouter au dictionnaire interne de *Motor*, nous avons utilisé deux stratégies. La première est de les identifier dans un concordancier, le deuxième est l'usage des ngrammes[39].

L'usage de concordancier nous permet d'identifier les noms propres à l'aide de leurs contextes gauche et droit. Pour les noms propres personnels, birmans et étrangers, ceci se fait à l'aide d'une liste de termes de respect habituellement préfixés aux noms personnels (voir le tableau 3.4) et pour les toponymes au moyen de listes de termes génériques des toponymes (voir le tableau 3.8). Après

39. Ce processus d'extraction de noms propres pourrait aussi se faire dans un logiciel d'analyse linguistique, tel que *Nooj* (Silberztein 2015), mais ces logiciels ne prennent pas (encore) en charge la technologie nécessaire pour l'affichage correct des polices birmanes. L'intégrité des données est assurée, mais nous avons trouvé qu'un usage approfondi à ce stade est trop fastidieux, nécessitant l'export des résultats à chaque manipulation pour la vérification.

Toponyme	**Traduction**
ရွာ ⟨rwă⟩	village
မြို့ ⟨mjiu¹⟩	ville
မြို့နယ် ⟨mjiu¹nyˣ⟩	commune
ခရိုင် ⟨kʰriuŋˣ⟩	municipalité
ပြည် ⟨pjŋˣ⟩	pays
နိုင်ငံ ⟨niuŋˣŋň⟩	nation
မြစ် ⟨mjṡˣ⟩	fleuve
ကန် ⟨knˣ⟩	lac
တောင် ⟨teaŋˣ⟩	montagne
စေတီ ⟨ṡetï⟩	stūpa
စေတီတော် ⟨ṡetïteaˣ⟩	stūpa
ဘုရား ⟨b̥uraˀ⟩	pagode
စေတီတော်ကြီး ⟨ṡetïteaˣkjïˀ⟩	pagode

TAB. 3.8 : Termes génériques permettant d'identifier les toponymes

une première segmentation, nous générons une concordance qui affiche toutes les occurrences du terme de respect en contexte. Ce terme devient le mot pôle de la concordance. Afin de vérifier la segmentation des noms propres au contexte droit, les résultats sont triés par ordre alphabétique du contexte droit pour grouper les mêmes noms. La figure 3.28 illustre une recherche dans le corpus BBC avec pour mot pôle le terme မစ္စတာ ⟨mṡ̥ṡ̥ta⟩ *Monsieur* au moyen du logiciel TXM (Heiden et al. 2010). Les noms personnels (ici, il s'agit de noms de famille) sont en gras. L'entité à identifier est identifiable par certaines syllabes qui la suivent fréquemment, telles que les marqueurs de thème, sujet ou complément d'objet direct (ဟာ ⟨ha⟩, က ⟨k⟩ ou ကို ⟨kiu⟩), ou conjonctions de coordination (နဲ့ ⟨nɛ¹⟩ *avec, et*). Dans le premier exemple, on voit que tous les éléments entre မစ္စတာ et က, **ကဲ နဲ့ ် ယာ တား** ne sont qu'une sursegmentation du nom de famille **ကဲနဲ့်ယာတား** (Kenyatta).

Bien que très précise, la méthode de détection à l'aide de concordancier est plutôt lente, car elle ne réunit pas automatiquement les noms sursegmentés. Elle repose sur l'œil humain pour détecter les erreurs et ne nous fournit aucune information statistique sur la fréquence ou la probabilité de retrouver certaines syllabes ensemble. Pour cette raison, nous avons eu recours à des méthodes qui utilisent des n-grammes qui regroupent des syllabes. En regardant les fréquences

text_id	Contexte gauche	Pivot	Contexte droit
1322	ဖြစ် တယ် လို့	မစ္စတာ	**ကဲ နဲ့ ယာ တား** က ဆို ပါ တယ်
0374	ဖြစ် နေ တဲ့	မစ္စတာ	**ကု ရှ ိ နာ** ဟာ ပြီး ခဲ့ တဲ့ နှစ်
0729	မဟုတ်ဘူး လို့ လည်း	မစ္စတာ	**ခွ တ ိ ရှ ိ နာ** က ဆို ပါ
0237	တွေ့ဆုံ ပွဲ မှာ	မစ္စတာ	**ထရမ့် ပ** ိ ကို ပေးအပ် ခဲ့ ပါ တယ် ။
0998	ခင် ဒီ ကိစ္စ	မစ္စတာ	**ထရမ့်** နဲ့ ဆွေးနွေး ခဲ့ တာ လို့ ပြင်သစ် သတင်းထောက် တွေ ကို

FIG. 3.28 : Exemple d'extraction de noms propres personnels par concordancier

de ces groupes de syllabes et les calculs statistiques qui indiquent la probabilité de cooccurrences de syllabes, nous pouvons en tirer des indications sur la lexicalisation, ou le figement d'unités lexicales dans la langue. Puisque nous n'avons utilisé cette méthode que pour les noms propres, nous la décrirons dans une section séparée.

3.3.7 *Utilisation de n-grammes pour identifier le vocabulaire polysyllabique*

Afin d'identifier quelles séquences de syllabes sont les plus fréquentes, c'est-à-dire les éléments qui apparaissent plus fréquemment ensemble, nous calculons les fréquences des n-grammes de différentes longueurs de syllabes et la probabilité d'apparition de ces syllabes ensemble par rapport au reste du corpus.

Un n-gramme est une séquence de tokens de longueur n découpée dans une suite ininterrompue du texte. Puisque nous nous intéressons au nombre de syllabes par vocable, un token équivaut à une syllabe. Un vocable à trois syllabes se présenterait comme un trigramme ($n = 3$), une séquence de trois tokens. Découper un texte en n-grammes signifie générer tous les n-grammes possibles d'une longueur n donnée. Pour illustrer ce procédé, voici un exemple (88), suivi de sa segmentation en syllabes. Dans la phrase segmentée, nous remplaçons l'espace typographique par le tiret bas, car l'espace se comporte plutôt comme une marque de ponctuation en birman et nous ne voulons pas perdre cette information.

(88) ဒီအခန်းကို လူတိုင်းသိကြတယ်။
ဒီ အ ခန်း ကို _ လူ တိုင်း သိ ကြ တယ် ။
'On connaît tous la scène.'[40]

40. Bien que la phrase birmane peut avoir plusieurs interprétations, il s'agit en l'occurrence

		ဒီ	အ	ခန်း	ကို	_	လူ	တိုင်း	သိ	ကြ	တယ်	။
bigrammes	1	ဒီ	အ									
	2		အ	ခန်								
	3			ခန်း	ကို							
	4				ကို	_						
	5					_	လူ					
	6						လူ	တိုင်				
	7							တိုင်း	သိ			
	8								သိ	ကြ		
	9									ကြ	တယ်	
	10										တယ်	။
trigrammes	1	ဒီ	အ	ခန်								
	2		အ	ခန်	ကို							
	3			ခန်း	ကို	_						
	4				ကို	_	လူ					
	5					_	လူ	တိုင်				
	6						လူ	တိုင်း	သိ			
	7							တိုင်း	သိ	ကြ		
	8								သိ	ကြ	တယ်	
	9									ကြ	တယ်	။

TAB. 3.9 : Illustration du découpage en n-grammes

Le tableau 3.9 illustre le processus pour le découpage en n-grammes de la chaîne de onze syllabes de l'exemple 88 en n-grammes. Quand $n = 2$, nous obtenons dix bigrammes, si $n = 3$, nous obtenons neuf trigrammes.

Pour la génération automatique de n-grammes, nous utilisons N-Gram Processor[41] (NGP) (Buerki 2017a), une version compatible avec l'Unicode du générateur de n-grammes du N-Gram Statistics Package (Text-NSP)[42] (Banerjee et Pedersen 2003), un module Perl qui permet de générer des n-grammes, mais aussi d'effectuer des analyses statistiques. Nous utilisons le module `statistic.pl` de Text-NSP par la suite.

d'une traduction en birman de la première ligne de la conférence TED *The Power of Simple Words* de Terin Izil qui est en anglais.

41. Version 0.6 http://buerki.github.io/ngramprocessor/

42. https://www.d.umn.edu/~tpederse/nsp.html

NGP permet à l'utilisateur de définir les tokens qui constituent les n-grammes[43] et aussi une liste de tokens qui ne devraient pas faire partie d'un n-gramme (une liste des mots outils, par exemple, ou dans le cas présent la ponctuation). L'usage simple consiste à exclure les n-grammes qui ne correspondent pas à des éléments du vocabulaire valables, comme les bigrammes 4, 5 et 10 et les trigrammes 3, 4, 5 et 9 du tableau 3.9, en rajoutant le tiret bas _ et les marques de ponctuation comme ။ à cette liste d'exclusion. Un usage plus élaboré sera détaillé plus loin.

3.3.7.1 Le module list.pl

L'outil de base, `list.pl`, fournit une liste de n-grammes par ordre de fréquences décroissantes. La sortie standard (sans options spécifiques) donne le nombre de n-grammes en premier, suivi de la liste des n-grammes et chaque n-gramme de la liste est suivi de sa fréquence dans le corpus. Dans cet exemple, qui affiche les premières lignes de la liste de bigrammes de notre corpus BBC Burmese, 562 707 bigrammes ont été extraits du corpus[44], le séparateur <> apparaît entre les éléments du bigramme[45] :

```
562707
ပါ<>တယ်<>13653
နိုင်<>ငံ<>4748
တွေ<>ကို<>3147
တယ်<>လို့<>2908
ပြော<>ပါ<>2355
အ<>တွက်<>2225
...
```

FIG. 3.29 : Premières lignes de la liste de bigrammes

Le module `list.pl` de NGP permet de créer la liste de n-grammes dans un

43. Cette étape de définition des tokens définit plutôt ce qui n'est pas un token, et utilise par défaut la ponctuation et les espaces typographiques comme délimiteurs. Puisque la segmentation en syllabes utilise le caractère espace comme délimiteur, notre prétraitement avant la segmentation remplace l'espace typographique par le tiret bas _, ainsi transformant l'espace du texte d'origine en token.

44. Ce chiffre comprend les doublons. Il représente donc la taille du corpus, pas le nombre de bigrammes uniques.

45. Cet usage de <> pour séparer les éléments de n-grammes (dans notre cas ce sont toujours des syllabes) ne doit se confondre avec les chevrons ⟨ ⟩, utilisés pour indiquer une translittération.

format compatible avec le module `statistic.pl` de Text-NSP. Le n-gramme est suivi de la fréquence de ses éléments ensemble, puis de la fréquence de chaque élément dans la même position dans tous les n-grammes du corpus. Si nous prenons comme exemple la première ligne de la liste dans la figure 3.29, la sortie compatible `statistic.pl` est la suivante :

ပါ<>တယ်<>13653 18043 18050

Le bigramme ပါ တယ် apparaît 13 653 fois dans le corpus, la première syllabe ပါ 18 043 fois en première position parmi les 562 707 bigrammes du corpus et la deuxième syllabe တယ် 18 050 fois en deuxième position parmi tous les bigrammes du corpus. Comme l'explique Banerjee et Pedersen (2003), en combinaison avec la valeur du nombre total de bigrammes dans le corpus (sur la première ligne de la liste) ce sont les seules valeurs requises pour calculer toutes les autres du tableau de contingence du bigramme 3.10. Par exemple, les bigrammes qui commencent par ပါ qui n'ont pas တယ် en deuxième position (indiqué par le point d'exclamation !တယ်) sont au nombre de 4390, soit le nombre d'occurrences de ပါ တယ်, 13 653, soustrait du nombre total d'occurrences de ပါ en première position, 18 043.

	တယ်	!တယ်	TOTAL
ပါ	13 653	4390	18 043
!ပါ	4397	540 267	544 664
TOTAL	18 050	544 657	562 707

TAB. 3.10 : Tableau de contingence

3.3.7.2 Le module statistic.pl

Ce sont ces valeurs qui sont utilisées pour estimer la dépendance entre les éléments d'un n-gramme par `statistic.pl`. Ce module permet de calculer si l'apparition des éléments d'un n-gramme ensemble est plus probable que le hasard, c'est-à-dire le degré de dépendance des éléments d'un n-gramme. Cette approche statistique pour évaluer la cohésion de mots composés a été utilisée par Aroonmanakun (2002). On peut considérer les statistiques qui en résultent comme une mesure de lexicalisation.

Le module propose plusieurs tests d'association, dix pour les bigrammes[46], quatre pour les trigrammes[47] et un seul pour les tétragrammes[48]. Nous avons concentré nos analyses sur le test du rapport de vraisemblance, considéré plus robuste comme mesure pour les fréquences les moins élevées (Dunning 1993), et aussi parce qu'il s'agit du seul test disponible pour les trois longueurs de n-gramme. Le rapport de vraisemblance mesure l'écart entre les données observées et ce que l'on attendrait si les syllabes étaient indépendantes. Plus le résultat est élevé, moins il y a de preuves en faveur de l'indépendance des syllabes.

3.3.7.3 L'outil Substring

En complément de N-Gram Processor, Buerki (2017b) propose un outil pour consolider les n-grammes, appelé `Substring`[49]. Cet outil réduit la fréquence de n-grammes par la fréquence des n-grammes dans lesquels ils occurrent. Autrement dit, cette procédure permet de réduire la fréquence des n-grammes qui font partie d'autres n-grammes plus longs. L'algorithme examine en premier les n-grammes les plus longs, et modifie les n-grammes $n-1$ qui s'y trouvent, puis $n-2$ et ainsi de suite. Dans l'exemple simple 3.30, on regarde d'abord le trigramme ဘီ ဘီ စီ ($n=3$ donc) avec 878 occurrences et on trouve les bigrammes ($n=3-1$ donc $n=2$) qu'il contient : ဘီ ဘီ a 879 occurrences et ဘီ စီ en a 883. Nous ne voulons pas compter les 878 occurrences de ces bigrammes qui font partie du trigramme, donc le nombre d'occurrences des bigrammes est réduit par le nombre de trigrammes : ဘီ ဘီ $879-878=1$; ဘီ စီ $883-878=5$.

L'algorithme commence toujours par les n-grammes les plus longs, en procédant vers les n-grammes les moins longs. Par exemple, le n-gramme à six syllabes ဘီ<>ဘီ<>စီ<>မြန်<>မာ<>ပိုင်း<>43 (*BBC Burmese Service*) va réduire ဘီ<>ဘီ<>စီ<>878 par 43 occurrences. Les deux n-grammes possèdent

46. Les tests d'association proposés pour les bigrammes sont le test exact de Fisher (le test unilatéral gauche ou droit et le test bilatéral), le coefficient de Jaccard, Log-likelihood ratio (*le rapport de vraisemblance*), l'information mutuelle, Odds Ratio (*le rapport des cotes*), Pointwise Mutual Information (*l'information mutuelle spécifique*), le coefficient Phi, le test du χ^2 de Pearson, la mesure Poisson-Stirling, et le T-score.

47. Les tests d'association proposés pour les trigrammes sont la Log-likelihood ratio (*le rapport de vraisemblance*), l'information mutuelle, Pointwise Mutual Information (*l'information mutuelle spécifique*) et la mesure Poisson-Stirling.

48. Le seul test d'association proposé pour les tétragrammes est la Log-likelihood ratio (*le rapport de vraisemblance*).

49. http://buerki.github.io/SubString/

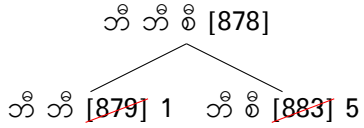


FIG. 3.30 : Exemple de consolidation de fréquences de n-grammes

toujours une fréquence élevée, 835 et 43 occurrences respectivement, indiquant que ဘီဘီစီမြန်မာပို့ငှ်း ⟨bïbïsïmǰnˣmapiuŋˣ²⟩ /bìbìsìmjànmàpáɪn/ (*BBC Burmese Service*) aussi bien que ဘီဘီစီ ⟨bïbïsï⟩ /bìbìsì/ (*BBC*) sont lexicalisés et devraient être idéalement inclus dans le dictionnaire de *Motor* pour être segmentés comme un token chacun. Le service de diffusion birman de la BBC est donc bien un concept d'un point de vue statistique. Nous constatons aussi que ဘီဘီစီမြန်မာပို့ငှ်း ⟨bïbïsïmǰnˣ-mapiuŋˣ²⟩ a sa propre page sur le Wikipédia birman, ce qui lui donne le statut d'une entité nommée d'un point de vue conceptuel de la part du wikinaute qui a créé la page. Cependant, nous voulons autant que possible réduire la taille du dictionnaire en évitant d'inclure les tokens dont les éléments sont fréquents et transparents. En d'autres termes, si un apprenant peut deviner le sens à partir de composants qui sont fréquents, les inclure dans les composés est redondant. Dans notre exemple, le bigramme des syllabes formant le nom du pays မြန်မာ ⟨mǰnˣma⟩ *Birmanie* est très fréquent, မြန်<>မာ<>1709, il est donc inutile de rajouter ဘီဘီစီမြန်မာပို့ငှ်း ⟨bïbïsïmǰnˣmapiuŋˣ²⟩ au dictionnaire, car ses composants ဘီဘီစီ ⟨bïbïsï⟩, မြန်မာ ⟨mǰnˣma⟩ et ပို့ငှ်း ⟨piuŋˣ²⟩[50] y figurent déjà et ils ont tous des fréquences élevées en dehors du composé. De plus, nous avons trouvé quelques instances de cette chaîne de caractères avec le caractère espace au milieu, ainsi : ဘီဘီစီ မြန်မာပို့ငှ်း ⟨bïbïsï mǰnˣmapiuŋˣ²⟩. Le choix de scinder en trois tokens assure la régularité de segmentation pour une même entité nommée. Il faut donc faire très attention avec les chaînes de caractères longues, même si d'un point de vue statistique elles pourraient être considérées comme une unité, dans la pratique, les segmenter est plus prudent. Finalement, on segmente cette chaîne de la même manière qu'en anglais (BBC<>Burmese<>Service<>), mais puisque la formation de mots composés n'est pas toujours identique dans les deux langues, regarder la segmentation dans d'autres langues n'est pas un critère utile pour cerner la segmentation. Cette asymétrie est illustrée par les noms des ministères

50. ပို့ငှ်း ⟨piuŋˣ²⟩ apparaît 551 fois dans le corpus BBC Burmese et 3636 fois dans l'ensemble de nos corpus authentiques.

en français et en birman. Un seul est formé sur le même modèle que le français *ministère des Affaires étrangères*, နိုင်ငံခြားရေး ဝန်ကြီးဌာန ⟨niuŋˣŋňkʰǰa²re² wnˣkǰï²ťan⟩. Ces noms sont habituellement écrits avec un caractère espace avant le composé qui veut dire *ministère* ဝန်ကြီးဌာန ⟨wnˣkǰï²ťan⟩ (ဝန်ကြီး ⟨wnˣkǰï²⟩ *ministre* + ဌာန ⟨ťan⟩ *département*), bien que d'un point de vue statistique, la fréquence du nom entier du ministère n'est pas insignifiant (voir tableau 3.11). Il est donc prudent de toujours le segmenter en deux parties. La première partie est structurée de façon identique pour tous les ministères avec comme terminaison la syllabe ရေး ⟨re²⟩ /jé/ qui dans ce contexte a le sens de *condition, affaires*. Ainsi, *les affaires étrangères*, နိုင်ငံခြားရေး ⟨niuŋˣŋňkʰǰa²re²⟩, est composé de နိုင်ငံ ⟨niuŋˣŋň⟩ *nation* + ခြား ⟨kʰǰa²⟩ *différent* + ရေး ⟨re²⟩ *affaire*.

Birman	Décomposition	Français	*f*
ကျန်းမာရေး ဝန်ကြီးဌာန	être en forme + affaire + ministère	ministère de la Santé	3
ကာကွယ်ရေး ဝန်ကြီးဌာန	défendre + affaire + ministère	ministère de la Défense	4
နိုင်ငံခြားရေး ဝန်ကြီးဌာန	nation + autre + affaire + ministère	ministère des Affaires étrangères	7
ပြည်ထဲရေး ဝန်ကြီးဌာန	pays + intérieur + affaire + ministère	ministère de l'Intérieur	7
ပြန်ကြားရေး ဝန်ကြီးဌာန	répondre + affaire + ministère	ministère de l'Information	8
ပညာရေး ဝန်ကြီးဌာန	éducation + affaire + ministère	ministère de l'Éducation	6

TAB. 3.11 : Comparaison de noms de ministères en français et en birman. Contrairement au français, la composition des noms de ministère en birman suit un schéma régulier. La fréquence *f* indiquée est celle trouvée dans le corpus *BBC Burmese*.

3.3.7.4 Utilisation de listes d'exclusion pour le birman

Le rôle d'une liste d'exclusion est d'écarter d'office des n-grammes qui contiennent des tokens (syllabes ou ponctuations) qui ne font pas partie des séquences de syllabes que nous voulons faire ressortir. Nous avons déjà illustré ce processus pour la ponctuation et l'espace dans le tableau 3.9. Son usage lors du processus de génération de n-grammes permet non seulement de réduire la taille du fichier résultant, mais aussi des fournir des informations concernant les frontières du vocabulaire non identifié, car lorsqu'on rajoute à la liste d'exclusion des éléments du vocabulaire que nous avons décidé avec certitude constituent des éléments valables du vocabulaire, les deux extrémités d'un tel élément contribuent

à définir les frontières d'autres éléments. Reprenons l'exemple 88, qui contient le segment ဒီ အခ ခန်း ကို. Si nous rajoutons aux signes de ponctuation déjà présents dans la liste d'exclusion (_ et ။) des unités monosyllabiques qui ont une fonction grammaticale, telles que ကို ⟨kiu⟩ /kò/ (-OBJ) ou d'autres unités monosyllabiques qui appartiennent à des catégories morphosyntaxiques closes comme le démonstratif ဒီ ⟨dï⟩ /dí/ (*ce*), cela permet d'identifier le bigramme အခ ခန်း comme unité significative[51]. Puisque le pluriel est marqué simplement par la suffixation d'une syllabe à un substantif (ဇတ္ထ ⟨tw̆e⟩ /twè/ ou များ ⟨mw̆ya²⟩ /mjá/) ou à un verbe (ကြ ⟨kw̆⟩ /tɕa̰/), l'ajout de ces syllabes à la liste d'exclusion permet de fusionner le singulier et pluriel, ainsi identifiant l'extrémité droite de ces éléments quand ils apparaissent au pluriel dans le corpus.

Bien que la mise en œuvre d'une liste d'exclusion soit a priori simple, son utilisation présente le risque d'exclure des syllabes qui font partie de vocables valides. La marque de fin de phrase တယ် ⟨tyˣ⟩ /tɛ̀/ est la première syllabe de တယ်လီဖုန်း ⟨tyˣlïpʰunˣ²⟩ /tɛ̀lìpʰóʊɴ/ (*téléphone*) par exemple. L'inclusion de cette marque de fin de phrase dans la liste d'exclusion va empêcher `list.pl` de générer le trigramme တယ်<>လီ<>ဖုန်း. A cause de cela, l'utilisation de ce procédé ne permet pas de générer automatiquement tout le vocabulaire du texte en même temps ; il est nécessaire d'utiliser plusieurs listes d'exclusion différentes et comparer les résultats. Nous l'utilisons pour découvrir le vocabulaire qui manque à notre dictionnaire interne de *Motor* qui contient déjà un dictionnaire, ce qui réduit considérablement le risque d'exclure le vocabulaire très fréquent avec une liste d'exclusion.

Potentiellement, tous les éléments monosyllabiques appartenant à des catégories morphosyntaxiques à liste finie (pronoms et éléments indiquant les relations syntaxiques) sont candidats pour faire partie de notre liste d'exclusion, mais au lieu d'appliquer une liste complète d'emblée, nous avons décidé d'examiner les n-grammes résultants de listes partielles, permettant ainsi d'éviter le problème d'exclusion trop extensive. Nous avons commencé par regarder les n-grammes contenant les noms propres pour identifier les éléments les plus fréquents qui les accompagnent. La figure 3.31 fournit deux exemples. Le premier montre les n-grammes à sept éléments qui contiennent le nom propre ဒေါ်အောင်ဆန်းစုကြည် (*Mme Aung San Suu Kyi*) et cinq différents éléments de structure qui le suivent

51. Dans le contexte de la phrase အခန်း ⟨ʔkʰnˣ²⟩ /ʔa̰kʰáɴ/ veut dire *scène* dans un film.

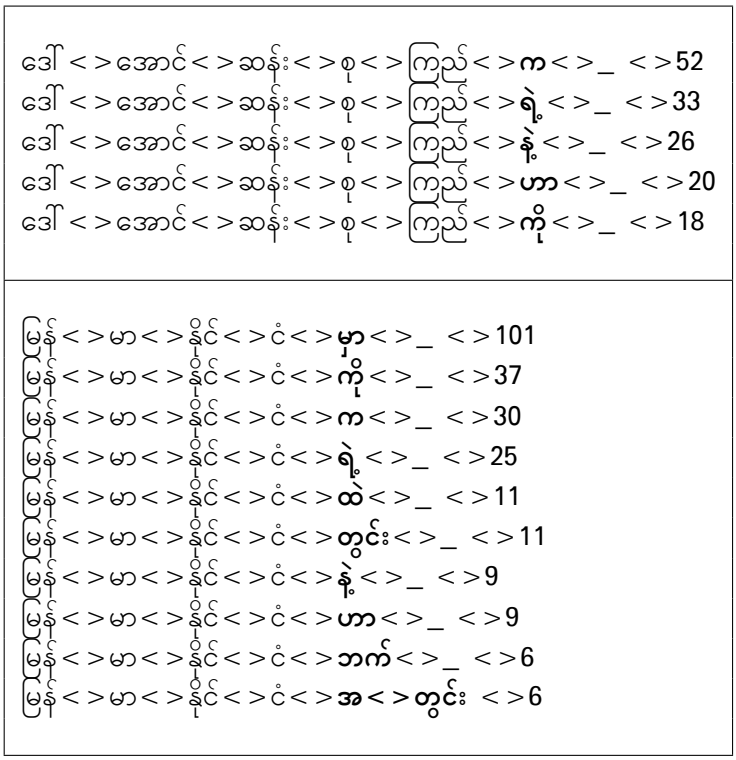


FIG. 3.31 : L'identification de candidats (en gras) pour une liste d'exclusion à l'aide de noms propres

(en gras), suivi de leurs fréquences dans nos corpus authentiques.

Puisque nous considérons le nom propre comme un ou plusieurs tokens, nous pouvons considérer l'élément qui le(s) suit, s'il est lui-même suivi de l'espace typographique, comme un token. De la même manière que l'exemple précédent (de la figure 3.31), dans le deuxième exemple, les n-grammes à six éléments contenant မြန်မာနိုင်ငံ (*Birmanie*), nous aident à identifier d'autres éléments de structure que nous pourrons ajouter à notre liste d'exclusion. Nous avons aussi parcouru les scores de rapport de vraisemblance (calculés avec `statistic.pl`) pour les bigrammes, d'abord pour ceux dont le deuxième élément est le tiret bas (le tableau 3.12).

Le dernier n-gramme de la figure 3.31 (*à l'intérieur de la Birmanie*) qui termine par အ < > တွင်း (les deux syllabes sont suivies de l'espace quand on les examine à l'intérieur de n-grammes à sept éléments) nous donne aussi une piste intéressante. La syllabe အ ⟨ʔ⟩ /ʔa̰/ est de loin la plus fréquente dans notre corpus de plus de

Bigramme	**Rang**	**Score**	***f*** **bigramme**	***f*** **1er élément**	***f*** **2e élément**
ကို< >_ < >	3	126464.5882	53710	68384	666486
က< >_ < >	6	78412.2593	47316	76101	666486
တဲ့< >_ < >	9	64373.1926	28857	37788	666486
မှာ< >_ < >	14	51281.0988	27059	39525	666486
ဟာ< >_ < >	16	48248.5885	17627	20100	666486
နဲ့< >_ < >	17	44872.6534	19668	25289	666486
ဖတ့< >_ < >	23	36280.6277	18517	26369	666486

TAB. 3.12 : Exemples de scores de rapport de vraisemblance pour bigrammes avec tiret bas en position finale (f = fréquence)

quatre millions de syllabes (voir le tableau B.1) avec 147 739 occurrences. Elle est considérée par Hopple (2003) comme un morphème lié. Nous avons trouvé qu'un prétraitement fusionnant celui-ci avec la syllabe qui suit dans nos corpus était en effet très utile pour améliorer les performances de la recherche de n-grammes et la segmentation.

Ces observations ont été comparées aux catégories d'éléments syntaxiques (et pronoms) proposées par Bernot, Cardinaud et al. (2001) (*Manuel de birman, volume 2 : Grammaire birmane*, voir le tableau 3.13 pour des exemples) et la liste de formes grammaticales triées par catégorie qui se trouve à la fin du *Burmese/Myanmar Dictionary of Grammatical Forms* (Okell et Allott 2017) (voir le tableau 3.14). Nous avons examiné non seulement les éléments monosyllabiques, mais aussi les éléments polysyllabiques, en comparant les scores de rapport de vraisemblance d'éléments polysyllabiques des ouvrages que nous venons de citer. Ensuite, un prétraitement qui reconstitue les éléments polysyllabiques dans les corpus a été effectué avant de regénérer les n-grammes.

Nous avons privilégié une approche conservatrice, dans la mesure où les formes qui ont un homographe avec un sens autre que grammatical ont finalement été exclues de la liste (comme les auxiliaires). Les entrées qui concernent les mots composés ou pourraient être interprétées comme des composants (*Common elements in compound verbs, Common elements in compound nouns* dans Okell et Allott (2017)) ont été écartées des listes d'exclusion.

Catégorie	Exemples
Marques modales	သည်; ပြီ; တယ်
L'impératif	နဲ့; နှင့်
Marques de fin de phrase	လား၊ လဲ (interrogation); ထာ (exclamation)
Auxiliaires	နေ *en train de*; ပေး *pour*
Marques verbales	ခဲ့ (passé ou changement d'état); ကြ (marqueur du pluriel)
Classificateurs	ခု (général) ကောင် (animaux)
Marqueurs du nom	က (origine); မှာ (lieu); ရဲ့ (possession)
Noms auxiliaires	အရ *selon*; အထဲ (l'intérieur)
Pronoms personnels	ငါ 1SG; ကျွန်မ 1SG (féminin)
Pronoms interrogatifs	ဘာ *que, quoi*
Pronoms indéfinis	တစ်ခုခု *quelque chose*; ဘာဖြစ်ဖြစ် *n'importe quoi*; ဘာမှာ *rien*
Démonstratifs	ဒါ၊ အဲဒါ၊ အဲဒီ *ceci*
Démonstratifs interrogatifs	ဘယ် *quel*
Subordonnées au nom	သော၊ သည့်၊ တဲ့၊ မည့်၊ မဲ့; သောအခါ *quand*
Subordonnées au verbe	လို့ *comme, parce que*; ပြီး *quand, après que*

TAB. 3.13 : Catégories de Bernot, Cardinaud et al. (2001) pour les listes d'exclusion

Catégorie	Exemples
1 Clauses and verb attributes	နေ *living in*; ရှိ *situated in*
6 Coordinate markers	နှင့်၊ နဲ့ *and, with*; သော *neither ... nor*
7 Location nouns	ပေါ် *upon*; အပေါ် *upon*
8 Noun attribute markers	၏ *'s, belonging to*
9 Noun markers	က *subject, from*; ကို *object*; ကတည်းက *ever since*
11 Selectives	ဒီ *this, that*; ဘယ်နှစ် *how many?*
12 Sentence final phrase particles	ပေါ့ *of course*; ပါ *polite*
13 Sentence markers	မည် (future statement)
17 Subordinate sentence markers	အတိုင်း *according to*; အကြောင်း *concerning*
18 Verb attribute markers	ရန် *in order to, for*; ဖို့ *for V-ing*

TAB. 3.14 : Catégories de Okell et Allott (2017) pour les listes d'exclusion

3.3.7.5 Evaluation de méthodes basées sur n-grammes

Afin d'évaluer l'efficacité et donc l'utilité de méthodes basées sur les n-grammes, nous les avons testées sur le corpus *myPOS* (Khin War War Htike et al. 2017) qui est fourni segmenté de deux façons différentes, l'une qui est qualifiée de segmentation en *words* (*mots*) et l'autre en *compound words* (*mots composés*). Les auteurs de ce corpus ne fournissent pas de définition de *word* ou *compound*

word, et il semblerait que les choix de segmentation ne soient pas totalement homogènes, mais à la lecture des fichiers nous constatons une certaine logique de différentiation. La figure 3.32 contient deux phrases du corpus pour illustrer les deux segmentations[52].

Segmentation en *Words*	**Segmentation en *Compound Words***
ဆရာကြီး ဦးစိန် ၏ မြန်မာ အထက်တန်း ကျောင်း မှ ဓ တန်း အောင်မြင် ခဲ့ ပြီး မင်းလှ မြို့ နော်မံ ကျောင်း မှ အလယ်တန်း ဆရာ ဖြစ် အောင်မြင် ခဲ့ ၏ ။	ဆရာကြီးဦးစိန် ၏ မြန်မာအထက်တန်းကျောင်း မှ ဓ တန်း အောင်မြင် ခဲ့ ပြီး မင်းလှမြို့ နော်မံကျောင်း မှ အလယ်တန်း ဆရာ ဖြစ် အောင်မြင် ခဲ့ ၏ ။
ရှစ်လေးလုံး လှုပ်ရှား မှု တွင် တက်ကြွ စွာ ပါဝင် ခဲ့ သလို ၊ ဗမာ နိုင်ငံ လုံး ဆိုင်ရာ ကျောင်းသား သမဂ္ဂ များ အဖွဲ့ချုပ် မှာ လည်း တက်ကြွ စွာ ပါဝင် လှုပ်ရှား ခဲ့ သည် ။	ရှစ်လေးလုံးလှုပ်ရှားမှု တွင် တက်ကြွ စွာ ပါဝင် ခဲ့ သလို ၊ ဗမာနိုင်ငံလုံးဆိုင်ရာကျောင်းသားသမဂ္ဂ များ အဖွဲ့ချုပ် မှာ လည်း တက်ကြွ စွာ ပါဝင် လှုပ်ရှား ခဲ့ သည် ။

FIG. 3.32 : Les deux segmentations du corpus myPOS. Les différences sont soulignées en jaune.

Dans la version *words*, les noms propres sont séparés le plus possible en composants. Par exemple, un titre de respect, comme ဆရာကြီး ⟨sʰrakǰï²⟩ /sʰa̰jàtɕí/ (*Professeur*) est séparé du nom personnel[53]. De la même manière, le nom d'une ville est séparé du token မြို့ ⟨mǰiu¹⟩ /mjo̰/ (*ville*). Les noms propres sont aussi séparés en composants dans la segmentation en tokens dans le corpus *words*, comme pour le *compound word All Burma Students' Union* ဗမာနိုင်ငံလုံးဆိုင်ရာကျောင်းသားသမဂ္ဂ ⟨bmaniuŋˣŋn̆lun̆²sʰiuŋˣrakɣ̆eaŋˣ²θa²θmg̊g̊⟩, séparé en six tokens : ဗမာ *birman* နိုင်ငံ *nation* လုံး *tout, entier* ဆိုင်ရာ *qui concerne* ကျောင်းသား *étudiant* သမဂ္ဂ *syndicat*.

Les syllabes qui font office de suffixes de dérivation sont également segmentées dans la version *words*. Les noms communs terminant en မှု ⟨mn̆u⟩ /mṵ/ (particule substantivante) sont presque systématiquement segmentés en composants dans la version plus segmentée *words*, mais réunis dans la segmentation

52. On peut observer toutes les différences à cette adresse : https://www.diffchecker.com/a5f7JeKM.

53. Ce terme pourrait lui-même être traité comme un *compound word*, mais dans les deux segmentations, il n'est pas toutefois divisé en ဆရာ ⟨sʰra⟩ /sʰa̰jà/ (*enseignant*) + ကြီး ⟨kǰï²⟩ /tɕí/ (*grand*). On ne peut pas donc considérer cette segmentation comme une segmentation en unités minimales de sens.

compound words. Ainsi, le token လှုပ်ရှားမှု ⟨lȟupˣrȟa²mȟu⟩ /l̥oʊʔʃámu̯/ (*mouvement*) du corpus *compound words* est segmenté en deux dans le corpus *words* : လှုပ်ရှား ⟨lȟupˣrȟa²⟩ /l̥oʊʔʃá/ (*déplacer*) + မှု ⟨mȟu⟩. Il y a une légère incohérence dans la segmentation de ဧရး ⟨re²⟩ /jé/ (*affaire*), qui constitue un token à part dans le corpus version *word* pour စီးပွား< >ဧရး< > ⟨sï²pw̆a²⟩⟨re²⟩ (*entreprise, prospérité* + *affaires*) et နိုင်ငံ< >ဧရး< > ⟨niuŋˣŋň⟩⟨re²⟩ (*nation* + *affaires*) (စီးပွားဧရး /sípwájé/ *économie* et နိုင်ငံဧရး /nàɪnŋànjé/ *politique* respectivement dans la version *compound word*), mais dans le cas de လူဝင်မှုကြီးကြပ်ရေး ⟨lùwŋˣmȟukj̃ï²kj̆pˣre²⟩ /lùwìnm̥u̯tɕítɕaʔjé/ (*immigration*) forme un seul token dans les deux corpus. Les deux façons de segmenter traitent les syllabes marquant le pluriel, telles que များ ⟨mў̆a²⟩ /mjá/ de la même manière : elles sont segmentées individuellement.

Le vocabulaire de chacune des versions segmentées du corpus *myPOS*, *words* et *compound words*, a été séparé en listes selon le nombre de syllabes afin de les comparer aux résultats de tests. Ensuite, nous avons segmenté tout le corpus en syllabes avec `sylbreak`[54] et nous avons généré les listes de n-grammes brutes avec la version de `list.pl` de Buerki (2017a), une liste pour chaque longueur de n-gramme (allant des bigrammes à deux syllabes, jusqu'aux n-grammes à neuf syllabes). La consolidation des n-grammes s'est faite avec l'algorithme de `Substring` Buerki (2017b) et le calcul du rapport de vraisemblance des bigrammes, trigrammes et tétragrammes avec `statistic.pl` de Text-NSP (Banerjee et Pedersen 2003).

Contrairement à Hla Hla Htay et Narayana Murthy (2008), nous n'avons pas trouvé dans les listes de n-grammes brutes une source fiable de tokens valables que l'on pourrait ajouter systématiquement au dictionnaire de *Motor*. Il se peut que les auteurs aient utilisé une liste d'exclusion plus extensive (au risque d'exclure des tokens valables), mais il semblerait qu'ils aient aussi eu recours à un tri manuel pour écarter les n-grammes qui ne représentent pas selon eux des mots valables. La comparaison entre leur expérience et la nôtre pâtit surtout du manque de caractères espaces dans le corpus que nous avons utilisé. Il est vrai que l'usage de n-grammes constitue une méthode plus rapide que celle à concordancier, mais selon nous le manque de fiabilité ne permet pas d'extraire des tokens de manière automatique. Ce problème de n-grammes non valables semble provenir principalement du fait qu'un grand nombre font partie d'un ou plusieurs

54. https://github.com/ye-kyaw-thu/sylbreak Consultée le 3 décembre 2018

n-grammes plus longs. Pour cette raison, la fréquence brute d'un n-gramme en elle seule n'est pas un indicateur fiable de la validité d'un token et l'étape de consolidation est très importante. Par exemple, le bigramme ဧဃ<>ရိုး<> ⟨me⟩⟨ri⟩, quarantième bigramme le plus fréquent du corpus *myPOS* avec une fréquence brute de 181, fait parti du n-gramme à quatre syllabes အ<>ဧဃ<>ရိုး<>ကန့်<> ⟨ʔ⟩⟨me⟩⟨ri⟩⟨kn^{x}⟩ (*États-Unis*, *américain*), qui a une fréquence de 126 occurrences dans le corpus. Le processus de consolidation des n-grammes réduit la fréquence de ဧဃ<>ရိုး<> ⟨me⟩⟨ri⟩ à zéro. Certains n-grammes qui représenteraient des tokens valables sont écartés par la liste d'exclusion, comme le cas du n-gramme အ<>များ<>စု<> ⟨ʔ⟩⟨mya^{ʔ}⟩⟨su⟩ (*majorité, plupart*) qui a pourtant une fréquence importante de 71 occurrences dans le corpus, qui se voit écarté, car il contient များ ⟨mya^{ʔ}⟩, une particule indiquant le pluriel qui se trouve dans la liste d'exclusion.

	word			***compound word***		
Longueur (syllabes)	**2**	**3**	**4**	**2**	**3**	**4**
Vocabulaire réf	5608	3705	2258	5084	5160	4053
Précision	0,452	0,185	0,131	0,416	0,268	0,211
Rappel	0,437	0,266	0.226	0,455	0,298	0,216
F-mesure	0,444	0,218	0,166	0,434	0,282	0,213

TAB. 3.15 : Comparaison entre n-grammes consolidés et vocabulaire du corpus *myPOS*

Même après le processus de consolidation, la comparaison entre listes de n-grammes et les listes de vocabulaire selon la segmentation s'avère décevante (voir le tableau 3.15) et il y a peu de différence entre l'efficacité de la méthode pour identifier les *words* et les *compound words*. La méthode est plus précise (moins de résultats incorrects) lors de la recherche de *words* à deux syllabes et les *compound words* à trois et quatre syllabes. La fraction du vocabulaire correctement identifiée est par contre plus élevée dans la recherche de *words* à quatre syllabes et les *compound words* à deux et trois syllabes. Même si la méthode est globalement peu convaincante, néanmoins on constate par les scores de F-mesure que la méthode est bien plus intéressante pour trouver les tokens valables à deux syllabes.

Nous n'avons comparé que les n-grammes et les tokens de deux à quatre syllabes, car en parcourant les listes de vocabulaire du corpus et de n-grammes plus longs, nous avons trouvé que bien peu peuvent se comprendre en les décomposant

en tokens de deux ou trois syllabes, ce qui serait utile si nous voulons réduire la taille du dictionnaire interne de *Motor*. A titre d'exemple, le n-gramme à neuf syllabes ဒု<>တိ<>ယ<>အင်္ဂ<>လိပ်<>မြန်<>မာ<>စစ်<>ပွဲ<> représente une chaîne de caractères qui apparaît six fois dans le corpus ဒုတိယအင်္ဂလိပ်မြန်မာစစ်ပွဲ ⟨dutiyʔŋˣglipˣmǰnˣmaśśˣpŭwɛ⟩ peut de décomposer en ဒု<>တိ<>ယ<> (*deuxième*) + အင်္ဂ<>လိပ်<> (*anglais*) + မြန်<>မာ<> (*birman*) + စစ်<>ပွဲ<> (*guerre*) (*La deuxième guerre anglo-birmane*). Cela vaut surtout pour les résultats après usage d'une liste d'exclusion. Sans liste d'exclusion, les listes de n-grammes longs sont ingérables et peu utiles ; par exemple la liste de n-grammes à sept syllabes a 258 725 éléments alors que le vocabulaire à sept syllabes du corpus séparé en *words* n'en a que 77. Il s'est donc avéré nécessaire de générer les n-grammes plusieurs fois en employant les listes d'exclusion multiples pour capter certains vocables. La syllabe ကို ⟨kiu⟩ (complément d'objet direct), par exemple, si elle est incluse dans une liste d'exclusion, écarte des n-grammes longs qui permettraient de bien segmenter certains noms propres tels que နီ**ကို**လပ်စ်ကော့ပါးနိုကပ်စ် ⟨nï**kiu**lpˣśˣkea¹päʔnikpˣśˣ⟩ (*Nicolas Copernic*) ou encore ချက်**ကို**ဆလိုဗားကီးယား ⟨kʰўkˣ**kiu**śʰliubaʔkïʔyaʔ⟩ (*La Tchécoslovaquie*).

En ce qui concerne les tests d'association, nous avons trouvé les listes d'exclusion indispensables pour générer du vocabulaire valable, à moins que la consolidation soit aussi utilisée en aval. Ce dernier procédé permet d'éliminer une bonne partie des n-grammes fréquents inutiles. Un exemple tiré du test de rapport de vraisemblance, le bigramme ဒု<>တိ<> ⟨du⟩⟨ti⟩ (rang 23, score de rapport de vraisemblance élevé 1649.736 et fréquence 152) a une fréquence nulle après un post-traitement de consolidation. Il s'agit en fait d'un morceau du trigramme ဒု<>တိ<>ယ<> ⟨du⟩⟨ti⟩⟨y⟩ (*deuxième* (qui seul a une fréquence de 48 après consolidation, mais fait partie d'autres n-grammes plus longs). Le calcul de scores de rapport de vraisemblance permet d'identifier des n-grammes qui forme des tokens valables peu fréquents, notamment les noms propres, par exemple လော့စ်<>အိန်<>ဂျယ်<>လီးစ်<> ⟨lea¹śˣ⟩⟨ʔinˣ⟩⟨gўyˣ⟩⟨lïʔśˣ⟩ (*Los Angeles*) qui n'apparaît qu'une fois dans le corpus *myPOs* avec un score relativement élevé de 67.7017. Toutefois, il nous semble que les listes résultant de ces calculs contiennent beaucoup de n-grammes inutiles. Les scores sont surtout intéressants pour comparer la pertinence des n-grammes entre eux.

3.3.8 *Observations*

Notre première observation principale est que quand en considère le lexique du birman en termes de lexies, il est (évidemment) beaucoup plus varié, plus vaste et plus intéressant que ne le suggèrent les dictionnaires existants.

Des lexies très fréquentes dans nos corpus ne sont pas répertoriées dans les dictionnaires, probablement comme nous l'avons déjà supposé en vertu de leur transparence morphologique. On peut citer comme exemples de ce type လူဝင်မှု /lùwìɴm̥ṵ/ *immigration* (composé de *personne*+*entrer*+PTC), မညီမျှမှု /mə̰ɲìmj̥ə̰m̥ṵ/ *inégalité*, မမေ့နိုင်စရာ /mə̰meɴàɪɴsə̰jà/ *inoubliable* et သူလျှိုမှု /θùʃòm̥ṵ/ *espionnage*, dérivé de သူလျှို /θùʃò/ *espion*. Certaines lexies que nous avons trouvées sont plus répandues que des synonymes se trouvant dans le MED, ainsi pour *heurt, attaque* တိုက်ခိုက်မှု /taɪʔkʰaɪʔm̥ṵ/ est beaucoup plus utilisé que အတိုက်အခိုက် /ʔə̰taɪʔʔə̰kʰaɪʔ/. D'autres lexies apparaissent plus fréquemment dans le corpus parce que certains sujets étaient d'actualité au moment de la création du corpus, tels que *les droits de l'homme* လူ့အခွင့်အရေး /lùʔə̰kwɪ̰ʔə̰jé/, မတရားမှု /mə̰tə̰jám̥ṵ/ *injustice* ou bien ဘက်မလိုက်မှု /bɛʔmə̰laɪʔm̥ṵ/ *neutralité*.

L'utilisation du vocabulaire anglais est assez surprenante, et donne parfois l'impression que n'importe quelle mot ou expression anglais est candidat légitime à l'inclusion dans le lexique birman. Le vocabulaire informatique est, sans surprise, très influencé par l'anglais, comme ကွန်ပျူတာ ⟨kʷǎnˣpʏ̌ùta⟩ /kʊ̀ɴpjùtà/ *computer*, ဝဘ်ဆိုဒ် ⟨wbˣsʰiudˣ⟩ /waʔsʰaɪʔ/ aussi écrit ဝက်(ဘ်)ဆိုက် ⟨wkˣ(bˣ)sʰiukˣ⟩ *website*, အွန်လိုင်း ⟨ʔʷǎnˣliuŋˣ²⟩ /ʔʊ̀ɴláɪɴ/ *online*, ဒစ်ဂျစ်တယ် ⟨dɨ́ˣgʏ̆ɨ́ˣtyˣ⟩ /dɪʔdʑɪʔtɛ̀/ *digital*, ဒေါင်းလုဒ် ⟨deäŋˣ²ludˣ⟩ /dáʊɴloʊʔ/ *download* et အပ်ဒိတ် ⟨ʔpˣditˣ⟩ /ʔaʔdeɪʔ/ *update*. Dans l'usage, ces deux derniers exemples ne semblent pas être employés en tant que verbes comme en anglais, mais empruntés comme des substantifs accompagnés du verbe လုပ် /loʊʔ/ *faire*. Un mot anglais, comme *blog* peut être combiné à un procédé de dérivation en birman, comme ဘလော့ဂ်ရေး /bə̰laʊʔjé/ *blogging*. Le vocabulaire simple anglais semble assez répandu et intégré (comme certains mots français), par exemple ချန်ပီယံ /tɕʰàɴpìjàɴ/ *champion*. Plus étonnant, ce sont des syntagmes ou mots composés quotidiens anglais qui pourraient très bien être rendus en birman, comme ဖိုင်နယ်ရီးယား ⟨pʰiuŋˣnyˣrï²ya²⟩ /pʰàɪɴnɛ̀jíjá/ *final year*. On constate la préférence pour une transcription du son bien plus que pour une translittération des lettres latines.

Nous n'avons pas vraiment trouvé de solution satisfaisante pour rendre compte

de la négation qui sépare les composants de verbes. Par exemple une forme négative de ကျေနပ် /tɕèna̰ʔ/ *être satisfait* peut être dérivée en préfixant la particule négative မ /mə̀/ à chaque syllabe, comme ceci : မကျေမနပ် /mə̀tɕèmə̀na̰ʔ/ *être mécontent*. Il nous semble qu'un étiquetage préalable, avec un système simple tel que celui du système NOVA, serait nécessaire pour les trouver de manière systématique.

Un aspect qui nous échappe, n'étant pas un locuteur qui maîtrise la langue, concerne les nuances de l'aspect stylistique. Nous soupçonnons un effet de traduction sur le vocabulaire de notre corpus, car, comme l'a fait remarquer Cunningham (2007), il y a une tendance dans les dictionnaires anglais-birman à utiliser des périphrases pour rendre des mots anglais, alors qu'il en existe un équivalent en birman. Elle donne l'exemple de *épidémie* qui est expliqué par *maladie transmissible* ကူးစက်တတ်သောရောဂါ /kúsɛʔtaʔθɔ́jɔ́gà/ au lieu de ကပ်ရောဂါ /kaʔjɔ́gà/. Il se peut que le fait que notre corpus contienne une quantité non négligeable de textes traduits (TED talks), ou potentiellement traduits (BBC) ait une influence sur le style des textes et le choix du vocabulaire.

L'orthographe des mots d'emprunt et d'entités nommés est un vrai sujet pour le traitement automatique du birman. Comme pour notre expérience avec le corpus *myPOS* ci-dessus, le calcul de scores de rapport de vraisemblance permettant d'identifier des n-grammes peu fréquents s'est avéré extrêmement utile pour identifier les variantes orthographiques. Fait intéressant, nous avons constaté dans notre corpus que l'orthographe pour (*Los Angeles*) le plus courant avec le score de rapport de vraisemblance le plus élevé (188.4544) n'est pas celui du corpus *myPOS*, mais လော့ < > အိန်း < > ဂျ < > လိစ် < > ⟨lea^{1}⟩⟨ʔin^{x2}⟩⟨gy̆⟩⟨lï^{2}s^{x}⟩. Nous n'avons pas été en mesure de décider quelle serait la meilleure façon de traiter ces variantes orthographiques. Faut-il normaliser les variantes ou préserver cette liberté orthographique ? La normalisation ou même les recommandations sur la manière de rendre les mots anglais et étrangers en birman nécessiteraient, à notre avis, l'analyse d'un corpus beaucoup plus vaste, plus équilibré et plus représentatif. La préservation de la variété orthographique nécessiterait un corpus de veille, une entreprise bien plus importante.

L'effet de confusion entre homographes est probablement moins problématique que l'on ne pense dans un corpus, surtout si on privilégie une segmentation en lexies qu'une segmentation en morphèmes. Comme l'a remarqué Wang et Nation

(2004), dans presque tous les cas, l'un des membres d'une paire ou d'un groupe d'homographes est beaucoup plus fréquent et répandu. Par exemple, si on prend le token ကာ ⟨k⟩ /kạ/, on ne peut pas distinguer entre le substantif *lettre ka* et le substantif ကာ ⟨k⟩ /kạ/ *selle de cheval*, mais on pourrait avec l'étiquetage distinguer s'il s'agit du verbe *danser* ou d'un marqueur syntaxique (appelé souvent postposition). Dans le corpus, le marqueur syntaxique est extrêmement fréquent et systématiquement suivi de l'espace typographique, et le verbe *danser* est plus répandu que le substantif *selle de cheval* ou la lettre de l'alphabet. Ceci dit, notre étude est bien sûr limitée par le fait de ne pas avoir eu recours à l'étiquetage morphosyntaxique, qui pourrait distinguer entre certains homographes. Une étude basée sur corpus d'homographes en birman, demanderait là aussi un corpus de taille plus importante, mais serait fort intéressante pour une meilleure évaluation qualitative du traitement automatique de la langue.

3.4 Résumé

Nous avons vu qu'il y a plusieurs facteurs importants à prendre en compte lors de la segmentation d'un corpus en birman destiné à être utilisé par des apprenants en langue. Certains de ces éléments sont pertinents pour tout travail sur corpus en birman (notamment les choix concernant la différence entre une erreur et une variante orthographique), mais d'autres sont plus spécifiques à l'apprentissage. Les limites de l'outil utilisé, les considérations culturelles telles que les croyances sur la langue et même la façon dont la langue est enseignée, ont toutes des conséquences sur l'utilisation et l'analyse futures du corpus.

La méthode de segmentation qui utilise un outil à dictionnaire interne présuppose donc qu'il y aura toujours un certain pourcentage d'erreurs, surtout en ce qui concerne les tokens les moins fréquents. Il y a toutefois un socle de vocabulaire dans une langue qui est relativement stable, avec une fréquence d'usage plus élevée et donc plus utile pour les apprenants. C'est cette partie du vocabulaire que nous voulons utiliser pour évaluer la difficulté textuelle pour les apprenants.

Deuxième partie

La lisibilité par la fréquence lexicale du birman langue étrangère

Chapitre 4

La fréquence lexicale

Dans le contexte d'apprentissage de langues étrangères, les données sur la fréquence lexicale contribuent surtout à l'identification du vocabulaire le plus « rentable » pour l'apprenant ; l'effort fourni pour apprendre les vocables que l'on rencontre fréquemment est davantage récompensé que pour ceux que l'on rencontre rarement. Ceci est crucial, car la quantité de vocabulaire à apprendre dans une langue étrangère présente une perspective intimidante et viser les compétences linguistiques d'un locuteur natif n'est pas toujours praticable ou réaliste. On estime, par exemple, qu'un locuteur natif éduqué anglophone possède un vocabulaire entre 16 000 et 20 000 vocables (Schmitt 2010). Un objectif plus réaliste pour un apprenant de langue étrangère est de connaître 95 à 98% du vocabulaire d'un texte, ce qui lui permet de deviner le vocabulaire restant. Les estimations quant à la taille du vocabulaire nécessaire pour atteindre ce but pour l'anglais varient entre 2000 et 3000 vocables pour couvrir 95% du discours parlé et entre 6000 et 7000 pour atteindre 98% (Schmitt 2010). En ce qui concerne la langue écrite, Nation (2006), se basant sur le British National Corpus (BNC), estime qu'un vocabulaire de 8000 à 9000 suffit pour comprendre 98% d'un texte écrit, mais la taille du vocabulaire nécessaire pourrait n'être que de 5000 si l'on se base sur le niveau C2 du Cadre européen commun de référence pour les langues (CECR) (Milton et Hopkins 2006). Encore plus encourageant, Nation et Meara (2010), citant Ringbom (1983), remarquent qu'en théorie ces chiffres pour l'anglais ne seraient pas forcement valables pour d'autres langues, du fait que les procédés morphologiques de création de vocabulaire sont moins productifs en anglais que dans d'autres langues, qui peuvent créer du vocabulaire nouveau à partir de vocables existants au moyen d'affixes ou de la composition. Un vocabulaire plus restreint, de 2000 à 3000 vocables, pourrait suffire.

Nation (2001) distingue quatre types de vocabulaire : le vocabulaire de haute

fréquence, le vocabulaire académique, le vocabulaire technique et le vocabulaire de basse fréquence. Selon lui, en anglais, le vocabulaire de haute fréquence couvre environ 80% d'un texte[1], et comprend non seulement les *mot-outils*, mais aussi des mots pleins, une liste de quelques 2000 vocables. Le vocabulaire académique anglais, identifié par Coxhead (2000), est commun aux textes académiques, peu importe le sujet du texte. Le vocabulaire technique, celui qui est propre au sujet en question d'un texte technique, ne couvre qu'environ 5% d'un texte technique. Enfin le vocabulaire de basse fréquence, lui aussi ne couvre qu'une faible partie d'un texte, environ 5%. Cette dernière catégorie est de loin la plus importante.

L'élaboration de listes de fréquence lexicale basée sur corpus a l'avantage de fournir un moyen de juger l'importance relative des vocables d'une manière empirique, sans s'appuyer sur les jugements subjectifs des locuteurs natifs. C'est un aspect non négligeable, car selon certains chercheurs (Alderson (2007), par exemple), même les capacités des locuteurs natifs éduqués à identifier correctement la fréquence lexicale relative ne seraient pas fiables. Cependant, ce constat ne fait pas l'unanimité, et n'est pas appuyé par les travaux de Schmitt et Dunham (1999) et Rogers et al. (2015). Il serait important donc de prendre aussi en compte l'intuition et les données générées par des études sur corpus.

4.1 L'état de l'art de la fréquence lexicale

L'intérêt pour la fréquence des mots dans les langues remonte au temps des Grecs anciens (DeRocher 1973). Au début du vingtième siècle, des chercheurs (essentiellement aux États-Unis) élaborèrent des listes de mots classés par ordre de fréquence conçues pour l'enseignement de l'anglais en tant que L1, le plus connu étant Thorndike (1921) et Thorndike et Lorge (1944)[2], mais aussi pour l'enseignement d'autres langues étrangères, comme le français (Henmon 1924 ; Vander Beke 1932), l'allemand (Engel 1931), l'espagnol (Buchanan 1927 ; Keniston 1941) et le portugais brésilien (Brown et al. 1945).

L'intérêt précoce pour la fréquence lexicale ne se limita pas aux chercheurs occidentaux. Pendant les années 1920, une équipe dirigée par Chen (1922) créa un corpus d'environ un demi-million de sinogrammes composé de six catégories

1. Le pourcentage de couverture d'un texte est le pourcentage de couverture de toutes les occurrences du vocabulaire, et non pas un pourcentage du vocabulaire.
2. Un grand nombre de ces études précurseurs est listé par Dale et Reichert (1957).

de textes afin de créer une liste de fréquence des sinogrammes à visée didactique, présentée par ordre des radicaux (l'équivalent de l'ordre alphabétique pour le français) en plus de l'ordre de fréquence (Xu 2015). En 1931, le chercheur bengalais Deb Chaudhury a publié une étude sur la fréquence lexicale de la variété littéraire du bengali (*Sadhu Bhasha*), afin d'améliorer les manuels de lecture pour des enfants (Dash 2007). Cette étude a été suivie par d'autres sur le bengali (Roy et Roy 1946 ; Bhattacharya 1965) et plus récemment par Mallick et al. (1998), focalisées elles aussi sur la langue littéraire, mais selon Dash (2007) présentant des anomalies sur le plan de représentativité. Ceci démontre la faiblesse de ces premières études, élaborées souvent à partir de corpus de textes écrits[3], considérés de taille importante avant l'avènement de l'informatique, mais qui manquaient de représentativité. En outre, ces listes ne tenaient pas toujours compte d'autres facteurs que la fréquence, tels que la dispersion ou la polysémie.

La *General Service List* de West (1953), une liste des mots les plus fréquents de l'anglais, basée sur une analyse plus rigoureuse, est devenue la liste de mots de référence pour l'anglais jusqu'à la publication de la *New General Service List* (Browne et al. 2013)[4]. Des listes analogues de mots de haute fréquence ont été élaborées pour le français par Haygood (1951) et l'italien par Bortolini et al. (1971). Plus récemment, la maison d'édition Routledge a publié une série de dictionnaires fréquentiels destinée aux apprenants de langues étrangères. À l'exception du dictionnaire de fréquence allemand (4034 mots), chaque dictionnaire représente les 5000 mots (lemmes) les plus fréquents de chaque langue[5]. L'élaboration de ces dictionnaires est faite à partir de corpus de grande taille, chaque chercheur étant soucieux de travailler à partir d'un corpus équilibré représentatif de la langue dans sa totalité qui comporte des textes oraux et écrits de sources variées et représentatives de la langue contemporaine.

Au-delà des mots de haute fréquence, des chercheurs ont élaboré des listes de mots fréquents dans des textes académiques en anglais, mais peu fréquents dans un corpus représentatif de la langue en général. Par exemple, la liste académique

3. À l'exception de la liste anglaise de Dewey (1923) dont le titre *Relative Frequency Of English Speech Sounds* laisse entendre que sa liste représenterait la fréquence lexicale de la langue orale.

4. http://www.newgeneralservicelist.org/

5. Les langues concernées sont l'anglais américain contemporain (Davies et Gardner 2013), l'allemand (Jones et Tschirner 2015), l'arabe (Buckwalter et Parkinson 2014), le chinois mandarin (Xiao et al. 2015), l'espagnol (Davies 2005), le français (Lonsdale et Le Bras 2009), le japonais (Tono et al. 2013), le néerlandais (Tiberius et Schoonheim 2013), le portugais (Davies et Raposo Preto-Bay 2007), le russe (Sharoff et al. 2014) et le tchèque (Čermák et Křen 2011).

en anglais élaborée par Coxhead (2000)[6] mentionnée ci-dessus couvre 8,5% du vocabulaire d'un corpus académique (Nation 2001). Des listes analogues ont été développées pour le suédois, le norvégien et le danois (Johansson Kokkinakis et al. 2012) et le portugais (Baptista et al. 2010), souvent à partir de la liste pour l'anglais ou s'en inspirant. Toutefois, il n'est pas certain que les textes académiques de toutes les langues aient le même profil de fréquence lexicale. Cobb et Horst (2004) argumentent contre une liste académique pour le français. Des tentatives non publiées, mais disponibles en ligne, ont été faites pour établir des listes de fréquence lexicale académique pour le thaï[7] basées aussi sur les travaux de Coxhead.

4.2 Prétraitements spécifiques

Certains prétraitements spécifiques à la création de listes de fréquence lexicale à visée didactique sont souhaitables voire nécessaires avant leur création. Puisque nous ne nous intéressons qu'à la fréquence du birman, nous ne retiendrons que les tokens qui contiennent des caractères birmans. Cela écarte le vocabulaire en lettres latines, qui est assez fréquent en texte birman, surtout quand il s'agit de noms propres, ainsi que les chiffres arabes. Toute la ponctuation est aussi éliminée qu'il s'agisse de signes de ponctuation birmane ou autres, à l'exception des parenthèses à l'intérieur du vocabulaire birman[8]. Les nombres en chiffres birmans, qui ont été segmentés en chiffres isolés par l'outil de segmentation (de ၀ ⟨0⟩ à ၉ ⟨9⟩) sont aussi éliminés. Nous avons choisi de ne pas inclure la marque du ton aukmyit, ့ ⟨ˀ⟩, (U+1037 MYANMAR SIGN DOT BELOW), séparée lors du traitement de la segmentation quand il s'agit d'un usage syntaxique, car cela concerne des connaissances syntaxiques et non pas lexicales.

4.3 L'élaboration de listes de fréquence lexicale

Au chapitre précédent, nous nous sommes concentrés sur la définition des tokens, et les enjeux concernant la correspondance entre token et lexie. Dans ce chapitre, nous supposons que le token représente une unité linguistique valable

6. Il s'agit de la AWL ou *Academic Word List*, composée de 570 familles de mots.
7. Voir http://www.sealang.net/thai/vocabulary/awl-2.htm.
8. Pour une explication de l'usage des parenthèses à l'intérieur des vocables en birman, voir la section 2.5.1.

et significative, et il est traité comme une chaîne de caractères uniquement. Le corpus est ainsi caractérisé par des calculs statistiques effectués sur les occurrences de ces tokens, qui forment les unités de base du corpus. La taille d'un corpus (N) est exprimée en nombre de tokens, alors que la taille du vocabulaire du corpus (V) exprime le nombre de tokens distincts, les *types*. Prenons un petit échantillon en français comme exemple :

> quand les étudiants nous viennent de l'étranger, nous ne leur accordons pas assez de crédit pour ce qu'ils savent, et ils le savent dans leur propre langue. Quand une langue meurt, nous ne savons pas ce que nous perdons avec cette langue.
>
> Patricia Ryan, *Don't insist on English !*, TED Talks

Pour le français, à la segmentation en tokens proprement dite s'ajoute leur lemmatisation[9], ici avec l'outil TreeTagger (Schmid 1994) configuré pour le français. Le corpus est transformé ainsi :

> quand le étudiant nous venir de le étranger , nous ne lui accorder pas assez de crédit pour ce que il savoir , et il le savoir dans leur propre langue . quand un langue mourir , nous ne savoir pas ce que nous perdre avec ce langue .

Quand la ponctuation est exclue, la taille du corpus N est de 43 tokens et la taille du vocabulaire V est de 27 types. Nous pouvons ranger les types par ordre de fréquence f dans un tableau (4.1) où les types sont numérotés du plus fréquent au moins fréquent, en attribuant le rang 1 au type le plus fréquent, le rang 2 au deuxième type le plus fréquent et ainsi de suite. On constate que le dernier rang, 27, correspond à V, la taille du vocabulaire. Les rangs de types de même fréquence sont arbitraires, on peut choisir aussi d'attribuer le même rang aux types avec la même fréquence (ici, les types avec la fréquence 3 pourraient tous avoir le rang 2 ou 5). Même en utilisant un petit échantillon on constate que les plus basses fréquences sont surreprésentées.

9. La lemmatisation transforme les tokens fléchis de même sens et même partie de discours par une forme unique de référence, appelée lemme. En français, on emploie l'infinitif pour les verbes et le masculin singulier pour les noms, les adjectifs et les articles.

Rang	***type***	***f***	**Rang**	***type***	***f***
1	nous	4	15	crédit	1
2	ce	3	16	dans	1
3	langue	3	17	et	1
4	le	3	18	étranger	1
5	savoir	3	19	étudiant	1
6	de	2	20	leur	1
7	il	2	21	lui	1
8	ne	2	22	mourir	1
9	pas	2	23	perdre	1
10	quand	2	24	pour	1
11	que	2	25	propre	1
12	accorder	1	26	un	1
13	assez	1	27	venir	1
14	avec	1			

TAB. 4.1 : Exemple de liste de fréquence

La distribution de fréquences de notre petit exemple (tableau 4.2) illustre le nombre de types *V(f)* qui possède une fréquence *f* donnée. 16 types sont des hapax, c'est-à-dire qu'ils n'apparaissent qu'une seule fois (fréquence 1), et seulement un type possède la fréquence 4. Cette configuration est appelée aussi un *spectre de fréquences* (Baroni 2008 ; Baayen 2001).

f	***V(f)***
1	16
2	6
3	4
4	1

TAB. 4.2 : Exemple de spectre de fréquences

Si les proportions exactes de fréquences de notre exemple ne sont bien sûr pas représentatives de la langue française et des langues en général, cette tendance, d'un grand nombre d'hapax dans le lexique d'une langue, est générale. La

majorité des tokens d'un texte est composée de types fréquents, mais l'inventaire des types est composé en majorité de types peu fréquents. Ceci est exemplifié par nos propres corpus. Le tableau 4.3 illustre ce phénomène pour notre corpus birman de textes littéraires. Sur les 758 553 tokens que composent les textes du corpus, les dix types les plus fréquents à eux seuls (131 346 tokens) font 17% du corpus, alors que les dix derniers rangs sont occupés par 12 807 types, seulement 37 719 tokens soient 5% du corpus.

Fréquences supérieures			**Fréquences inférieures**		
Rang	***f***	***type***	**Fourchette de rangs**	***f***	**Exemples de types**
1	20 721	ကို	4221–4472	10	အဖိုး ရေခဲသေတ္တာ ပိုးဟပ်
2	17 093	မ	4473–4801	9	ကဖီး ထမင်းကြော် ရှေးဟောင်း
3	16 423	က	4802–5190	8	ကော်မတီ စာအုပ်စင် မျက်လှည့်ဆရာ
4	14 073	တယ်	5191–5654	7	စီးပွားရေး ဆက်ဆံမှု ဇော်ဂျီ
5	12 769	သည်	5655–6256	6	ကာရာအိုကေ အရွှေးအမှား ငြိမ်သက်မှု
6	12 532	နေ	6257–7024	5	အားလပ်ချိန် လွတ်လပ်ရေး မြို့သူမြို့သား
7	10 443	တွေ	7025–8063	4	ကစားကွင်း ကားမှတ်တိုင် ဓာတ်ပုံဆရာ
8	9399	မှာ	8064–9507	3	ကယ်ဆယ်ရေ ဇန်နဝါရီလ ပန်းဆီ
9	8988	တဲ့	9508–11 924	2	ဦးဦးဖျားဖျား သဘောတူညီမှု ဓာတ်ပုံဆရာ
10	8905	တာ	11 925–17 028	1	ကန်ထရိုက်တာ စစ်သူရဲ အော့ကြောလန်

TAB. 4.3 : Fréquences supérieures et inférieurs de la liste de fréquence du corpus littéraire Lotaya de taille N= 758 553 tokens

Cette distribution de types dans la langue fut remarquée pour la première fois par Jean-Baptiste Estoupe en 1916 (Sorell 2014), mais c'est Zipf (1949) qui a commenté de manière extensive la distribution de fréquences de mots. Il a remarqué que la relation entre fréquence et rang est (à peu près) constante. C'est ce qu'on appelle la loi de Zipf. La courbe rang-fréquence 4.1 de rangs et fréquences du corpus *Lotaya* illustre cette relation. En haut, les premiers rangs de haute fréquence sont entassés près de l'axe vertical à gauche, et au fil que l'on monte dans les rangs, rapidement on tombe dans les tokens de basse fréquence. Les mêmes données sont plus faciles à interpréter tracées à l'échelle logarithmique (base dix) dans le graphique du dessous : la courbe, assez droite surtout au milieu, est typique d'une courbe zipfienne. En réalité, la relation n'est pas exactement constante, car le début et la fin de la courbe

sont légèrement inclinés. Ceci est prévu par la modification à la loi de Zipf, le modèle Zipf-Mandelbrot (Sorell 2014).

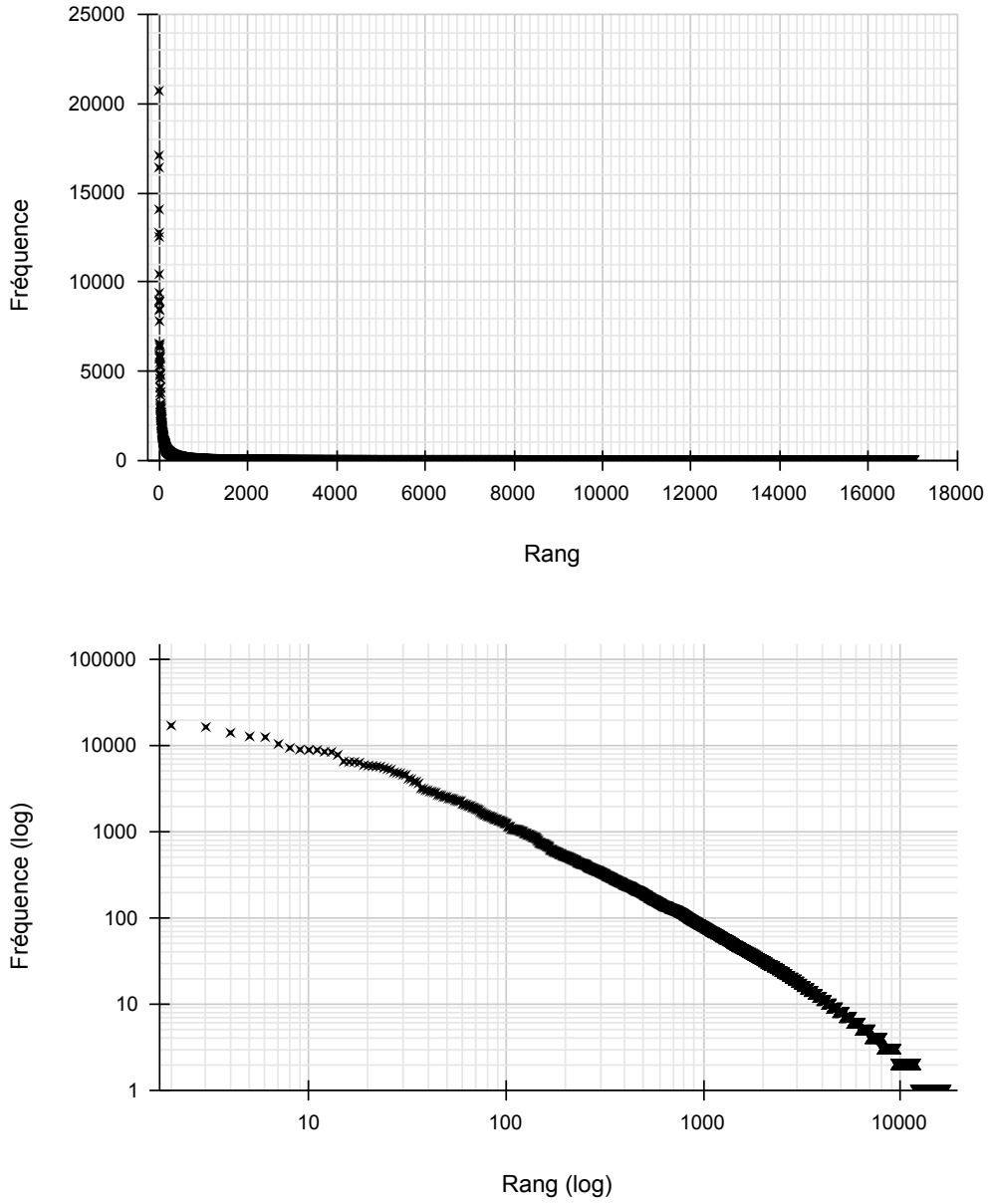


FIG. 4.1 : Représentations non logarithmique et logarithmique de la courbe zipfienne du spectre de fréquences du corpus littéraire *Lotaya*

4.3.1 *Fréquence absolue et fréquence relative*

Jusqu'alors nous n'avons parlé que de la fréquence absolue d'un type, sans prendre en compte la taille du corpus. Il s'agit du nombre d'occurrences d'un type dans un corpus donné. Si l'on souhaite comparer le même type dans deux corpus différents, les valeurs de fréquence ne sont pas comparables quand les corpus n'ont pas la même taille. Le type အချိန် /ʔa̰tɕʰèɪɴ/ (*le temps*) a 478 occurrences dans le corpus de presse BBC, mais dans le corpus de sous-titres TED, il apparaît 1284 fois. Il est apparemment plus de deux fois et demie plus fréquent dans l'un que dans l'autre. Si toutefois la fréquence est pondérée par la taille du corpus par un calcul de fréquence relative (FR), la différence est moindre.

$$\text{fréquence relative} = \frac{\text{fréquence absolue}}{\text{taille du corpus}}$$

$$\text{FR(t)} = \frac{\text{f}(t)}{\text{N}}$$

où t (comme « type ») désigne la fréquence du type qui nous intéresse dans le corpus.

La FR de အချိန် (t = 478) dans le corpus BBC serait

$$\text{FR} = \frac{478}{518179} = 0,000922461 \tag{4.1}$$

La FR de အချိန် (t = 1284) dans le corpus TED serait

$$\text{FR} = \frac{1284}{1014744} = 0,001265344 \tag{4.2}$$

La FR de အချိန် (t = 736) dans le corpus Lotaya serait

$$\text{FR} = \frac{736}{756132} = 0,000973375 \tag{4.3}$$

Pour que le résultat soit plus lisible et intuitif, on applique une normalisation. Pour les corpus de petite taille on peut utiliser une base de normalisation de mille, pour un corpus de plus d'un million de tokens, on utilise 1 million comme base de normalisation (McEnery et Hardie 2012).

$$\text{fréquence relative normalisée} = \frac{\text{fréquence absolue}}{\text{taille du corpus}} \times \text{base de la normalisation}$$

$$\text{FRN}(t) = \frac{\text{f}(t)}{\text{N}} \times \text{base de la normalisation}$$

La FRN de အချိန် dans le corpus BBC serait

$$\begin{aligned} \text{FR} &= \frac{478}{518179} \times 1000000 \\ &= 922 \end{aligned}$$

Lorsque la base de normalisation est plus élevée que la taille du corpus, ceci pose problème, car toutes les valeurs sont grossies. La base de mille nous donne les valeurs autour de un : 0,922461, 1,265344 et 0,973375 occurrences pour mille tokens (pour les corpus BBC, TED et Lotoya respectivement), un résultat qui n'est pas très intuitif.

Nous avons choisi donc 500 000 comme base de normalisation, le chiffre rond le plus élevé en dessous de la taille de notre plus petit corpus. Cela est suffisant pour comparer les types de nos corpus authentiques, mais il faut garder à l'esprit qu'il ne s'agit pas d'une base de normalisation habituelle en linguistique de corpus quand on compare nos résultats à d'autres corpus.

La FRN de အချိန် dans le corpus BBC serait donc

$$\begin{aligned} \text{FRN} &= \frac{478}{518179} \text{ x } 500000 \\ &= 461 \end{aligned}$$

Nous pouvons maintenant comparer les fréquences relatives d'un même type entre les différents corpus.

အချိန် dans le corpus de sous-titres TED :

$$\begin{aligned} \text{FRN} &= \frac{1284}{1014744} \times 500000 \\ &= 633 \end{aligned}$$

အခြိန် dans le corpus littéraire Lotaya :

$$\begin{aligned} \text{FRN} &= \frac{736}{756132} \times 500000 \\ &= 486 \end{aligned}$$

On constate que les fréquences relatives sont plus similaires que les fréquences absolues et en fait les valeurs de FR des corpus de presse et corpus littéraire sont assez similaires.

4.3.2 *La dispersion*

La dispersion permet de juger plus finement la fréquence d'usage du vocabulaire ; deux types peuvent avoir la même fréquence relative, mais une répartition très différente. Un type peut apparaître un même nombre de fois dans une seule partie du corpus (dans un échantillon de langue spécialisée par exemple), alors que d'autres apparaissent dans toutes les parties du corpus (Gries et Newman 2014). Les proportions de genres textuels inégales contribuent aussi à des fréquences biaisées si l'on ne considère que les fréquences absolues. Différentes mesures de dispersion fournissent des informations de nature différente sur l'homogénéité des occurrences du vocabulaire d'un corpus. Nous passerons en revue différentes mesures afin d'expliquer comment elles peuvent servir dans l'élaboration de listes de fréquence lexicales, notamment l'élaboration d'une liste générale à partir de corpus de taille et de genres textuels différents.

4.3.2.1 L'étendue

L'*étendue*$_2$[10] est une mesure très élémentaire et grossière de la dispersion d'un type dans un corpus. Il s'agit du nombre de parties du corpus qui contient le type étudié. Bien sûr, cela veut dire que la mesure est dépendante de la structure du corpus. Elle est souvent exprimée en pourcentage (Brezina 2018).

$$\text{étendue}_2\% = \frac{\text{nombre de parties du corpus contenant le token}}{\text{nombre total de parties du corpus}} \times 100$$

10. Ici le sens du terme *étendue* n'est pas le même de celui généralement utilisé en statistique, la différence entre la valeur la plus grande et la valeur la plus petite d'une série.

Par exemple, dans le corpus littéraire Lotaya, de 295 textes, le token le plus fréquent est la postposition ကို /kò/ avec 20 721 occurrences. Ce token apparaît dans tous les textes, ainsi

$$\text{étendue}_2\% \text{ de } \text{ကို} = \frac{295}{295} \times 100 = 100\%$$

Moins fréquent, ထမင်းကြော် /tʰəmínʁtɕɔ̀/ (*riz sauté*), avec seulement neuf occurrences, n'apparaît que dans cinq textes, donc dans seulement 1,7% des parties du corpus :

$$\text{étendue}_2\% \text{ de } \text{ထမင်းကြော်} = \frac{5}{295} \times 100 = 1,7\%$$

Cette mesure a plus d'utilité quand le corpus est divisé en un grand nombre de parties, mais elle ne prend pas en compte la taille des parties, on sait seulement si un type est présent ou non dans chaque partie. Elle donne une idée de dispersion, mais doit s'accompagner d'autres mesures plus précises.

4.3.2.2 L'écart-type et le coefficient de variation

Pour préciser l'homogénéité de la dispersion d'un type dans le corpus, on peut utiliser la distance des valeurs de fréquence dans chaque partie par rapport à leurs moyennes sur toutes les parties (Brezina 2018). Il s'agit de l'*écart-type de population*.

$$\text{écart-type}_{population} = \sqrt{\frac{\text{somme des distances au carré de la moyenne}}{\text{nombre total de parties du corpus}}}$$

L'écart-type évalue ici l'uniformité de répartition du token dans les diverses parties : s'il est nul, la répartition est uniforme, s'il est important, il sera très spécifique à certaines parties. Prenons l'exemple de ဘူတာရုံ /bùtàjòʊɴ/ (*gare*), qui apparaît quinze fois dans huit des 295 textes du corpus Lotaya. Pour les besoins de la démonstration, notre corpus n'a que huit textes, présentés dans le tableau 4.4. Nous nous arrondissons les valeurs des tailles de parties au millier le plus proche. Le token qui nous intéresse, ဘူတာရုံ /bùtàjòʊɴ/, sera de nouveau désigné par *t* (comme « type »).

	Partie 1	**Partie 2**	**Partie 3**	**Partie 4**	**Partie 5**	**Partie 6**	**Partie 7**	**Partie 8**	**Corpus entier**
Tokens	4000	2000	2000	4000	5000	3000	2000	1000	23000
f(t) absolue	2	4	1	2	3	2	1	1	16
f(t) relative par 1000	0,5	2	0,5	0,5	0,4	0,7	0,5	1	0,7
distance de la moyenne	−0,3	1,2	−0,3	−0,3	−0,2	−0,1	−0,3	0,2	

TAB. 4.4 : Fréquences de t dans un échantillon de corpus (d'après Brezina (2018))

La moyenne des fréquences relatives de m est $0,8$.

$$\begin{aligned}\text{moyenne}(t) &= \frac{0,5+2+0,5+0,5+0,4+0,7+0,5+1}{8}\\ &= 0,8\end{aligned}$$

On obtient la distance de la moyenne pour chaque partie en soustrayant la moyenne de la fréquence relative, les valeurs de la dernière ligne du tableau 4.4.

Le calcul de l'écart-type, σ, pour t serait donc

$$\begin{aligned}\sigma(t) &= \sqrt{\frac{(-0,3)^2+(1,2)^2+(-0,3)^2+(-0,3)^2+(-0,2)^2+(-0,1)^2+(-0,3)^2+(0,2)^2}{8}}\\ &= 0,24\end{aligned}$$

Cette valeur, par rapport à la moyenne $0,8$, n'est pas très importante. Une valeur plus proche de la moyenne signifierait plus de variation. L'écart-type n'est utile que pour comparer la dispersion de types de même fréquence, car l'écart-type doit être considéré par rapport à la moyenne (Brezina 2018).

Afin de pouvoir comparer la dispersion du vocabulaire qui possède des moyennes de fréquences différentes, on calcule le *coefficient de variation*, *CV*, ou écart-type relatif. Le *CV* est le quotient de l'écart-type par la moyenne, donc une valeur indépendante de la moyenne. Il a l'avantage d'être indépendant de l'unité de mesure (Muller 1973). Plus la valeur du coefficient de variation s'approche

de zéro, plus la dispersion du type est régulière.

$$\begin{aligned} \text{CV}(t) &= \frac{\sigma(t)}{\text{moyenne}(t)} \\ &= \frac{0,24}{0,8} \\ &= 0,3 \end{aligned}$$

Brezina (2018) explique que le *CV* est souvent simplement multiplié par 100 et présenté comme un pourcentage. Cela pose un problème lorsque l'écart-type est supérieur à la moyenne, ce qui rend le *CV* supérieur à un. Il propose de diviser le *CV* d'abord par sa valeur maximale, qui dépend du nombre de parties du corpus. Cette valeur maximale du coefficient de variation est égale à la racine carrée du nombre de parties moins un. Voici le calcul pour notre exemple :

$$\begin{aligned} \text{CV\%}(t) &= \frac{\text{CV}(t)}{\sqrt{\text{nombre total de parties du corpus} - 1}} \times 100 \\ &= \frac{0,3}{\sqrt{8-1}} \times 100 \\ &= 11,3\% \end{aligned}$$

La valeur maximale serait atteinte si le type n'apparaissait que dans une partie du corpus. Le type en question, ဘူတာရုံ /bùtàjòʊɴ/ (*gare*) est réparti de manière assez homogène dans notre corpus exemple, car son coefficient de variation n'est que 11,3%.

4.3.2.3 Le coefficient D de Juilland

Le coefficient D de Juilland et al. (1970) permet de comparer la dispersion du vocabulaire avec des moyennes de fréquences différentes. Il fut développé pour la création de dictionnaires de fréquence et repose sur l'écart-type de fréquences du vocabulaire dans des parties du corpus (Burch et al. 2017).

$$\text{D} = 1 - \frac{\text{coefficient de variation}}{\sqrt{\text{nombre total de parties du corpus} - 1}}$$

Une valeur de D proche de 1 est significative d'une distribution homogène, alors qu'une valeur proche de zéro désigne une distribution très inégale. Selon notre exemple avec huit parties, le D de Juilland se calcule :

$$\begin{aligned} \mathrm{D} &= 1 - \frac{0,3}{\sqrt{8-1}} \\ &= 0,89 \end{aligned}$$

Si nous recalculons avec le nombre réel de parties de notre corpus, 295, il y a 287 textes où ဘူတာရုံ /bùtàjòʊɴ/ (*gare*), n'apparaît pas du tout. La moyenne de fréquences relatives serait donc $0,02$, ce qui nous donne un coefficient D de Juilland proche de 1. Ceci illustre les critiques de Biber et al. (2016) sur la dépendance du D de Juilland sur le nombre de parties du corpus. Plus le nombre de parties est important, plus le coefficient est élevé.

Ce coefficient est largement utilisé lors de la création de dictionnaires de fréquence, notamment pour le russe (Sharoff et al. 2014), le turc (Aksan et al. 2017) et le chinois mandarin (Xiao et al. 2015), où il est précisé pour chaque entrée en même temps que la fréquence normalisée. Xiao et al. (2015) fournit également une mesure qui indique le taux d'utilisation, *U*, qui combine le D de Juilland et la fréquence.

$$\mathrm{U} = f \times \frac{\mathrm{D}}{100}$$

4.3.2.4 La déviation de proportions

La déviation de proportions ou *DP* de Gries (2008) (utilisé par Lonsdale et Le Bras (2009) dans la création de *A Frequency Dictionary of French*) est aussi exprimé comme un chiffre entre 0 et 1, zéro signifiant une distribution parfaitement homogène et 1 une distribution très hétérogène[11]. Ce coefficient repose aussi sur la fréquence du vocabulaire dans les parties du corpus, comme le D de Juilland, mais cette fois en exploitant les différences absolues des proportions de fréquences observées et attendues dans les parties. Les valeurs attendues sont calculées à partir des tailles de parties rapportées à la taille du

11. C'est-à-dire l'inverse du D de Juilland.

corpus.

$$\text{DP} = \frac{\text{somme des valeurs (observées} - \text{attendues)}}{2}$$

Appliqué à notre petit exemple, le calcul est expliqué dans le tableau 4.5. Après division du dénominateur $(0,41)$ par deux, DP = $0,20$, ce qui est proche de zéro est donc signifie une dispersion homogène, en accord avec le D de Julliard.

	Partie 1	**Partie 2**	**Partie 3**	**Partie 4**	**Partie 5**	**Partie 6**	**Partie 7**	**Partie 8**	**Corpus entier**
Tokens	4000	2000	2000	4000	5000	3000	2000	1000	23000
f(t) absolue	2	4	1	2	2	2	1	1	15
proportion attendue	$\frac{4000}{23000}$	$\frac{2000}{23000}$	$\frac{2000}{23000}$	$\frac{4000}{23000}$	$\frac{5000}{23000}$	$\frac{3000}{23000}$	$\frac{2000}{23000}$	$\frac{1000}{23000}$	1
proportion observée	$\frac{2}{15}$	$\frac{4}{15}$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	1
différences absolues	$0,04$	$0,18$	$0,02$	$0,04$	$0,08$	$0,0$	$0,02$	$0,02$	$0,41$

TAB. 4.5 : Calcul de la déviation de proportions (d'après Brezina (2018))

4.3.3 *La fréquence et la dispersion combinées*

Ces calculs de dispersion d'un type fournissent des mesures informatives sur son utilité, mais n'ont pas d'incidence sur les rangs de fréquence. Idéalement, la dispersion devrait être incluse dans l'élaboration de la liste même, afin d'assurer l'utilité des vocables non seulement d'un point de vue de la fréquence, mais aussi de la dispersion. La fréquence réduite moyenne permet de combiner ces deux types d'informations. La fréquence d'un type qui n'est pas très fréquent dans certaines parties du corpus sera ainsi réduite, favorisant donc les types qui ont une grande dispersion.

4.3.3.1 La fréquence réduite moyenne

L'inconvénient de ses deux mesures, le coefficient D de Juilland et la déviation de proportions est que l'on ne peut pas facilement comparer les valeurs d'un corpus à un autre, car elles sont dépendantes de la structure du corpus. Lorsqu'il est difficile de collecter une grande quantité de textes de genres textuels variés

en proportions égales, ce qui est souvent le cas pour les langues peu dotées, ces méthodes de calcul de la dispersion peuvent poser problème et on doit tronquer certains textes pour s'assurer d'un corpus équilibré, réduisant la taille du corpus. Pour cette raison, nous pensons que la méthode de Savický et Hlavácová (2002), la fréquence réduite moyenne (FRM), qui corrige les fréquences selon la dispersion des occurrences dans le corpus entier, sans références aux parties du corpus, est très intéressante pour élaborer les listes de fréquence pour les langues peu dotées.

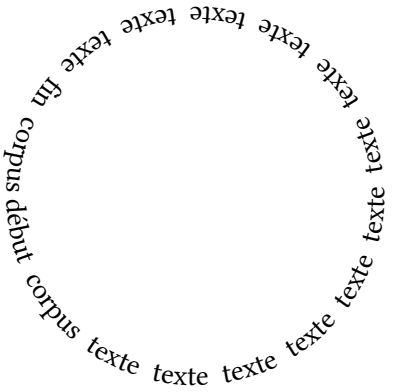


FIG. 4.2 : Conceptualisation de texte pour calcul de FRM

La fréquence réduite moyenne est obtenue en corrigeant les valeurs de fréquence absolue selon la dispersion des types dans le corpus entier. Si la dispersion est homogène, la fréquence corrigée est proche de la fréquence absolue et inversement. Afin de traiter le corpus, on l'imagine comme une seule ligne de texte déployée sur toute la circonférence d'un cercle, obtenue par la concaténation de tous les textes du corpus, le dernier token du dernier texte placé devant le premier token du premier texte (voir figure 4.2).

Pour un type, le flot de textes est divisé en autant de parties que sa fréquence absolue : sa fréquence réduite est alors le nombre de parties qui le contiennent. De cette façon, on ignore les occurrences qui sont proches les unes des autres.

La FRM est calculée à partir des distances entre les occurrences. Chaque occurrence est numérotée selon sa position, *p*. La taille du corpus en tokens, *N*, est divisée par la fréquence du type, *f(t)*, pour obtenir la dispersion moyenne. Si la distance entre occurrences adjacentes est plus grande que la dispersion moyenne, c'est cette dernière qui est retenue. Brezina (2018) l'illustre avec un

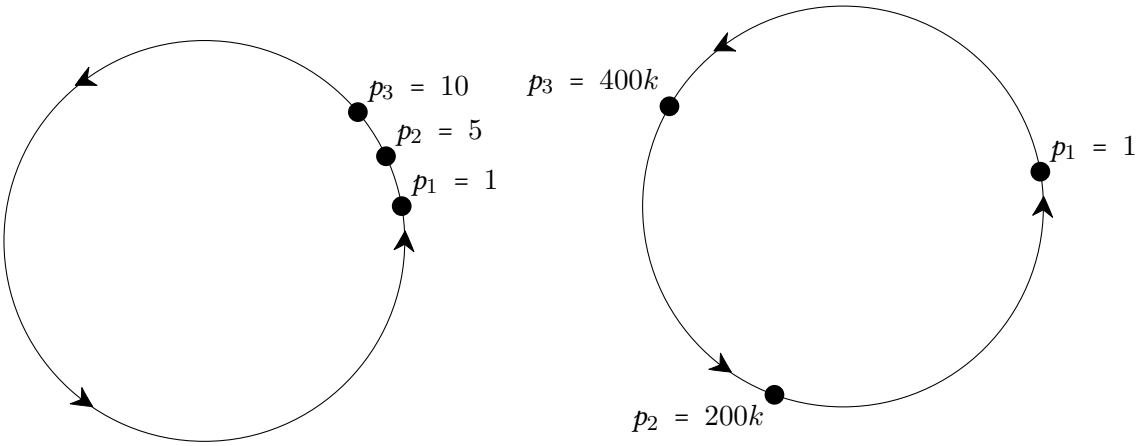


FIG. 4.3 : Comparaison schématique de deux calculs de la fréquence réduite moyenne pour des types avec dispersions différentes dans un corpus. On conceptualise le texte d'un corpus sur une seule ligne en forme de cercle avec les positions du type dans le corpus, p, indiquées. A gauche, un type, t, peu dispersé dans le corpus, à droite un type très dispersé. Afin de calculer la dispersion, le corpus est divisé en parties égales, le nombre de parties étant égale à la fréquence du type $f(t) = 3$. Le début de chaque partie est indiqué par une flèche.

schéma similaire à celui de la figure 4.3. Dans les deux cas de figure, le corpus a une taille de $N = 600\,000$ et il est divisé en parties égales par la fréquence du type, $f(t) = 3$, afin de déterminer la longueur théorique des parties, v, qui équivaut la longueur des parties entre types distribués de façon parfaitement régulière, $v = 200\,000$.

$$\text{dispersion moyenne(t)} = \frac{\text{taille du corpus, } N}{\text{fréquence absolue de t}}$$

$$\begin{aligned} v(t) &= \frac{N}{f(t)} \\ &= \frac{600000}{3} \\ &= 200000 \end{aligned}$$

A gauche, les positions, p, très proches, se trouvent dans la même partie, et les positions ne comptent que pour une seule. A droite, les positions, p, sont dispersées de façon parfaite, et les trois sont incluses dans le calcul de la

fréquence.

$$\text{FRM} = \frac{1}{v} \times (\min(\text{distance}_1,\ v\ ) + \min(\text{distance}_2,\ v\ ) + ...\ )$$

On calcule la distance entre les positions des occurrences, et on additionne la plus petite valeur soit la distance, soit v la longueur théorique. La première distance, $distance_1$, couvre le début et la fin du corpus. On additionne donc la première position, p_1, à la taille du corpus, moins la dernière position, p_3. On peut l'exprimer $distance_1 = p_1 + (N - p_3)$. La deuxième serait simplement $distance_2 = p_2 - p_1$ et ainsi de suite pour les distances suivantes. Nos deux exemples qui ont tous les deux une fréquence absolue de 3, se calculent ainsi :

$$\text{FRM} = \frac{1}{v} \times (\min((p_1 + N\ - p_3),\ \text{v}) + \min((p_2 - p_1),\ \text{v}) + \min((p_3 - p_2),\ \text{v}))$$

Dans le cas de gauche, avec des positions serrées, la fréquence serait réduite de 3 à pratiquement 1 :

$$\begin{aligned}
\text{FRM} &= \frac{1}{200k} \times (\min((1 + (600k - 10)), 200k) + \min((5 - 1), 200k) + \min((10 - 5), 200k)) \\
&= \frac{1}{200k} \times (\min((599991, 200k) + \min(4, 200k) + \min(5, 200k)) \\
&= 0,000005 \times (200k + 4 + 5) = 1,000045
\end{aligned}$$

En revanche, le cas de droite, où les positions des types dans le corpus sont équidistantes, la fréquence reste quasi inchangée.

$$\begin{aligned}
\text{FRM} &= \frac{1}{200k} \times \Big(\min((1 + 600k - 400k), 200k) + \min((200k - 1), 200k) \\
&\quad + \min((400k - 200k), 200k)\Big) \\
&= \frac{1}{200k} \times (\min(200001, 200k) + \min(199999, 200k) + \min(200k, 200k)) \\
&= 0,000005 \times (200000 + 199999 + 200000) = 2,999995
\end{aligned}$$

Lorsque l'on dresse la liste par ordre de fréquence après l'application de cette méthode, le classement du vocabulaire change. Dans le tableau 4.6, nous avons des exemples de changement de rang du vocabulaire du corpus littéraire Lotaya après la prise en compte de la dispersion. La fréquence de mots comme

ကလေး /kə̰lé/ (*enfant*), ကျွန်တော် /tɕəɴɔ̀/ (*je* (m)), ကျွန်မ /tɕəɴmə̰/ (*je* (f)), အိမ် /ʔèɪɴ/ (*domicile*) et မိန်းမ /méɪnmə̰/ (*femme*) a baissée après prise en compte de la dispersion des occurrences plus fréquentes dans certains textes, alors que pour ရဲ /jɛ́/ (*police*) et ကိစ္စ /keɪʔsə̰/ (affaires) la fréquence a augmentée sous l'effet d'une plus grande dispersion.

Vocabulaire	Fréquence absolue	Rang	FRM	Nouveau rang
ကလေး	2635	47	1165	↓ 55
ကျွန်တော်	3534	36	1065	↓ 61
ကျွန်မ	2926	42	565	↓ 102
အိမ်	1512	81	592	↓ 98
မိန်းမ	580	185	235	↓ 207
ရဲ	367	285	180	↑ 253
ကိစ္စ	350	298	175	↑ 262

TAB. 4.6 : Changement de rang dans le corpus Lotaya après calcul de FRM qui prend en compte la fréquence et la dispersion

Pour récapituler, l'usage de la fréquence réduite moyenne est une méthode pratique d'élaborer une liste de fréquence lexicale qui prend en compte la fréquence et la dispersion du vocabulaire d'un corpus, d'autant plus qu'il n'est pas nécessaire de créer un corpus divisé en plusieurs parties. Cette méthode ne nécessite que la taille du corpus en tokens, la fréquence absolue de chaque type et les positions des types dans le corpus. C'est cette méthode que nous avons choisie pour nos listes de fréquence par genre.

4.3.3.2 Fréquence réduite moyenne normalisée globale

La fréquence réduite moyenne normalisée globale (FRMNG), utilisée par Čermák et Křen (2011) dans la création de *A Frequency Dictionary of Czech*, permet de combiner les corpus de tailles différentes. Cette méthode repose sur la fréquence réduite moyenne décrite ci-dessus. Après calcul des fréquences réduites pour chaque corpus, les fréquences sont normalisées à une base inférieure à la taille du plus petit corpus. Nos corpus littéraires Lotaya et d'articles de presse BBC Burmese étant moins d'un million de tokens, nous avons choisi 500 000 comme

base de normalisation. Ensuite, on prend simplement la moyenne des valeurs pour chaque corpus.

t = ꜝ̀	Corpus		
	BBC	**Lotaya**	**TED**
Taille corpus	518 185	756 210	1 014 749
Fréquence absolue	13 473	20 621	40 207
Rang absolu	5	1	1
FRM	8990	13 361	26 878
FRM normalisée 500k	8675	8835	13 244
Fréquence globale		10 252	
Rang global		1	

TAB. 4.7 : Calcul de la fréquence réduite moyenne normalisée globale pour ꜝ̀ /kò/

L'effet de la dispersion n'est pas très marqué pour les vocables les plus fréquents, comme ꜝ̀ /kò/, particule qui marque le complément d'objet direct. Ceci est illustré dans le tableau 4.7, seul le rang du corpus BBC est modifié. Par contre, pour le vocabulaire très présent dans certains domaines, l'importance de prendre en compte la dispersion devient plus évidente. Le tableau 4.8 montre l'importance de la dispersion pour diminuer l'importance d'un vocable plus fréquent dans un genre textuel et moins dans d'autres. Ici le type ꜝ̀ʃ̀ɔ̀ /nàɪnŋàn/ (*nation*) est très fréquent dans le corpus BBC par rapport aux autres corpus.

t = ꜝ̀ʃ̀ɔ̀	Corpus		
	BBC	**Lotaya**	**TED**
Taille corpus	518 185	756 210	1 014 749
Fréquence absolue	3014	160	654
Rang absolu	24	557	196
FRM	1568	44	257
FRM normalisée 500k	1512	29	127
Fréquence globale		556	
Rang global		62	

TAB. 4.8 : Calcul de la fréquence réduite moyenne normalisée globale pour နိုင်ငံ

Voici un aperçu de la liste globale résultat de la combinaison de nos trois corpus authentiques.

Fréquences supérieures			**Fréquences inférieures**		
Rang	***f***	***V***	**Fourchette de rangs**	***f***	**Exemples de types**
1	10 252	ကို	2336–2515	10	ချန်ပီယံလိဂ် ဩစတြေးလျ တိရိစ္ဆာန်
2	9232	တယ်	2516–2736	9	အရှည် ဤသို့ဖြင့် ဆက်ဆံမှု
3	8794	ပါ	2737–3042	8	ရှေးဟောင်း တစ်စုံတစ်ယောက် ဘောင်းဘီ
4	7772	က	3043–3426	7	သဘောတူညီမှု ဒီဂျီတယ် သံဃာ
5	7411	တွေ	3427–3871	6	ပလပ်စတစ် လျှို့ဝှက်ချက် ဒစ်ဂျစ်တယ်
6	5937	တဲ့	3872–4526	5	အရူးအမူး ဇန်နဝါရီလ မက်ခရွန်
7	5452	မှာ	4527–5450	4	ပဲခူး စကားဝိုင် ပြုပြင်ပြောင်းလဲရေး
8	4853	မ	5451–7031	3	ရေခဲသေတ္တာ ထမင်းကြော် ငြိမ်းငြိမ်းချမ်းချမ်း
9	4011	တာ	7032–10 323	2	ဓာတ်ပုံဆရာ ပျက်ကျမှု ယူနီဖောင်း
10	3678	နေ	10 324–	1	ကန်ထရိုက်တာ စစ်သူရဲ အော့ကြောလန်

TAB. 4.9 : Fréquences supérieures et inférieurs de la liste de fréquence générale avec fréquences normalisées à $500k$

Un inconvénient de cette méthode provient de la normalisation. Les fréquences des hapax ($f(t) = 1$) des corpus les plus grands sont divisées lors de la normalisation. Dans ce cas, nous normalisons les fréquences en dessous de 1 en 1. Ceci n'est pas très problématique pour nos besoins, car nous sommes plutôt intéressée par le vocabulaire le plus fréquent.

4.4 Observations sur corpus

4.4.1 *Taille de liste de fréquence et couverture lexicale*

L'intérêt majeur de listes de fréquence est d'évaluer la taille de vocabulaire nécessaire pour qu'un étudiant puisse comprendre un texte. Selon une étude de Hu et Nation (2000) sur la relation entre couverture lexicale et compréhension de la lecture chez les étudiants de l'anglais en L2, une compréhension adéquate n'est obtenue qu'avec une connaissance de 98% du vocabulaire du texte, bien que certains apprenants parviennent à comprendre le texte avec seulement une

connaissance de 90 à 95%. Ces résultats ont été globalement confirmés par Laufer et Ravenhorst-Kalovski (2010), qui ont conclu que le seuil de 98% est nécessaire pour les étudiants L2 qui veulent poursuivre des études universitaires en anglais. Néanmoins, ils ont aussi trouvé que 95% (y compris environ 2% de noms propres) peuvent s'avérer suffisant pour comprendre à ce même niveau, mais à l'aide de dictionnaires ou d'un encadrement spécifique. Nation (2006), partant de ce principe qu'il faut connaître 98% d'un texte anglais pour le comprendre facilement, maintient qu'un vocabulaire de 8000 à 9000 types serait nécessaire pour couvrir 98% de texte. La structure du birman étant différent de l'anglais, on ne peut être certain que ces mêmes seuils s'appliquent pour la compréhension de textes par les étudiants de BLE. Toutefois, à partir de nos listes lexicales, nous pouvons identifier le vocabulaire fondamental de la langue générale (FRMNG) et, plus précisément, le vocabulaire fondamental pour certains genres textuels spécifiques. Pour ce faire, nous nous sommes intéressée à la couverture lexicale de nos listes de fréquence lexicales créées à partir de notre liste globale et de chacun de nos trois corpus authentiques.

Afin de visualiser cette notion de couverture lexicale, le texte de la figure 4.4 est balisé de couleurs différentes selon les tranches successives de mille types de notre liste de fréquence lexicale globale. Le vocabulaire présent dans les mille premiers types les plus fréquents de notre liste est balisé en jaune, celui dans les deuxièmes mille types en vert et ainsi de suite[12]. En l'occurrence, il s'agit d'un extrait d'article de presse de notre corpus BBC Burmese. Tout d'abord, nous remarquons que la plupart du texte est couvert par la première tranche de mille types. L'importance de cette tranche de mille types les plus fréquents est en fait moins marquée pour les textes de presse que pour d'autres genres textuels, car le lecteur de la presse rencontre souvent du vocabulaire étranger et des noms propres dont les types sont moins fréquents. Cependant, les deux sont souvent connus du lecteur et donc constituent une aide à la compréhension. Ce texte balisé contient 3,6% de tokens en lettres latines et 6,4% de chiffres. 10% du texte est donc déjà facile à déchiffrer pour le lecteur. 11,8% des tokens du texte sont absents des premiers cinq milles types de notre liste, mais parmi ceux-ci, il y a aussi des noms propres, en lettres birmanes, comme အက်ပ်ဲလ် ⟨ʔkˣpɛlˣ⟩ *Apple*, ဆီးရီး ⟨s̥ʰɪ̈²rɪ̈²⟩ *Siri* et un emprunt assez transparent မိုဘိုင်း ⟨miulˣbiuŋˣ²⟩ *mobile*, 5,5%

12. La clé de couleurs se trouve sous le texte. Noter que les statistiques concernent le texte entier qui n'est que partiellement reproduit.

Clé couleur	Rangs types	Couverture tokens (%)
	$1 - 1000$	$62,7$
	$1001 - 2000$	$7,3$
	$2001 - 3000$	$0,9$
	$3001 - 4000$	$3,6$
	$4001 - 5000$	$3,6$
	TOTAL :	$78,2$

FIG. 4.4 : Couverture d'un texte par tranches d'une liste de fréquence

des tokens absents. Cela signifie que seulement 6,2% de tokens sont inconnus d'un apprenant qui connaît les premiers cinq mille types de notre liste globale. Le total de types connus est finalement assez proche du seuil de couverture nécessaire pour la compréhension mentionnée ci-dessus. Sans identification explicite du vocabulaire d'emprunt et de noms propres en lettres birmanes, il est difficile d'estimer l'apport moyen de ces catégories de vocabulaire, seuls les types en lettres latines sont facilement identifiables. L'importance de ces derniers sont de fait en moyenne moins importante que pour notre texte exemple. Pour le corpus BBC, les emprunts en lettres latines ne constituent finalement que 0,6% du texte, et pour les corpus TED et Lotaya, ils ne représentent que 2% et 1,1% respectivement. Par prudence, pour ne pas surestimer les capacités de couverture

de nos listes dans nos calculs, nous considérons que ce vocabulaire hors liste est inconnu.

4.4.1.1 Couverture moyenne par liste globale

En premier lieu, nous avons mesuré les capacités de couverture de notre liste globale sur nos corpus authentiques. Le tableau 4.10 et figure 4.5 montrent les résultats de couverture moyenne par corpus.

Rangs de types	Couverture moyenne (%)		
	BBC	Lotaya	TED
$1--1000$	76.3 (σ 6.6)	76.4 (σ 4.6)	80.0 (σ 5.9)
$1001--2000$	7.9 (σ 3.2)	7.3 (σ 1.8)	6.2 (σ 2.1)
$2001--3000$	4.1 (σ 2.5)	3.7 (σ 1.1)	3.1 (σ 1.2)
$3001--4000$	2.4 (σ 1.8)	2.3 (σ 0.9)	1.9 (σ 1.1)
$4001--5000$	1.6 (σ 1.5)	1.8 (σ 1.0)	1.2 (σ 0.6)
$5001--6000$	1.0 (σ 1.2)	1.2 (σ 0.7)	0.8 (σ 0.5)
$6001--7000$	0.8 (σ 1.0)	1.1 (σ 0.7)	0.7 (σ 0.4)
$7001--8000$	0.5 (σ 0.8)	0.6 (σ 0.6)	0.4 (σ 0.3)
$8001--9000$	0.4 (σ 0.7)	0.7 (σ 0.6)	0.4 (σ 0.3)
$9001--10000$	0.5 (σ 0.8)	0.5 (σ 0.3)	0.4 (σ 0.4)

TAB. 4.10 : Couverture moyenne des textes du corpus BBC, Lotaya et TED par tranches d'une liste de fréquence globale

Comme on pouvait s'y attendre, un grand nombre des tokens de nos corpus est composé des types les plus fréquents. Beaucoup de ce vocabulaire fréquent en birman est composée de vocabulaire à fonction grammaticale. Nous avons trouvé que plus de 50% de ces types figure au *Burmese/Myanmar Dictionary of Grammatical Forms* de Okell et Allott (2017)[13]. Plus précisément, les formes grammaticales représentent 51,7% des mille premiers types de la liste BBC, 57,3% pour le corpus Lotaya, et 59% pour le corpus TED. Si l'on considère les nombres totaux des formes grammaticales des corpus, 52,3% (BBC), 59,5% (Lotaya) et 59,6%

13. Cette liste (en Unicode) de formes grammaticales nous a été fournie par un des auteurs, John Okell.

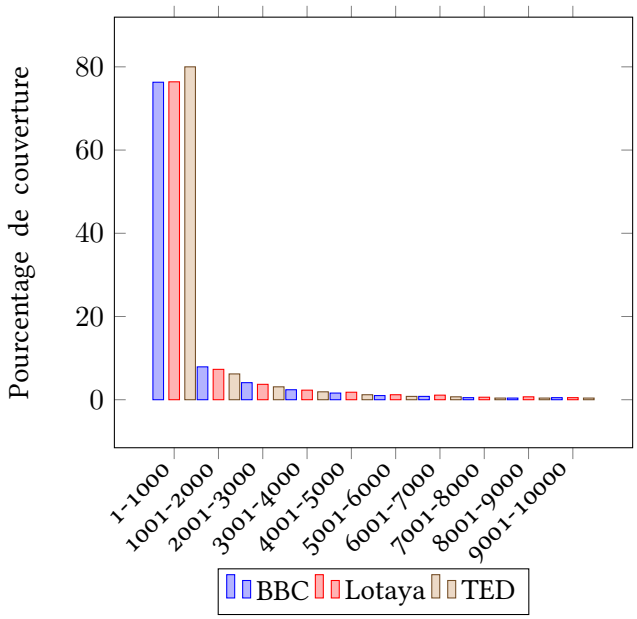


FIG. 4.5 : Couverture moyenne des textes du corpus BBC, Lotaya et TED par tranches d'une liste de fréquence globale

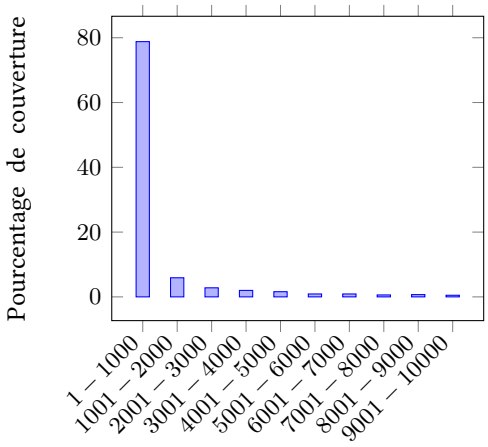
(TED), on voit que presque la totalité de formes grammaticales se trouve parmi les mille premiers types les plus fréquents des textes birmans.

Puisqu'il s'agit des textes à partir desquels nous avons construit nos listes, il était possible que ces résultats ne soient pas représentatifs de la langue en général. Nous avons donc décidé de comparer ces résultats avec la couverture d'un échantillon de textes externes à nos corpus.

4.4.1.2 Couverture moyenne de nouveaux textes par liste globale

Nous avons pris un petit échantillon de quinze textes : des articles de presse, des articles littéraires et des discours politiques informels. Cinq du site de BBC Burmese, de la rubrique ဆောင်းပါး /sʰáʊɴpá/ (*articles*), qui semble-t-il ont moins de noms propres potentiellement problématiques pour notre outil de segmentation et les dix autres proviennent de Okell (2018) : cinq articles littéraires de l'écrivaine Ludu Daw Amar et cinq discours spontanés de Aung San Suu Kyi datant des années quatre-vingt[14].

14. Taille totale du corpus échantillon : 39 657 syllabes ; 1151 phrases. Taille moyenne : 2643.8 syllabes par texte (écart-type : 1464.7) ; 76.7 phrases par texte (écart-type : 25.1)



Rangs de types	Couverture moyenne (%)
1 – 1000	78.8 (σ 5.9)
1001 – 2000	5.9 (σ 1.5)
2001 – 3000	2.8 (σ 1.0)
3001 – 4000	2.0 (σ 0.9)
4001 – 5000	1.6 (σ 0.9)
5001 – 6000	0.9 (σ 0.5)
6001 – 7000	0.9 (σ 0.6)
7001 – 8000	0.6 (σ 0.4)
8001 – 9000	0.7 (σ 0.5)
9001 – 10000	0.5 (σ 0.3)

FIG. 4.6 : Couverture moyenne d'un petit échantillon de textes externes aux corpus par tranches d'une liste de fréquence globale

Les résultats sont résumés dans la figure 4.6. Les mille premiers types de notre liste offrent de loin la plus grande couverture (78,8%). Ce résultat n'est guère étonnant, étant donné que les textes sont de genres similaires aux corpus utilisés pour créer notre liste globale.

On note aussi que, comme pour nos grands corpus authentiques, 59,0% des mille premiers types sont composés de formes grammaticales, seulement 0,9% de formes grammaticales se trouvant en dehors des mille premiers types.

4.4.2 *Liste de fréquence lexicale globale et vocabulaire de manuels didactiques*

S'il est difficile de fournir une évaluation de l'utilité de notre liste de fréquences globale, nous avons cependant pensé qu'il serait intéressant de la comparer avec les listes de vocabulaire des manuels d'apprentissage du birman, car elles nous donnent une indication de ce que les instructeurs en birman considèrent comme du vocabulaire important. L'un des problèmes majeurs de la comparaison entre les manuels d'enseignement et un corpus basé sur des matériaux authentiques essentiellement textuels est que l'enseignement ne se concentre pas simplement sur la lecture, et qu'une grande partie des bases d'une langue repose sur la compréhension et l'expression orale. On peut donc s'attendre à ce qu'il y ait un

inévitable déficit de représentativité. Cela semble être un problème dans notre comparaison, mais il ne semble pas être aussi important que nous l'aurions pensé et les différences pourraient s'expliquer par d'autres facteurs, principalement liés à la segmentation. Le tableau 4.11 et les histogrammes de la figure 4.7 qui donne un aperçu global des résultats des comparaisons représentent le pourcentage du vocabulaire de chaque manuel qui apparaît dans chaque tranche de mille types de notre liste. Par exemple, 21,6% du vocabulaire du manuel *Burmese by Ear* (BBE) figure dans les mille premiers types de notre liste de fréquence lexicale globale, 7,8% dans la deuxième tranche de mille types et ainsi de suite.

Pour chaque manuel, on constate que le vocabulaire le plus partagé se situe dans la première tranche de mille types de notre liste globale. Cette tranche est composée en grande partie de formes grammaticales. Si l'on regarde la courbe de l'histogramme, nous pouvons constater une diminution constante du pourcentage de vocabulaire partagé dans presque tous les cas. Les manuels où la courbe est la moins cohérente sont *Burmese by Ear* (BBE) et *Myanmar Newspaper Reader* (MNR), essentiellement dû à la proportion d'entités nommées dans les textes.

Rangs de types	**Vocabulaire dans liste globale (%)**							
	BBE	**CB**	**MdB1**	**MFG**	**TALPCo**	**TUFSD**	**MNR**	**AiB**
$1-1000$	$21,6$	$36,4$	$37,9$	$37,4$	$44,5$	$35,1$	15	$36,1$
$1001-2000$	$7,8$	13	$10,4$	$10,7$	$15,9$	$11,1$	$7,8$	$14,8$
$2001-3000$	$6,7$	$8,9$	$6,7$	$6,6$	$9,8$	$6,2$	5	$7,6$
$3001-4000$	$4,6$	$5,3$	$5,1$	$5,8$	$5,2$	$3,3$	$4,3$	6
$4001-5000$	$3,1$	$4,2$	$3,7$	$3,6$	$3,6$	$2,4$	$2,7$	$3,8$
$5001-6000$	$3,4$	$2,6$	$2,3$	$1,7$	$2,2$	$1,6$	$2,1$	3
$6001-7000$	$2,9$	$2,2$	$2,2$	$3,1$	2	$1,7$	2	$1,9$
$7001-8000$	$2,7$	$1,3$	$1,5$	$0,7$	$1,4$	1	$0,8$	$1,4$
$8001-9000$	$1,9$	$1,6$	2	$1,7$	$1,5$	1	$0,9$	$1,7$
$9001-10\,000$	$1,8$	$1,1$	$0,8$	$0,3$	$1,1$	$0,8$	$1,2$	$1,1$

TAB. 4.11 : Couverture de liste de fréquence lexicale globale du vocabulaire de huit ressources didactiques

Pour nous, le vocabulaire qui n'est pas partagé entre la liste et les manuels est encore plus révélateur. Le tableau 4.12 résume le pourcentage de ce vocabulaire

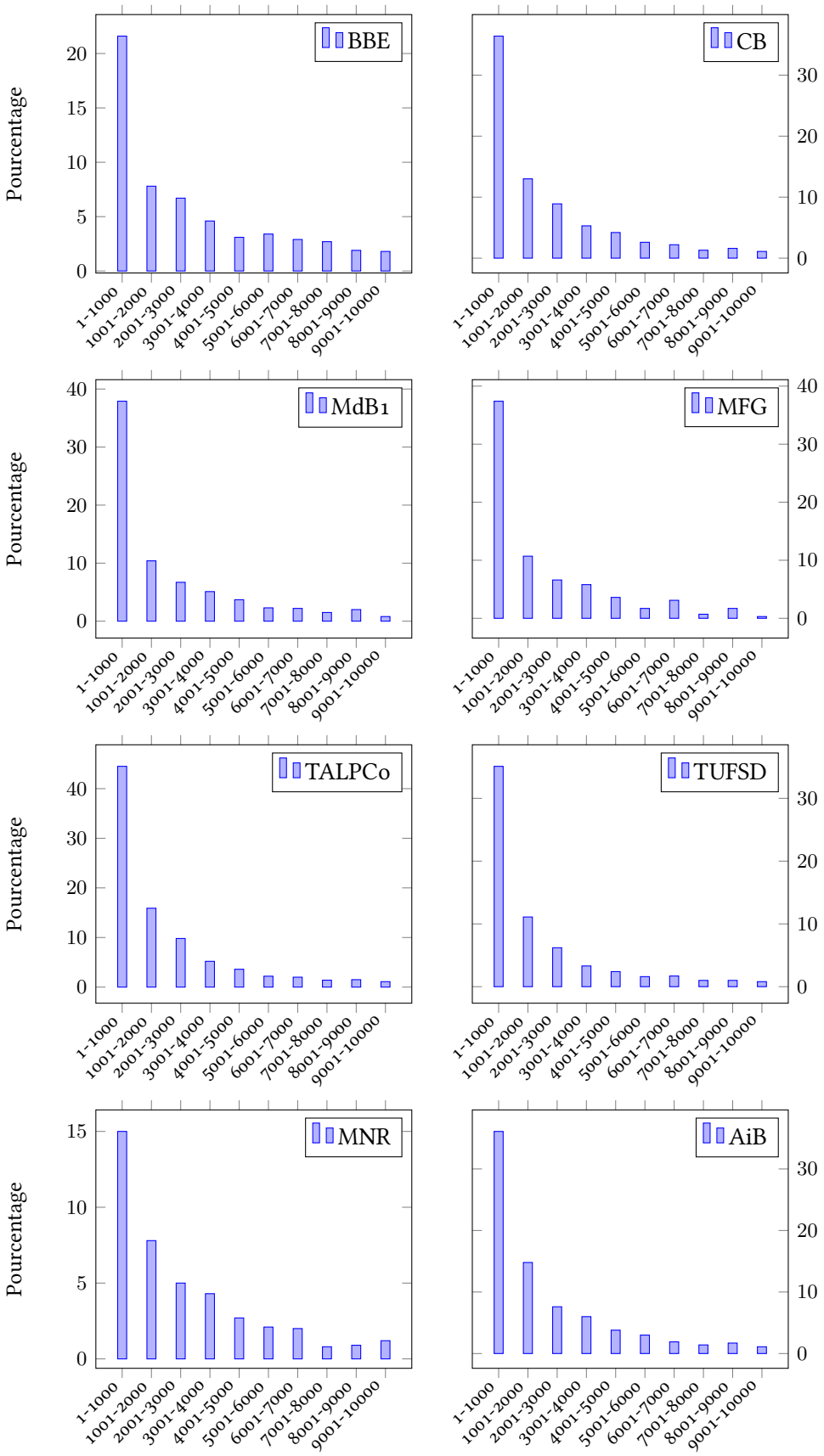


FIG. 4.7 : Comparaison de couverture de liste globale du vocabulaire de huit ressources didactiques : *Burmese by Ear*, *Colloquial Burmese*, *Manuel de birman 1*, *Myanmar Flower Grammar*, *TALPCo-TUFS Vocabulary*, *TUFS Dialogues*, *Myanmar Newspaper Reader*, *Advancing in Burmese*

manquant pour chaque ouvrage.

La première explication concerne les entités nommées, celles qui sont soit absentes de notre liste, soit segmentées autrement. Par exemple, les vocables de BBE, ကမ်းနားလမ်း /kánnálán/ (*Strand Road*) et မဟာဗန္ဓုလလမ်း /mə̰hàbànd̪ṵlə̰lán/ (*Maha Bandula Road*) sont chacun répertorié dans le dictionnaire de *Motor* sans လမ်း /lán/ qui veut dire *rue* et que nous avons donc segmentés en deux tokens. D'autres noms propres ne figurent pas du tout dans notre liste, comme မဟာမြတ်မုနိ /mə̰hàmja?mṵnḭ/ (nom du grand bouddha au temple Mahamuni) (du manuel CB), ou ခိမုရ /kʰḭmṵjə̰/ (*Kimura*, un nom de famille japonais) (TALPCo).

Comme les entités nommées, certains vocables dans les manuels sont moins segmentés par les auteurs des manuels que par notre outil de segmentation. Dans BBE par exemple, nous avons အစိုးရအမှုထမ်း /ʔə̰sójə̰ʔə̰m̥ṵtʰán/ (gouvernement + personnel) *fonctionnaire*, ဂျပန်ဖိနပ် /dʑə̰pànpʰḭnaʔ/ (Japon + chaussure) (*geta* chaussures traditionnelles japonaises). Dans TUFSD, nous avons trouvé une tendance d'exprimer la quantité en accolant le nombre au classificateur, comme တစ်ခွက် /tɪʔkʰʊ̰weʔ/ *un verre*.

Code corpus	**Vocabulaire absent (%)**
BBE	42,4
CB	23,4
MdB1	27,4
MFG	28,4
TALPCo	12,4
TUFSD	35,9
MNR	58,3
AiB	22,6

TAB. 4.12 : Pourcentage du vocabulaire des manuels de birman absent de la liste de fréquence lexicale globale

Certains vocables directement empruntés à d'autres langues ne figurent pas dans notre liste comme ဟီရဂန /hìjə̰gə̰nə̰/ (*hiragana*, un des systèmes d'écriture japonaise), ကာရာအိုကေ /kàjàʔòkè/ (*karaoke*) (TALPCo) et du vocabulaire anglais courant de la langue parlée ဆောရီးပဲ /sʰɔ́jípɛ́/ (*sorry*) ဘီစကစ် /bìsə̰kɪʔ/ (*biscuit*)

(BBE).

Les manuels donnent souvent des phrases entières courantes comme vocabulaire à assimiler tel quel, alors que nous avons choisi de les segmenter, ainsi မဟုတ်လား /mə̰hoʊʔlá/ *n'est-ce pas ?* (CB), ကိစ္စမရှိပါဘူး /keɪʔsə̰mə̰ʃḭpàbú/ (*ce n'est pas important, ce n'est pas grave*) (CB) et သွားလို့ကိုးမယ် /θwálaɪʔʔùmɛ̀/ (*allons-y !*) (TUFSD).

Le faible pourcentage de vocabulaire du corpus TALPCo qui n'est pas partagé avec notre liste, 12%, s'explique simplement par la nature répétitive du contenu du corpus.

Étant donné que notre liste est composée à partir de seulement trois types de textes, nous ne pouvions pas attendre à ce qu'elle soit parfaitement représentative de la langue en général, mais il est vrai que l'absence de contenu oral et l'importance relative des textes traduits, plutôt que des textes originaux en birman, semblent en effet compromettre la représentativité et donc avoir un effet limitatif sur l'utilité de la liste pour les apprenants. Cette comparaison a mise en évidence des omissions de vocabulaire qui concernent des aspects importants de la culture birmane, comme သင်္ကြန်ရေသဘင်ပွဲ /θɪ̀ɴtɕàɴjèθə̰bɪ̀ɴpwé/ *le festival de l'eau de Thingyan*, certains aliments très quotidiens en Birmanie comme သရက်သီး /θə̰jɛʔθí/ (*mangue*) (MFG). Il manque aussi du vocabulaire utile au quotidien, comme အားကစားရုံ /ʔákə̰sájòʊɴ/ (*gymnase*), ပင်လယ်ကမ်းခြေ /pɪ̀ɴlɛ̀káɴtɕʰè/ (*mer + rive*, c'est-à-dire *plage*) (MdB1), ရုံးပိတ်ရက် /jóʊɴpeɪʔjɛʔ/ (*vacances*) (CB) et même la salutation la plus courante မင်္ဂလာပါ /mɪ̀ɴgə̰làpà/ (*Bonjour*) qui est présente dans tous les manuels.

Ce manque de représentativité pourrait en partie être compensé par l'ajout d'un corpus de sous-titres de films et séries télévisées. Depuis le premier corpus de ce type en français de New et al. (2007), de tels corpus ont été créés pour analyser les effets de la fréquence lexicale (surtout en psycholinguistique) pour un nombre de langues[15], car il a été observé que le contenu se rapproche beaucoup du langage courant (Brysbaert et New 2009). Lors de la création de nos corpus,

15. Ces corpus sont souvent créés à partir de sous-titres du site *Open Subtitles* https://www.opensubtitles.org et par convention sont toujours appelé SUBTLEX plus un code de langue ou pays ainsi : SUBTLEX-UK en anglais britannique (Van Heuven et al. 2014), SUBTLEX-DE en allemand (Brysbaert, Buchmeier et al. 2011), SUBTLEX-NL en néerlandais (Keuleers et al. 2010), SUBTLEX-PL en polonais (Mandera et al. 2014), SUBTLEX-CH en chinois mandarin (Cai et Brysbaert 2010), SUBTLEX-PT en portugais (Soares et al. 2015), SUBTLEX-ESP en espagnol (Cuetos et al. 2012), SUBTLEX-CAT en catalan (Boada et al. 2019), SUBTLEX-GR en grec (Dimitropoulou et al. 2010), SUBTLEX-AL en albanais (Avdyli et Cuetos 2013) et SUBTLEX-VIET en vietnamien (Pham et al. 2019).

nous nous sommes heurtés à un problème de disponibilité, car il y avait très peu de fichiers de sous-titres exploitables disponibles en birman. Bien qu'il y ait maintenant beaucoup plus de fichiers disponibles, nous avons toujours des réserves quant à leur utilisation, car il s'agit exclusivement de traductions, presque entièrement de superproductions cinématographiques américaines ou chinoises, ce qui ne va guère résoudre les omissions de vocabulaire qui concerne le monde birman.

4.5 Résumé

Ce chapitre s'est concentré sur le développement d'une méthode permettant de créer une liste de fréquences globale issue de corpus créés à partir de ressources authentiques de taille limitée ou différente, l'idée étant de pouvoir utiliser ces informations pour évaluer la difficulté d'un texte pour des apprenants de langue étrangère en se basant uniquement sur la fréquence lexicale. On espère que cette méthode pourra s'avérer utile non seulement pour le birman, mais aussi pour d'autres langues moins bien pourvues en ressources. Comme nous l'avons vu, il existe des limites à la comparaison de notre liste avec le vocabulaire utilisé dans les manuels d'enseignement de langues, qui se concentrent non seulement sur la lecture, mais aussi sur les compétences culturelles et orales insuffisamment représentées par nos corpus. Nous constatons également que les différences dans la manière dont le vocabulaire est présenté aux étudiants et la façon dont nous avons défini nos tokens exagèrent les différences entre notre liste de fréquence globale et les listes de vocabulaire des apprenants. Il a été décidé de ne pas inclure le contenu des listes de vocabulaire des manuels des apprenants dans le dictionnaire de l'outil de segmentation *Motor*, afin d'essayer de maintenir une certaine cohésion dans la façon dont le vocabulaire est segmenté, mais cela n'est peut-être pas le meilleur choix et l'inclusion de ce vocabulaire améliorerait certainement les résultats du point de vue des apprenants. Étant donné que nous nous intéressons à l'évaluation du matériel textuel pour la lecture, nous ne pensons pas que l'absence de contenu oral s'avérera nécessairement problématique pour calculer la difficulté de textes à l'aide de la liste de fréquence globale créée par cette méthode.

Chapitre 5

La lisibilité par la fréquence lexicale

Les mesures de la lisibilité sont utilisées dans une grande variété de contextes, de l'éducation des enfants à l'apprentissage des langues, en passant par les études médicales, le journalisme et la bibliothéconomie. Bien que l'histoire des études de lisibilité en général soit longue (elle remonte à plus d'un siècle), pour autant que nous puissions en juger, il n'y a pas encore eu d'études sur la lisibilité du birman, que ce soit pour les locuteurs natifs ou les apprenants du birman comme langue étrangère.

La recherche sur la lisibilité consiste à mesurer la difficulté de lecture d'un texte. Ces recherches reposent sur des hypothèses sous-jacentes concernant la nature de la difficulté et les capacités et besoins du lecteur. Pour pouvoir faire correspondre un texte à un certain niveau de difficulté à un certain lecteur, les chercheurs font des généralisations sur ce qui constitue la difficulté et regroupent les lecteurs en populations traitées comme homogènes. Des présentations détaillées des approches spécifiques de la difficulté des textes se trouvent dans DuBay (2004), François (2015) et Bailin et Grafstein (2016). Bailin et Gradstein considèrent la lisibilité du point de vue de la bibliothéconomie, en mettant l'accent sur la communication, mais nombre de leurs commentaires et critiques sur la recherche en matière de lisibilité sont intéressants à examiner du point de vue de l'apprentissage des langues étrangères et la mise en œuvre de notre travail.

5.1 Les inconvénients des formules de lisibilité

Les premières méthodes de la mesure de la difficulté des textes étaient fondées sur de simples statistiques textuelles et/ou des listes de mots à haute fréquence, comparant les caractéristiques de textes dont la difficulté avait déjà été estimée pour créer des formules de lisibilité, facilement applicables pour évaluer la

difficulté de nouveaux textes. Lively et Pressey (1923) ont utilisé les données de la liste de Thorndike (1921) et le nombre de types pour mille tokens pour évaluer ce qu'ils appellent la charge ou le fardeau du vocabulaire d'un texte. Le système de classement de Thorndike utilisé est basé sur la fréquence et la dispersion dans un corpus fortement orienté vers les textes bibliques et littéraires. Outre le problème évident de la représentativité de ce corpus, cette approche est critiquée par Bailin et Grafstein (2016) dans la mesure où elle suppose qu'un mot plus fréquent est plus facile à comprendre sur une échelle incrémentale. Cette critique est probablement plus valable pour les lecteurs de langue maternelle que pour les apprenants de langue seconde, surtout lorsque l'utilisation de la lisibilité pour les apprenants de langue seconde est motivée par l'acquisition progressive de vocabulaire plutôt que par la compréhension[1].

Bailin et Grafstein (2001) analysent les formules classiques de lisibilité, démontrant comment les hypothèses linguistiques qui les sous-tendent sont déficientes. Les formules qui calculent les indices de lisibilité pour l'anglais telles que l'indice FKGL[2], les indices de Dale et Chall (1948), l'indice GFI[3], ou l'indice SMOG[4] tentent toutes de prendre en compte la difficulté du vocabulaire et la complexité syntaxique s'appuyant sur des mesures de caractéristiques superficielles des textes.

La difficulté du vocabulaire est mesurée soit par la présence ou l'absence des vocables dans une liste de mots fréquents, soit par la longueur des vocables (mesurée en syllabes). La plus grande faiblesse de l'utilisation d'une liste de mots à haute fréquence est que le vocabulaire d'une langue évolue constamment, de sorte que de telles listes doivent toujours être réévaluées, en utilisant un corpus de textes plus récent. Cette critique s'applique également à notre liste de fréquences lexicales, car une partie de celle-ci peut être considérée comme contemporaine (les articles de journaux) et une autre moins (les textes littéraires). L'autre critique

1. Comme exemple de cette approche d'apprentissage progressive du vocabulaire on peut citer l'application au chinois mandarin par Bellassen (2008). On note que l'auteur lui-même (Bellassen 2010) met en garde contre une approche progressive qui n'utilise pas les mots comme base de l'apprentissage du chinois, mais les sinogrammes (comme Lyssenko et Weulersse (1986)), une méthode qui selon lui « s'oriente vers une compétence communicative faible ». Le choix de l'unité minimale qui représente le vocabulaire est donc primordial.

2. FKGL = Flesch-Kincaid Grade Level (Kincaid et al. (1975), une version modifiée de l'indice de Flesch (1948))

3. GFI = Gunning-Fog Index (Gunning 1952)

4. SMOG = Simple Measure of Gobbledygook (McLaughlin 1969)

de Bailin et Grafstein (2001) est que des groupes socio-économiques différents et des personnes d'âges différents auront des niveaux de familiarité différents avec le même vocabulaire. D'une certaine manière, nous pouvons également appliquer cette critique à la familiarité des apprenants de langue seconde avec certains éléments de vocabulaire, tels que les noms propres (que nous avons choisi d'analyser de la même manière que tout autre vocabulaire) ou les emprunts à des langues connues de l'apprenant. Par exemple, quand l'apprenant rencontre အင်တာနက် /ʔɪ̀ɴtànɛʔ/ pour la première fois, il doit apprendre qu'il s'agit du mot habituel pour *internet* en birman. Un autre problème qu'ils mentionnent et qui est effectivement pertinent pour notre travail est la question de la polysémie. Les mots ont des significations différentes dans des contextes différents, et le fait de les découper en listes de mots constitue une simplification grossière de leur signification. Ils donnent l'exemple du mot anglais *set* qui aurait 430 significations selon le contexte[5]. La longueur des vocables dans ces formules est présumée être en corrélation avec leur difficulté, ce qui est particulièrement problématique, car cela ne tient pas compte de la connaissance qu'a un lecteur des processus de dérivation, comme l'affixation. En effet, Richaudeau (1976) dans ses expériences sur la lisibilité a trouvé « une corrélation fortement négative entre le nombre de mots retenus par phrase et la longueur des mots ». On pourrait utiliser cette observation pour ne prendre en compte que les morphèmes les plus simples du birman et enseigner les affixes à part, mais comme nous l'avons remarqué pour les emprunts, l'apprenant d'une langue étrangère qui voit un vocable formé par affixation dont l'équivalent n'est pas dérivé de la même manière dans sa propre langue doit apprendre quand un processus est utilisé. Par exemple, en birman လူသတ်မှု /lùθaʔmṵ/ est formé de *personne* + *tuer* + မှု → *meurtre*. Même si l'affixation aide dans la compréhension, l'apprenant doit avoir vu le vocable pour l'apprendre. Le fait que l'apprenant d'une langue étrangère lise un texte pour apprendre une langue et non pour son contenu propre ou pour apprendre à lire est l'une des différences essentielles qui caractérisent les besoins des lecteurs de langue maternelle et des lecteurs de langue étrangère.

En général, les formules de lisibilité classiques assimilent la longueur de la phrase à la complexité syntaxique. A cet égard, Richaudeau (1976) confirme les conclusions de Pearson (1974) : que les phrases les plus courtes n'augmentent pas

5. Au moins selon l'Oxford English Dictionary (Moore 1997)

systématiquement la mémorisation ou la compréhension, et que, au contraire, des structures syntaxiques complexes sont plus facile à mémoriser et à comprendre. Ceci reflète le fait qu'un texte bien structuré est plus facile à retenir et qu'une augmentation du contexte facilite la compréhension des mots ou (pour l'apprenant d'une langue étrangère) permet de deviner plus facilement le sens du vocabulaire.

Les indices de lisibilité sont généralement spécifiques à une langue, car ils s'appuient sur une combinaison de ces analyses linguistiques superficielles avec l'évaluation de la difficulté de textes par les informateurs, que ce soit les jugements de spécialistes (créateurs de manuels scolaires par exemple) ou de simples lecteurs. Ces corpus de textes classifiés par niveau de difficulté sont utilisés pour calculer les coefficients de la formule. Les indices ont été adaptés à de nombreuses langues, comme le français (Henry 1975), l'espagnol (Spaulding 1956) et l'arabe (Daud et al. 2013).

La seule formule de lisibilité qui ne repose pas sur des variables calculées pour une langue spécifique est celle de Tuldava (1993) qui utilise uniquement la longueur moyenne des mots en syllabes et la longueur moyenne des phrases en mots. Cette méthode, qui a été testée sur des textes allemands, russes et estoniens, ne prétend pas être très précise, mais simplement indicative du niveau de difficulté probable pour le lecteur. Cette formule semblerait intéressante à tester pour les langues peu dotées, mais étant donné la difficulté de déterminer de façon systématique ce qu'est un mot en birman, nous n'avons pas encore tenté de l'appliquer au birman[6].

5.2 La lisibilité comme classification

L'apprentissage automatique offre de plus grandes perspectives pour calculer la lisibilité en se basant sur beaucoup plus de caractéristiques que les traits lexicaux ou syntaxiques, comme les caractéristiques sémantiques, discursives (Pitler et Nenkova 2008) et les facteurs psycholinguistiques comme la cohérence (Crossley et al. 2008). En apparence, cela promet de répondre au désir de Bailin et Grafstein (2016) de rendre compte d'une théorie plus complète de la lisibilité.

Bien que ces méthodes soient en elles-mêmes moins spécifiques à la langue

6. La difficulté d'appliquer une telle formule à des textes en thaï serait d'autant plus importante, car non seulement les mots ne sont pas délimités, mais il n'y a pas de marque de fin de phrase explicite non plus.

que les formules de lisibilité, elles posent des problèmes pour les langues peu enseignées, car elles nécessitent toujours des corpus d'entraînement, et des corpus annotés de plus en plus finement (François 2015). Les corpus déjà classifiés par niveaux de difficulté pour les apprenants de langue étrangère reposent sur des manuels de langue conformes au *Cadre européen commun de référence pour les langues*, le CECR (Conseil de l'Europe 2021). Ce cadre de référence décrit les niveaux selon des compétences d'apprenants dans des situations spécifiques qui ne sont pas expressément conçus pour être finement associés à des niveaux de difficulté, même s'il est évident que le fait de progresser dans les différents niveaux accroît les compétences linguistiques des apprenants. Puisque l'apprentissage automatique applique un modèle statistique sur les corpus classifiés afin de créer un modèle de classification, nous revenons à la question de la définition de la difficulté et la qualité des corpus.

De tels corpus classifiés en plusieurs niveaux sont difficiles à trouver pour une langue peu dotée, et en outre, les niveaux de difficulté de langues peu enseignées ne sont pas définis de façon standardisée comme le CECR. On attribue un chiffre sur une échelle, un niveau de scolarité, ou bien des étiquettes telles que *débutant*, *faux débutant*, *intermédiaire* ou *avancé*. En conséquence de ce manque de standardisation, les niveaux choisis sont souvent difficiles à attribuer lors de la création d'un corpus classifié et difficilement interprétables pour l'utilisateur. Si le créateur du corpus n'a pas bien classifié les textes, ou l'utilisateur du modèle estime que la classification du corpus d'entraînement ne lui convient pas, la classification est peu utile.

5.3 Le classement par ordre de difficulté

Une alternative au classement par niveau de difficulté est le classement de textes par ordre de difficulté. Plutôt que d'attribuer une classe de difficulté à un texte, un ensemble de textes est ordonné par difficulté croissante. Cette approche a été utilisée par Ghadirian (2002) pour créer un parcours de lecture de textes pour les apprenants de l'anglais langue étrangère. L'idée est de classer des textes selon un ordre de lecture qui facilite l'acquisition du vocabulaire de manière progressive par la répétition lexicale. Le tri de textes est basé sur la couverture textuelle de listes lexicales : une liste de vocabulaire connu et une liste de vocabulaire cible constituée du vocabulaire des textes de haute fréquence qui

ne se trouve pas dans la liste de vocabulaire connu. Le texte le plus facile est celui qui contient le plus de vocabulaire connu. Le vocabulaire cible du texte est ensuite rajouté à la liste de vocabulaire connu qui augmente au fur et mesure de la création du parcours de lecture. Une approche analogue a été utilisée par Huang et Liou (2007) pour classifier des textes anglais, et par Lewis-Wong et Mkhitaryan (2016) pour un classement ordonné de textes hindis et thaïs. Cette approche relativement simple à mettre en œuvre ne permet pas toutefois d'insérer un nouveau texte sans recréer le parcours. Ce classement ne concerne exclusivement que le vocabulaire et sa présence ou non parmi le vocabulaire connu.

Le classement avec un tri par insertion binaire basé sur un apprentissage automatique permet non seulement de classifier des textes par ordre de difficulté, mais aussi de rajouter un nouveau texte dans une suite de textes déjà rangés par ordre de difficulté. Cette méthode ne nécessite qu'un modèle de classification binaire, autrement dit un corpus avec deux catégories de difficulté, *facile* et *difficile*, une tâche plus aisée à mettre en œuvre pour une langue peu dotée. Bien que nous n'utilisons que la fréquence lexicale, toute caractéristique de lisibilité seule ou en combinaison avec d'autres pourrait aussi être utilisée pour la comparaison.

Dans une étude sur la qualité des textes, Pitler et Nenkova (2008) ont employé une méthode de comparaison de textes sur un corpus de trente textes journalistiques anglais du *Wall Street Journal* qui ont été évalués par trois étudiants pour la qualité de leur rédaction. Elles ont combiné des caractéristiques lexicales, syntaxiques et discursives pour produire leur modèle par le biais de machines à vecteurs de support (SVM). L'application qui utilise aussi la comparaison de textes, en rajoutant un algorithme de tri binaire pour les classer par ordre de difficulté est celle de Tanaka-Ishii et al. (2010), qui classe des textes avec la fréquence lexicale comme seule métrique. Au lieu d'essayer de rendre compte d'une théorie complète de la lisibilité qui tiendrait compte de sa complexité, nous choisissons de simplifier notre définition de la lisibilité à une lisibilité du seul vocabulaire, sans prétendre tester autre chose que la fréquence lexicale. Pour les apprenants et les formateurs d'une langue étrangère peu enseignée, la question n'est pas tant de savoir si un texte est difficile pour l'apprenant que de savoir s'il est utile. En d'autres termes, quels sont les textes

à utiliser pour apprendre la langue. Cette technique de lisibilité devient donc une mesure de l'utilité du vocabulaire du texte plutôt que sa difficulté. Compte tenu de la forte corrélation entre le vocabulaire et la compétence dans l'apprentissage d'une langue étrangère (Laufer 1992 ; Alderson 2005 ; Milton 2010) et le défi que représente le vocabulaire, cette approche semble particulièrement prometteuse. En outre, certaines des hypothèses qui sous-tendent d'autres méthodes de lisibilité ne s'appliquent plus, car le modèle ne repose pas sur la comparaison de textes à d'autres textes, mais plutôt sur la comparaison de textes à la langue elle-même (représentée par les fréquences dans le corpus).

C'est donc cette dernière que nous avons appliquée aux textes birmans. En mettant en œuvre cette méthode, nous ne cherchons pas à en tester la validité, mais à comprendre les enjeux de son utilisation sur la langue birmane, étant donné que nous disposons de moins de données pour le birman que les créateurs de la méthode n'en avaient pour les langues qu'ils ont testées, l'anglais et le japonais. Avec une plus grande quantité de textes, cette méthode aurait pu être utilisée pour attribuer des niveaux de difficulté aux textes. Sato (2014) a d'ailleurs utilisé une méthode similaire à celle de Tanaka-Ishii et al. (2010) pour classer les textes d'un grand corpus, puis il a attribué un niveau de difficulté à chaque texte au moyen d'une méthode de classement appelée *Stanine* (pour *STAndard NINE*) qui divise les textes d'un corpus en neuf niveaux de difficulté quand ils présentent une distribution gaussienne.

5.3.1 *Le tri par insertion binaire*

Le tri par insertion binaire fait appel à un algorithme de recherche binaire qui trie par comparaison de deux textes. L'algorithme commence avec deux listes : la liste des textes à trier et une liste vide qui sera utilisée pour insérer les textes à la bonne position. A la fin du tri cette liste vide sera la liste des textes ordonnés. A chaque itération un texte est sorti de la liste à trier pour le comparer au contenu actuel de la liste triée par difficulté croissante et l'y placer à la bonne position.

On peut conceptualiser l'algorithme comme un système qui compare toujours le texte à positionner avec le texte au milieu de la liste triée. A chaque itération l'algorithme commence la comparaison avec le texte médian de toute la liste triée, puis selon le résultat de la comparaison, choisit l'une ou l'autre moitié de

la liste triée pour poursuivre la comparaison. Autrement dit, si le texte est plus facile que celui au milieu de la liste, on réduit la liste à sa moitié inférieure et on recommence le même processus de comparaison avec le milieu de cette moitié de la liste, plus *facile* que le texte que l'on vient de juger *difficile*. Inversement, si le texte à positionner est jugé plus difficile que celui au milieu de la liste triée, on le compare ensuite avec le milieu supérieur de la moitié de la liste triée qui est plus difficile. On répète le processus jusqu'à ce que le texte choisi soit placé entre un texte plus *facile* et un texte plus *difficile*, ou bien s'il se trouve à l'un ou l'autre extrémité de la liste. Cette méthode de recherche binaire est plus rapide qu'une recherche linéaire, car le nombre de comparaisons nécessaires est réduit.

La première itération prélève deux textes de la liste des textes, les compare pour décider lequel des deux est le plus *facile*, puis on les range avec le texte le plus *facile* en début de liste (à gauche) et le plus *difficile* à sa suite (à droite). Un troisième texte est ensuite prélevé de la liste de textes à trier, et on commence la comparaison avec le texte à gauche. S'il est jugé plus facile, il est placé à gauche de ce texte. S'il est jugé plus difficile, il est ensuite comparé avec le texte à sa droite et soit laissé en position, soit placé à la fin de la liste. A ce stade, la liste triée est ainsi composée de trois textes rangés par ordre de difficulté. La figure 5.1 donne un exemple schématique simplifié de l'algorithme où chaque lettre représente un texte à trier. Le résultat en ordre alphabétique serait l'équivalent de notre liste de textes ordonnés du plus facile (A) à plus difficile (I). Ceci est un exemple simplifié pour expliquer le concept de tri par insertion binaire. En réalité, nous comparons le texte à insérer avec plus d'un texte au milieu à chaque fois, en suivant l'exemple de Tanaka-Ishii et al. (2010), car ils préviennent que l'utilisation d'une SVM comme dispositif de comparaison peut entraîner un positionnement incorrect.

5.3.2 *Comparaison binaire de lisibilité par SVM*

Suivant la méthode de Tanaka-Ishii et al. (2010), le dispositif qui compare deux textes décidant lequel des deux est le plus difficile est une SVM entraînée sur deux deux groupes de textes, l'un de textes faciles et l'autre de textes difficiles. Les textes sont représentés sous forme numérique utilisant les données de notre liste de fréquence lexicale globale issue de corpus créés à partir de

Liste à trier	→	**Liste ordonnée**
FCIEABHGD		
CIEABHGD	F	F
IEABHGD	C	F CF
EABHGD	I	CF C | F CFI
ABHGD	E	CFI C | FI CEFI
BHGD	A	CEFI C | EFI ACEFI
HGD	B	ACEFI AC | EFI A | C | EFI ABCEFI
GD	H	ABCEFI ABC | EFI ABC | EF | I ABCEFHI
D	G	ABCEFHI ABCE | FHI ABCE | F | HI ABCEFGHI
	D	ABCEFGHI ABC | EFGHI AB | C | EFGHI ABCDEFGHI

FIG. 5.1 : Exemple schématique de tri par insertion binaire. On parcourt la liste à trier de gauche à droite. Les comparaisons effectuées entre l'élément pris de la liste à trier et la liste ordonnée sont indiquées en rouge. Si la liste ordonnée est composée d'un nombre pair d'éléments, on choisit celui à gauche. Les barres verticales indiquent où on coupe la liste ordonnée en deux pour cibler l'élément à comparer et la partie restante de la liste qui n'est plus utilisée est grisée.

ressources authentiques pour la création de vecteurs. Avec cette méthode donc, nous n'avons pas besoin d'un corpus annoté en plusieurs niveaux de difficulté, seulement deux niveaux et une liste de fréquence lexicale.

5.3.2.1 Le corpus d'entraînement

Le corpus d'entraînement est composé de deux niveaux. Le niveau *facile* est composé de 90 textes (19 483 tokens) de notre corpus didactique : 17 textes de *Advancing in Burmese : A Drill Book for Intermediate to Advanced Learners* (Yadana Aung 2020), 36 de la *Myanmar Flower Grammar* (Niyomtham et al. 2017), 9 textes du niveau intermédiaire du site Seasite[7], 7 de *Colloquial Burmese* (Hnin Tun et McCormick 2015), 8 du niveau débutant du site Seasite et 13 dialogues du site TUFS (Kawaguchi 2007). Le niveau *difficile* est constitué de 90 textes (180 163 tokens) de nos corpus authentiques et du manuel de lecture avancée composé d'articles de presse authentiques : 32 articles du site BBC Burmese, 28 articles du *Myanmar Newspaper Reader* (Luzoe 1996) et 30 conférences TED Talks.

5.3.2.2 Le calcul de valeurs numériques des textes

Comme Tanaka-Ishii et al. (2010), nous utilisons deux caractéristiques de fréquence lexicale de nos textes pour construire et utiliser le modèle de notre machine à vecteurs de support. Pour chaque *type* d'un texte, nous calculons la *fréquence locale*, la fréquence relative à l'intérieur du texte lui-même :

$$\text{fréquence locale} = \frac{\text{nombre d'occurrences du token dans le texte}}{\text{taille du texte en tokens}}$$

L'autre caractéristique est la *fréquence globale*, la fréquence logarithmique relative de ces éléments dans un grand corpus de référence, c'est-à-dire les valeurs de fréquence de notre liste de fréquence lexicale globale créée selon la méthode décrite dans le chapitre précédent. Puisque la taille de nos corpus de référence utilisés pour créer la liste de fréquence lexicale globale pour le birman est sensiblement plus modeste que celle des corpus japonais et anglais de Tanaka-Ishii et al. (2010), nous avons décidé de ne pas écarter le vocabulaire des textes qui ne figure pas dans la liste. Dans le calcul de la fréquence globale relative au

7. http://www.seasite.niu.edu/burmese/

corpus de référence, nous attribuons à ce vocabulaire la valeur de fréquence la plus petite (1) et nous rajoutons 1 à toutes les autres valeurs. Cette opération, appelée *lissage de Laplace*, est celle recommandée par Brysbaert et Diependaele (2012), qui avertit qu'il faut aussi modifier la taille du corpus en rajoutant le nombre de types au nombre d'occurrences avant de calculer la fréquence relative (et le logarithme). Ceci évite le logarithme de zéro, qui est indéfini. Voici la formule de calcul des fréquences globales :

$$\text{fréquence globale} = \log_{10}\left(\frac{\text{nombre d'occurrences du token dans le corpus} + 1}{\text{taille du corpus en tokens} + \text{nombre de types du corpus}}\right)$$

5.3.3 *Entraînement et usage d'une SVM*

Nous utilisons la bibliothèque Python `scikit-learn` (Pedregosa et al. 2011), une bibliothèque open source d'apprentissage automatique pour créer notre SVM, un algorithme qui apprend à partir de données fournies lors d'une étape d'entraînement. Nous transformons d'abord les textes en vecteurs, et ensuite nous entraînons la SVM. Nous avons utilisé la concaténation pour la création des vecteurs, car Tanaka-Ishii et al. (2010) n'ont trouvé aucune différence significative en utilisant un autre opérateur.

La machine suppose que les vecteurs de deux classes d'entraînement, *facile* et *difficile*, sont séparables par un hyperplan de séparation et apprend son emplacement, maximisant la marge entre les deux classes (Garreta et Moncecchi 2013). Cette SVM entraînée sur nos corpus d'entraînement devient notre outil de comparaison utilisé dans le tri par insertion binaire.

5.3.3.1 Tests avec mini-corpus didactique

Bien que nous ne disposions pas de corpus de textes classifiés finement, et nous n'avons pas d'autre méthode de lisibilité à utiliser comme référence pour évaluer la classification par notre méthode, nous avons tout de même essayé de tester le système sur les seuls textes classifiés dont nous disposons, les leçons du site Seasite (voir le tableau 2.8). Ces textes sont d'abord transformés en vecteurs de la même manière que les textes d'entraînement. Ces petits tests préliminaires

nous ont permis de comparer les performances du modèle de classification à l'intuition d'un instructeur de birman langue étrangère.

Nous avons été assez étonnée par les résultats de notre évaluation, car le rapport généré par la SVM de `scikit-learn` nous fournit toujours un rappel, précision et f-mesure de 1.

Le modèle départage aisément les deux niveaux du site Seasite : Beginner (B) et Intermediate (I), plaçant les textes des leçons dans l'ordre suivant (de plus facile en plus difficile) :

> B06 B03 B05 B10 B08 B17 B13 B01 B02 B04 B09 B16 B20 B14 B18 B11 I09 B19 B15 B23 B12 B22 B21 B07 I03 I01 I02 I05 I14 I11 I06 I13 I12 I10 I04 I08 I07 I15 I16

On remarque que seul le texte de la leçon 9 du niveau intermédiaire se situe parmi les textes du niveau débutant. Il est intéressant de noter qu'il s'agit de la seule leçon de ce niveau au format de dialogue.

Un deuxième test incluant cette fois les textes littéraires du Seasite place ses deux textes de poésie au milieu des textes intermédiaires (entre I06 et I013), et à l'exception du texte L10 မောင်တင့်တယ် /màʊɴtɪ̰ɴtè/ (traduit en anglais comme *The Story of Maung Tint De*) qui est placé avant le dernier texte intermédiaire (I16), tous les textes littéraires se trouvent placés après les textes intermédiaires :

> B06 B03 B05 B10 B08 B17 B13 B01 B02 B04 B09 B16 B20 B14 B18 B11 I09 B19 B15 B23 B12 B22 B21 B07 I03 I01 I02 I05 I14 I11 I06 LITPOE15 LITPOE16 I13 I12 I10 I04 I08 I07 I15 LIT10 I16 LIT11 LIT09 LIT07 LIT12 LIT02 LIT01 LIT05 LIT06 LIT13 LIT08 LIT04 LIT03 LIT14

Nous avons supposé qu'il était très probable que les textes littéraires soient placés vers la fin de la classification, attribuant cela à l'utilisation du style écrit birman dans ces textes, qui n'étaient également présents que dans le niveau difficile des données d'entraînement, le *Myanmar Newspaper Reader*. (Luzoe 1996). Cependant, cela ne peut être la seule explication, car tous les textes ne sont pas écrits dans ce style. Sans un corpus d'entraînement et un corpus de test composés uniquement de style parlé ou de style écrit, il est très difficile de se prononcer. Nous nous heurtons donc à nouveau au problème de la petite taille de corpus disponibles.

5.4 Résumé

Nous avons démontré qu'il est en effet possible d'évaluer la difficulté relative des textes uniquement à travers la fréquence lexicale, une méthode bien adaptée aux langues disposant de peu de ressources en traitement automatique des langues, à condition de disposer d'un corpus de taille suffisante et de disposer des moyens pour que la segmentation en tokens puisse se faire de manière satisfaisante. Cette méthode est cependant difficile à évaluer, car il faudrait disposer d'un corpus didactique plus important pour le faire.

Chapitre 6

Conclusion et perspectives

6.1 Contributions

Cette thèse constitue une contribution significative à l'avancement de la recherche à base de corpus en langue birmane, car elle détaille les enjeux et les méthodes spécifiques au birman qu'il est nécessaire de prendre en compte lors de la création et de l'analyse d'un corpus. Il s'agit de la première tentative d'aborder le traitement automatique du birman dans un but particulier, l'évaluation de la difficulté des textes pour l'étude d'une langue étrangère, en tenant compte des besoins d'un apprenant de la langue. Jusqu'à présent, les travaux dans ce domaine ont été menés par des informaticiens plutôt que par des linguistes et peu d'attention a été accordée à certains des problèmes propres à la langue que nous avons étudiée, en particulier la segmentation et ses conséquences. A cet égard, notre objectif a été d'identifier et d'expliquer les difficultés importantes pour l'amélioration du traitement automatique du birman, et bien que notre méthodologie et nos choix soient souvent spécifiques à notre recherche, beaucoup de nos méthodes sont pertinentes pour la création de corpus birmans en général.

Une particularité des pages précédentes est que nous avons choisi d'aborder notre sujet en nous plaçant du point de vue d'un apprenant en langues qui cherche des explications aux problèmes et aux phénomènes rencontrés, plutôt que de nous appuyer sur une théorie préconçue. Cette perspective nous a conduit à prendre en compte non seulement les observations issues de nos corpus authentiques (dont le contenu a été choisi en tenant compte de l'expérience de l'apprenant), mais aussi la manière dont une langue est présentée par les enseignants et les créateurs de ressources utilisées par les apprenants en langues. D'après nos observations, il semble clair que les enseignants savent intuitivement que les apprenants ont besoin d'un « vocabulaire d'apprenant », quelque chose

qui est difficile à définir, mais qui se rapproche le plus de notre notion de « mot » ou de lexie (qu'elle soit simple ou complexe) que toute autre unité linguistique. Nous avons essayé de concilier ce vocabulaire de l'apprenant avec l'élément de base du TAL, le token, de manière pratique. Une langue étant rarement enseignée séparément de la culture qu'elle véhicule, nous avons également pris en compte le point de vue des Birmans sur leur propre langue et la façon dont elle est utilisée. Ces facteurs semblent souvent négligés en TAL, et nous espérons que notre travail incitera d'autres chercheurs à considérer les langues avec lesquelles ils travaillent en contexte, plutôt que de présumer que toutes les méthodes peuvent être appliquées de manière similaire pour toutes les langues.

Nous avons essayé de présenter notre travail d'une manière moins technique, afin que les questions que nous soulevons puissent être comprises par des linguistes travaillant sur le birman en conjonction avec des informaticiens qui n'ont aucune connaissance de cette langue.

6.1.1 *Corpus*

Nous avons créé deux types de corpus : un corpus de textes didactiques et trois corpus de textes authentiques, un corpus de textes journalistiques, un corpus de traductions de conférences TED Talks et un corpus de textes littéraires. Compte tenu des restrictions en matière de droits d'auteur, aucun de ces documents ne peut être distribué librement, mais ils seront partagés sur demande spécifique à des fins de recherche.

La contribution la plus importante ne concerne cependant pas les corpus eux-mêmes, mais la méthodologie sur la façon de créer et d'analyser un corpus en birman. Nous avons l'intention d'écrire un guide plus concis en anglais sur la création d'un corpus en birman pour une utilisation générale.

6.1.2 *Listes de fréquence*

En utilisant nos corpus, nous avons créé des listes de fréquences lexicales basées sur nos corpus de textes authentiques et une liste de fréquences lexicales globale utilisant les données des trois corpus. Ici encore, la méthodologie utilisée est plus importante que les résultats (qui seront bien sûr partagés sur demande) et nous espérons à l'avenir créer un corpus plus représentatif, en incluant certains corpus du domaine public pour améliorer notre liste de fréquences lexicales

globale.

6.1.3 *Évaluation de textes par fréquence lexicale*

Notre dernière contribution a été l'application d'une méthode d'évaluation de textes adaptée aux langues peu documentées comme le birman qui est assez simple d'un point de vue conceptuel et peu difficile à mettre en œuvre. Comme nous l'avons vu, la validité de l'évaluation de cette méthode pour le birman n'était pas totalement satisfaisante en raison du manque de données, mais nos petits tests démontrent une efficacité certaine.

6.2 Discussion et perspectives

Notre étude repose sur l'hypothèse qu'il existe une la corrélation entre la fréquence des mots et leur importance pour l'apprenant. La familiarité du vocabulaire pour un locuteur natif n'est pas forcément corrélée avec la fréquence et peut aussi constituer un critère important pour l'apprentissage du vocabulaire d'une langue étrangère. Le concept de familiarité n'est pas très bien défini, mais il concerne la perception subjective que les locuteurs ont sur l'importance du vocabulaire dans leur langue. Certains chercheurs ont d'ailleurs mené des expériences psycholinguistiques auprès de larges panels de locuteurs leur demandant d'attribuer un score de familiarité à des mots, aboutissant ainsi à la constitution de listes de familiarité lexicale (Coltheart 1981 ; Nusbaum et al. 1984 ; Amano et Kondo 2000). Dans une étude comparant des listes de fréquence lexicale avec des listes de familiarité lexicale créées à partir de jugements humains pour le japonais et pour l'anglais, Tanaka-Ishii et Terada (2018) ont trouvé une plus forte corrélation entre familiarité et fréquence quand le corpus est de grande taille. Ils ont trouvé que la corrélation augmente de manière log-linéaire jusqu'à une taille de corpus d'un milliard de tokens. Plus important encore, ils ont trouvé une forte corrélation entre la familiarité et la fréquence lexicale lorsque le corpus était constitué de données parlées plutôt qu'écrites. Notre réalisation s'appuie principalement sur des données écrites et sur une quantité relativement faible de textes par rapport à d'autres études de corpus. Cela montre l'importance non seulement de créer un plus grand corpus, mais aussi la nécessité d'inclure plus de données orales.

Par ailleurs, l'enracinement du vocabulaire dans l'esprit de l'apprenant dépend aussi du temps entre expositions (Behrens et Pfänder 2016). Un défaut de notre corpus littéraire est que les textes ne sont pas datés. Nous ne pouvons donc pas considérer tous ces textes comme aussi contemporains que les textes des conférences ou les textes journalistiques. Il est fort possible qu'il y ait un effet de surreprésentation de vocabulaire ou de conventions orthographiques qui ne sont pas actuellement fréquents. Sans une étude diachronique basée sur corpus composé de textes datés précisément, nous ne pouvons pas savoir si cela constitue réellement un problème.

Un autre aspect dont nous avons conscience concerne le fait que notre méthode ne prend pas en compte l'influence interlinguistique de la L1 de l'apprenant dans la lecture de textes en langue étrangère. Cela inclut les noms propres, mais aussi, et c'est plus important, le vocabulaire apparenté à des mots dans des langues que l'apprenant connaît déjà. Une étude de Beinborn et al. (2014) a démontré que les apprenants peuvent déduire le sens des nouveaux mots par leur connaissance d'autres langues. La fréquence L1 des mots empruntés au L1 rencontrés dans la langue étrangère (par exemple les mots anglais en birman) pourrait être utilisée pour modifier la fréquence brute, comme nous l'avons fait pour la dispersion. Il serait donc utile de développer des moyens automatiques d'identifier au moins les mots anglais en birman, une tâche un peu intimidante, étant donné la variation de l'orthographe des mots anglais et étrangers en écriture birmane. Cela nous amène à un point important : devrait-il y avoir une manière standard plus simple d'épeler les mots étrangers en birman ? Ou la souplesse même dans l'orthographe que nous avons observée fait-elle partie de la culture linguistique birmane ? Etant donné que la MLC est actuellement peu active, il est possible que l'orthographe des noms propres sur le Wikipédia birman devienne un standard de fait, agissant comme un dictionnaire des noms propres. Une étude plus approfondie, basée sur des corpus, de la façon dont les mots anglais sont orthographiés en birman et l'élaboration d'un ensemble de recommandations basées sur ces observations aiderait à harmoniser l'orthographe de ces entités et faciliterait ainsi le traitement automatique du birman.

Nous émettons quelques réserves sur certaines caractéristiques de nos corpus de textes authentiques et sur l'inclusion de textes traduits ou potentiellement traduits. Comme nous l'avons déjà évoqué, le birman possède deux registres de langue,

appelés *parlé* et *écrit*, qui ne sont pas totalement distincts, les deux pouvant apparaître dans les textes écrits. Ce serait utile de développer un système de détection automatique de ces deux registres de langue afin d'étiqueter les textes selon le registre et les séparer lors des calculs de fréquence. Une alternative serait d'écarter tout simplement les éléments grammaticaux qui forment les principales caractéristiques des deux styles, mais il serait prudent de partitionner un corpus selon les deux registres d'abord pour mieux étudier et caractériser la différence entre les deux. Nous soupçonnons que la traduction pourrait également avoir un effet sur le style d'écriture et le choix du vocabulaire. Nous n'avons pas eu la possibilité d'étudier le phénomène, mais il nous semble que les textes traduits, notamment les traductions de conférences, utilisent davantage de vocabulaire créé à partir de procédés de dérivation productive que les textes écrits directement en birman. Là encore, cet aspect de la langue mérite plus d'étude, afin de vérifier s'il existe effectivement une différence stylistique significative entre les textes d'un même genre textuel qui ont été rédigés en birman et ceux qui sont traduits.

Une dernière question concerne l'aspect sémantique, plus précisément la prise en compte de la polysémie. Ceci n'est pas nécessairement un problème pour les apprenants, car ils apprendront naturellement les sens les plus courants du vocabulaire en contexte. C'est pour cette raison que nous avons choisi de traiter les homographes de la même manière. Notre décision de nous concentrer sur les lexies plutôt que sur les morphèmes a réduit la quantité d'homographes dans nos listes de fréquences, mais il serait souhaitable de comparer notre méthode de création de listes de fréquences avec des listes prenant en compte l'homographie. La question de la polysémie pourrait être en partie résolue en créant un outil d'étiquetage des parties du discours. Cela créerait bien sûr le besoin d'un jeu d'étiquettes utile et appliqué de manière cohérente avec un guide d'annotation complet, donnant des exemples. Nous pensons que cela devrait être fait en utilisant des exemples de données de corpus pour illustrer l'utilisation réelle, et pas seulement l'intuition du locuteur et les dictionnaires comme cela a été le cas jusqu'à présent. Cela nécessite là encore la constitution d'un corpus plus large, et c'est cela qui devrait désormais être une priorité pour la recherche de corpus en birman.

Annexes

Annexe A

Écriture : translittération et encodage

Tab. A.1 : Système de translittération et points de code Unicode

Translit.	Lettre	Code	Nom Unicode
Consonnes indépendantes			
k	က	U+1000	MYANMAR LETTER KA
k^h	ခ	U+1001	MYANMAR LETTER KHA
g	ဂ	U+1002	MYANMAR LETTER GA
ġ	ဃ	U+1003	MYANMAR LETTER GHA
ŋ	င	U+1004	MYANMAR LETTER NGA
ṡ	စ	U+1005	MYANMAR LETTER CA
$ṡ^h$	ဆ	U+1006	MYANMAR LETTER CHA
ż	ဇ	U+1007	MYANMAR LETTER JA
ẓ	ဈ	U+1008	MYANMAR LETTER JHA
ṅ	ဉ	U+1009	MYANMAR LETTER NYA
ɲ	ည	U+100A	MYANMAR LETTER NNYA
ṫ	ဋ	U+100B	MYANMAR LETTER TTA
ẗ	ဌ	U+100C	MYANMAR LETTER TTHA
ḋ	ဍ	U+100D	MYANMAR LETTER DDA
ḍ	ဎ	U+100E	MYANMAR LETTER DDHA
ṅ	ဏ	U+100F	MYANMAR LETTER NNA
t	တ	U+1010	MYANMAR LETTER TA
t^h	ထ	U+1011	MYANMAR LETTER THA
d	ဒ	U+1012	MYANMAR LETTER DA
ḏ	ဓ	U+1013	MYANMAR LETTER DHA
n	န	U+1014	MYANMAR LETTER NA

Translit.	Lettre	Code	Nom
p	ပ	U+1015	MYANMAR LETTER PA
p^h	ဖ	U+1016	MYANMAR LETTER PHA
b	ဗ	U+1017	MYANMAR LETTER BA
ḅ	ဘ	U+1018	MYANMAR LETTER BHA
m	မ	U+1019	MYANMAR LETTER MA
y	ယ	U+101A	MYANMAR LETTER YA
ṙ	ရ	U+101B	MYANMAR LETTER RA
l	လ	U+101C	MYANMAR LETTER LA
w	ဝ	U+101D	MYANMAR LETTER WA
θ	သ	U+101E	MYANMAR LETTER SA
h	ဟ	U+101F	MYANMAR LETTER HA
l̇	ဠ	U+1020	MYANMAR LETTER LLA
Voyelles indépendantes			
ʔ	အ	U+1021	MYANMAR LETTER A
i̤	ဣ	U+1023	MYANMAR LETTER I
ï̤	ဤ	U+1024	MYANMAR LETTER II
ṳ	ဥ	U+1025	MYANMAR LETTER U
ṳ̈	ဦ	U+1026	MYANMAR LETTER UU
e̤	ဧ	U+1027	MYANMAR LETTER E
o̤	ဩ	U+1029	MYANMAR LETTER O
ö̤	ဪ	U+102A	MYANMAR LETTER AU
Signes vocaliques dépendantes			
ä	◌ါ	U+102B	MYANMAR VOWEL SIGN TALL AA
a	◌ာ	U+102C	MYANMAR VOWEL SIGN AA
i	◌ိ	U+102D	MYANMAR VOWEL SIGN I
ï	◌ီ	U+102E	MYANMAR VOWEL SIGN II
u	◌ု	U+102F	MYANMAR VOWEL SIGN U
ú	◌ူ	U+1030	MYANMAR VOWEL SIGN UU
e	◌ေ	U+1031	MYANMAR VOWEL SIGN E
ɛ	◌ဲ	U+1032	MYANMAR VOWEL SIGN AI
Signes divers			
n̆	◌ံ	U+1036	MYANMAR SIGN ANUSVARA
¹	◌့	U+1037	MYANMAR SIGN DOT BELOW

Translit.	Lettre	Code	Nom
²	◌း	U+1038	MYANMAR SIGN VISARGA
Virama et « signe qui tue »			
ˣ	◌္	U+1039	MYANMAR SIGN VIRAMA
ˣ	◌်	U+103A	MYANMAR SIGN ASAT
Signes consanantiques dépendantes			
ў	◌ျ	U+103B	MYANMAR CONSONANT SIGN MEDIAL YA
ǰ	◌ြ	U+103C	MYANMAR CONSONANT SIGN MEDIAL RA
ŵ	◌ွ	U+103D	MYANMAR CONSONANT MEDIAL WA
ĥ	◌ှ	U+103E	MYANMAR CONSONANT MEDIAL HA
Consonne			
θ:	ဿ	U+103F	MYANMAR LETTER GREAT SA
Chiffres			
0	၀	U+1040	MYANMAR DIGIT ZERO
1	၁	U+1041	MYANMAR DIGIT ONE
2	၂	U+1042	MYANMAR DIGIT TWO
3	၃	U+1043	MYANMAR DIGIT THREE
4	၄	U+1044	MYANMAR DIGIT FOUR
5	၅	U+1045	MYANMAR DIGIT FIVE
6	၆	U+1046	MYANMAR DIGIT SIX
7	၇	U+1047	MYANMAR DIGIT SEVEN
8	၈	U+1048	MYANMAR DIGIT EIGHT
9	၉	U+1049	MYANMAR DIGIT NINE
Ponctuation			
,	၊	U+104A	MYANMAR SIGN LITTLE SECTION
.	။	U+104B	MYANMAR SIGN SECTION
Signes divers			
(loc)	၌	U+104C	MYANMAR SYMBOL LOCATIVE
(sub)	၍	U+104D	MYANMAR SYMBOL COMPLETED
(afore)	၎	U+104E	MYANMAR SYMBOL AFOREMENTIONED
(gen)	၏	U+104F	MYANMAR SYMBOL GENETIVE

Annexe B

Syllabes les plus fréquentes en birman

Rang	Syllabe	f	Rang	Syllabe	f	Rang	Syllabe	f
1	_	667 924	21	ကျွန်	27 112	41	သည်	17 665
2	အ	147 739	22	ဖြစ်	26 420	42	စ	17 391
3	။	111 201	23	ပြီး	25 743	43	လုပ်	17 387
4	က	83 329	24	ဒီ	25 450	44	ခု	16 788
5	ကို	78 580	25	ဟာ	24 376	45	များ	16 126
6	မ	68 627	26	သ	23 988	46	ကြီး	15 895
7	ပါ	68 126	27	တို့	23 824	47	ပြော	15 832
8	တယ်	63 224	28	လို့	23 009	48	ရေး	15 206
9	တွေ	61 473	29	တ	22 737	49	လိုက်	14 517
10	နေ	45 119	30	ရှိ	22 578	50	ပဲ	13 528
11	တဲ့	44 959	31	နိုင်	22 047	51	လေး	13 328
12	မှာ	44 469	32	ရာ	22 040	52	လည်း	13 181
13	တာ	44 116	33	ဆို	21 666	53	ဖို့	13 164
14	၊	40 304	34	လို	20 571	54	.	12 912
15	ရ	35 719	35	လာ	20 388	55	သွား	12 776
16	တော့	29 405	36	မှ	19 235	56	ဘူး	12 566
17	သူ	28 997	37	ဒါ	18 935	57	လူ	12 332
18	နဲ့	28 849	38	တော်	18 866	58	ထဲ	12 282
19	ခဲ့	27 878	39	ကြ	17 996	59	ထား	12 242
20	တစ်	27 629	40	ရဲ့	17 963	60	ပေး	12 148

TAB. B.1 : Les syllabes les plus fréquentes. (Fréquences (f) brutes. Taille corpus en syllabes : 4 263 757 tokens)

Annexe C

Corpus birmans

C.1 Titres des textes du corpus littéraire

Tab. C.1 : Textes du corpus littéraire Lotaya

N°	Titre	Auteur
1	နေရာ၏တိရစ္ဆာန်သဘော	အောင်သူရ
2	သက်တံ့	ပျိုးလက်ဟန်
3	ဗုဒ္ဓဟူးသမီး	နေဝင်းမြင့်
4	တွင်းနက်များ၏ လိင်သဘာဝ	အောင်သူရ
5	အင်နာကရီနား	ပျိုးလက်ဟန်
6	အလင်း၏ ဝိရောဓိ	အောင်သူရ
7	ထိုမြို့	နေမျိုး
8	တချို့က နီပြာပြာ၊ တချို့က ဖြူဝါဝါ၊ တချို့မှာ ချိတ်ပွင့်၊ တချို့မှာ ပန်းရင့်	နေဝင်းမြင့်
9	အပြင်မှာ တကယ်ရှိတဲ့ ဇာတ်ကောင်	ချိုပိန်းနောင်
10	ဖွင့်ဆောင်းမရတဲ့ ထီး	ချိုပိန်းနောင်
11	လွန်ခဲ့တဲ့ အနှစ်တစ်သန်းက ကြယ်တွေ	ပျိုးလက်ဟန်
12	Democratic Jails	ချိုပိန်းနောင်
13	အိမ်နံပါတ် ၁၇	နေမျိုး
14	အင်္ဂေ၏ ဘာသာဗေဒ	အောင်သူရ
15	နိုင်ငံကျော် မင်းသမီးတစ်ဦးနှင့် အိပ်စက်ခြင်း	ချိုပိန်းနောင်
16	မေးရိုး၏ ပြောကြားချက်များ	နေမျိုး
17	လမ်းတို့၏ သားရဲ	အောင်သူရ
18	ကွဲအက်နာရီ	နေဝင်းမြင့်
19	ချိုင်းထောက်	အရိုး
20	မာရ်နတ်	ပျိုးလက်ဟန်

N°	Titre	Auteur
21	စာအုပ်ဆိုင်ကလေး	နေမျိုး
22	Demomorphosis (ဒီမိုမော်ဖိုးဆစ်)	ချိုပိန်းနောင်
23	လပြည့်ည ကဗျာဆရာ	မြင့်သန်း
24	မုန်လာချဉ်ပုလင်းကလေး	နေမျိုး
25	အသွေးအသား၏ နယ်နိမိတ်	အောင်သူရ
26	ထိန်းချုပ်လိုမရတဲ့ အရာတွေ	ပျိုးလက်ဟန်
27	ရုပ်ရှင်ရုံက နီယွန်မီး	ပျိုးလက်ဟန်
28	တတိယမြောက် နည်းပရိယာယ်	တာရာမင်းဝေ
29	အမှောင်ကဖေး	သူဝေး
30	စက္ကူမင်းသား	နေမျိုး
31	မြေခွေး	နေဝင်းမြင့်
32	ပင်လယ်ထဲက မြစ်	သစ္စာနီ
33	ညှပ်စလီ	နေဝင်းမြင့်
34	သစ်ခြောက်ပင်ကိစ္စ	သစ္စာနီ
35	သမီး	ချိုပိန်းနောင်
36	အစက်ကလေး တစ်စက်	မိုးသက်ဟန်
37	လေယာဉ်ပျံ	ပျိုးလက်ဟန်
38	အဘိဓာန်၏ ဆင့်ကဲဖြစ်စဉ်	အောင်သူရ
39	မိုးကုန်ပြီးစ တစ်နေ့၌ ကျေနပ်စွာ လွမ်းဆွတ်ရခြင်း	မြင့်သန်း
40	အပွင့်သဘော၏ ဝေဒနာ	အောင်သူရ
41	ထွန်းထွန်း	ကြည်အေး
42	ညကဖေး	နေမျိုး
43	ဘယားအိတ်	ခင်ခင်ထူး
44	နာရီဟန်ချက် ဥပဒေသ	တာရာမင်းဝေ
45	လမ်းလျှောက်သူ	တေဇာ (လရောင်လမ်း)
46	ဝဲသဖြစီးခြင်း	မိုးသက်ဟန်
47	ကမ္ဘာ	တာရာမင်းဝေ
48	ရေကူး အသင်း	မိုးသက်ဟန်
49	ကမ္ဘာကြီးမှာ တွေးစရာတွေများတယ် ။ ကမ္ဘာကြီးဟာ သိပ်သနားဖို့ကောင်းတယ် ။	ချိုပိန်းနောင်
50	ကျာကျူး	ခင်ခင်ထူး
51	အင်္ဂါပွဲဟ်သားနှင့် ကဗျာဆရာ စကားပြောခြင်း	နေမျိုး
52	တင်းအားတို့၏ မော်တာ	အောင်သူရ

N°	Titre	Auteur
53	နက်(စ်)ကော်ဖီ	နေဝင်းမြင့်
54	မြင့်မြတ်သောကြောင်များ	ချိုပိန်းနောင်
55	လူရွင်တော်အကြောင်း မှတ်စု	နေမျိုး
56	အဝိဇ္ဇာတောလား	သစ္စာနီ
57	သဘော်ကြီးထဲက ရေထည့်ထားတဲ့ ဖန်ခွက်	ချိုပိန်းနောင်
58	သစ္စာမဲ့သူ	ကြည်အေး
59	စိတ္တဇ ရိုစီ	သူဝေး
60	ရွှေမျက်နှာရဲ့ မြအသည်း	သစ္စာနီ
61	အထမမြောက်သော သေကြောင်းကြံစည်မှု	သစ္စာနီ
62	စံပယ်ကို မျှားခြင်း	တေဇာ (လရောင်လမ်း)
63	မိမိကိုယ်ကိုခွေးကျွေးခြင်း	သူဝေး
64	အဖေ	မိုးသက်ဟန်
65	ကျာကျူး (အပိုင်း (၃) ဇာတ်သိမ်း)	ခင်ခင်ထူး
66	အဲဒီနေ့က ကျွန်တော် အရပ်ပိုမြင့်နေခဲ့တယ်	တေဇာ (လရောင်လမ်း)
67	လမ်းဆုံး	မြင့်သန်း
68	နွံအညစ်ဝယ်	ကြည်အေး
69	မေ့ချစ်သူ	ကြည်အေး
70	ကိုသော်တာ	ခင်ခင်ထူး
71	နွေပုံပြင်	တေဇာ (လရောင်လမ်း)
72	ကျွန်တော်၏ ကြောင်ကလေး	တေဇာ (လရောင်လမ်း)
73	ခြေသလုံး ချစ်သူ	သုမောင်
74	ပန်းရဲ့ ရိုက်ချက်	သုမောင်
75	မထူးဆန်းသော ရထား	သူဝေး
76	တကယ့် အချစ်ဝတ္ထု	သစ္စာနီ
77	မိုးရေစက်များ၏ သုည‌တ္တအင်အား	တာရာမင်းဝေ
78	Touch me if you can	အရိုး
79	ဝိုးတဝါး ဘာသာဗေဒ	တာရာမင်းဝေ
80	ပေါ့ပေါ့ပါးပါးပဲ ပြော	မိုးသက်ဟန်
81	ကျွန်တော်နှင့် အသံခုနစ်ထပ်	တာရာမင်းဝေ
82	သနားခါးရေကျဲကျဲ	သုမောင်
83	၂၁ ၏ အချစ်ပုံပြင်	သုမောင်
84	လေကိုသာ အမွေဆက်ခံခွင့်ရရှိသူ	မြင့်သန်း
85	ဘဝသည် ဟာသတစ်ပုဒ်	တေဇာ (လရောင်လမ်း)

N°	Titre	Auteur
86	ပဉ္စရူပ	သစ္စာနီ
87	ဆင်းရဲခြင်းပုံပြင်	ကြည်အေး
88	မှတ်ဉာဏ်ဆိုင်	မိုးသက်ဟန်
89	ပူပါတယ်ဆိုမှ	အရိုး
90	အာဝါဟသဘင်	သုမောင်
91	ဦးထုပ် နှင့် လူ	သစ္စာနီ
92	တိရိစ္ဆာန် သုဘရာဇာ	သူဝေး
93	ဘာစီးနေလဲ	မင်းလူ
94	ခရေပင်လမ်း	တေဇာ (လရောင်လမ်း)
95	ချောင်းမကြီး ရေလျှံတုန်းက သန်းခ ဘာလုပ်နေလဲ	မြင့်သန်း
96	သုံးမျက်နှာ	အရိုး
97	ကျာကျူး (အပိုင်း ၂)	ခင်ခင်ထူး
98	ဆယ့်နှစ်ရာသီ တောလား	သစ္စာနီ
99	စခန်းသာ	ကြည်အေး
100	ပေါင်မုန့်တစ်ဖဲ့	မိုးသက်ဟန်
101	ပိုးတုံးလုံး	နေဝင်းမြင့်
102	ဆွဲငင်ခြင်းနဲ့ တွန်းကန်ခြင်း	ကြည်အေး
103	ရာသီမိုး	ကြည်အေး
104	ကျယ်ကျယ့်ကြယ်	သူဝေး
105	အလှ နှင့် ဘဝ	ကြည်အေး
106	လူမှုနှင်ရင် သိပ္ပံညာရှင်လောက်တော့ ဖြစ်သင့်တဲ့အကြောင်း	မြင့်သန်း
107	ဆေးဆိုး၍ မရပါ	သုမောင်
108	မစဉ်းစားချင်တော့ဘူးဆိုပြီး မစဉ်းစားပဲနေတဲ့အခါ	မိုးသက်ဟန်
109	ပြာဖြစ်သွားသော ပန်းသီး	သစ္စာနီ
110	ကျူး	ခင်ခင်ထူး
111	အနီရောင်နေ့စွဲ	တာရာမင်းဝေ
112	ချစ်-ကြည့်	သုမောင်
113	စက်ရုပ်ဆိုသည်မှာ	အရိုး
114	ကြိုးတန်း	ခင်ခင်ထူး
115	Invention ဆိုသည်မှာ	အရိုး
116	သီတို့ရွာ	ကြည်အေး
117	King of the Writers စာရေးဆရာ	နေမျိုး

N°	Titre	Auteur
118	နွေးသောနှင်း	သုမောင်
119	တန်ပြန်သံသရာ	တေဇာ (လရောင်လမ်း)
120	ဂျော်နီဆိုတဲ့ ကောင်	မိုးသက်ဟန်
121	ခေါင်းလောင်းသံ မကြားလိုက်တဲ့ ကျီးကန်း	တာရာမင်းဝေ
122	ပန်း	မြင့်သန်း
123	လူ၏ ကွန်တိုမျဉ်းများ	သစ္စာနီ
124	သက်တူ ရွယ်တူ	တာရာမင်းဝေ
125	အကြည့်တို့၏ ခန္ဓာ	အောင်သူရ
126	ကိုရိုန် Virus	အရိုး
127	ဝတ်လစ်စားလစ်	ပျိုးလက်ဟန်
128	အရီးနှင့် ကြီးတော်	ခင်ခင်ထူး
129	ကောင်းကင်ကိုဖွင့်မယ့်ကြယ်	မေဇွန်အေး
130	ကားတိုက်မိခြင်း	ပျိုးလက်ဟန်
131	ဒိုင်ယာရီစာအုပ် ပြန်လာပေးသူ	သူဝေး
132	ထိုည	ကြည်အေး
133	ကျည်ခွံတွေကို အိပ်မက်ခြင်း	သူဝေး
134	ကျွန်မနာမည် မပျော်ရွှင်သူကြီး လို့ ခေါ်ပါတယ်	မနော်ဟရီ
135	သပြေနုချိန်	တေဇာ (လရောင်လမ်း)
136	ကျွန်မဘေးကလူ ဘယ်သူလဲ	ခက်မာ
137	မေးခွန်း၏ ဖြစ်ရှိမှု	အောင်သူရ
138	ရထားဘီးနှင့် ခဲလုံးများ	ကြည်အေး
139	ချိုချဉ်	မြင့်သန်း
140	သေဆုံးသွားတဲ့ လျှို့ဝှက်ချက်	တာရာမင်းဝေ
141	ဖြန့်ထားတဲ့ လက်ဝါးကြီး	နေမျိုး
142	သွေးတွေ	သူဝေး
143	ခန္ဓာဗေဒ	မင်းလူ
144	အထီးကျန် တနင်္ဂနွေ	မိုးသက်ဟန်
145	ပိုးကောင်ကလေးကို တောင်းပန်ခြင်း	သူဝေး
146	လနီနီကို မြင်တဲ့လူ	ချိုပိန်းနောင်
147	လူ ၊ လှေ နဲ့ လမင်း	သူဝေး
148	ပိုးဖလံ	ခင်ခင်ထူး
149	ချစ်ခြင်းတွေနဲ့အိမ်	ဧပရယ်နွေ
150	ချစ်သူ ဟောစာတမ်း	မင်းလူ

N°	Titre	Auteur
151	ခြံခုန်ကျော်နွား	လှိုင်းထက်
152	စတုတ္ထမြောက်လက်ချောင်း	ဇူးကျော်
153	စပါယ်သရှယ်လစ်စ်	အရိုး
154	ဆားတစ်တို့	မင်းလူ
155	ဆေးခါးကြီး	ခင်မျိုးချစ်
156	ဆင်ခြေဖုံး	အရိုး
157	ဇာတ်ဝင်ခန်း	မင်းလူ
158	တနင်္ဂနွေများ	မနော်ဟရီ
159	တနင်္ဂနွေရောင် မြူခိုးငွေ့များ	မနော်ဟရီ
160	တက္ကသိုလ်နောက်ခံ ဗယာကြော်ဇာတ်လမ်း	မင်းလူ
161	တစ်ပွင့်တည်းသော ပန်းကုံကော်	ဦးတင့်ဇော်
162	တစ်ယောက်ယောက်	မနော်ဟရီ
163	ဒေါင်လိုက်ဖြတ်တိုက်သွားတဲ့ လေ	အိန္ဒြာ
164	နက္ခတ်တို့၏ အရှင်သခင်	နတ်မှူး
165	နောက်ဆုံးအဖြေ	တက္ကသိုလ်မြစိမ်း
166	နှလုံးသား ၃၈၀/၅၄	မင်းလူ
167	နှုတ်ခမ်း၏ ထောင့်စွန်းများ	မင်းလူ
168	ပန်းစားဘီလူး	နီနီနိုင် (သထုံ)
169	ပန်းဆိုပေမယ့် စက္ကူပန်း	ကျော်မောင် (အညာတက္ကသိုလ်)
170	ပုံတူပန်းချီ	ခင်လှိုင်ကြူ
171	ပယ်မယား တစ်ယောက်သောကား	မင်းလူ
172	ပွင့်သစ်ပန်း	ကြည်အေး
173	ဗိုင်းရပ်(စ်)တစ်ကောင်ရဲ့ အဖျား	နေမျိုး
174	ဘုံကြိုးပြတ်ကြသူများ	မြတ်ကိုကို
175	မိုးတွေရွာနေတဲ့လမ်းကလေး	မနော်ဟရီ
176	မောင့်ကိုယုံပါသည်	သစ်ဦးဇင်
177	မြစ်	မြင့်သန်း
178	ယနေ့ခတ် အချစ်ဝတ္ထုတစ်ပုဒ်	မောင်လွန်းကြင်
179	ရင်ထုမနာ	ပပသင်းစည်
180	ရယ်မောခြင်း အပိုင်းအစများ	မင်းလူ
181	လေဒီရှူးနှင့် ခုံဖိနပ်	မောင်ရင့်မာ ကျောင်းကုန်း
182	လှတစ်မျက်နှာ	ဂျာနယ်ကျော်မမလေး

N°	Titre	Auteur
183	သဘာဝပိုင်ဆိုင်မှု	အရိုး
184	သမီးရဲ့သမိုင်း	စမ်းစမ်းနဲ့
185	သံပရာသီးများ	နေမျိုး
186	အချစ်စစ်ခေါ်သံ	ဝိုင်းယဉ်နွယ်
187	အချစ်ဟု အဓိပ္ပာယ်သက်ရောက်ခြင်းမရှိစေရ	မင်းလူ
188	အစိမ်းရောင်ဇာတ်လမ်း	မနော်ဟရီ
189	အစွဲ	အောင်ကြွေတင့်
190	အပျိုကြီးဖြစ်ရခြင်း အကြောင်း	မစန်းပွင့်
191	အပျိုကြီးဖြစ်ရခြင်းအကြောင်းအရင်း	မစန်းပွင့်
192	အပျောက်အရှုများ	မနော်ဟရီ
193	အဖိုးထိုက်တန်ပစ္စုပ္ပန်	မြကြာငုံ
194	အရောင် အသံ အနံ့ အရိပ်	မအိ
195	အလွမ်းနှင့် ရေးသော ပြဇာတ်	မနော်ဟရီ
196	အဝေးက လက်ဆောင်	ချိုပိန်းနောင်
197	အဲဒီနေ့က ကျွန်မ ခရီးထွက်နေခဲ့တယ်	မနော်ဟရီ
198	အိမ်မြှောင်	အရိုး
199	အိမ်အပြန်	မနော်ဟရီ
200	ဝမ်းမီးစွမ်းအင်	ညီပုလေး
201	ဒိုင်မင်းရှင်းလျောင်အိမ်	အောင်သူရ
202	မြစ်တစ်စင်းကို ရှာဖွေတွေ့ရှိခြင်း	ကျူရီ
203	အလောင်းကောင်ယန္တရား	အောင်သူရ
204	ကျင်းပျောက်သူ	ဂျိုဇော်
204	ပိတောက်ပင်ကို အိပ်မက်မက်တဲ့ ကောင်လေး	တာရာမင်းဝေ
205	ပင်လယ်ထဲခုန်ချတဲ့သူည	တာရာမင်းဝေ
206	ပန်းပျိုးလျက်ပဲနေ	ကိုကြွယ်
207	ဖေဖေ့ သူငယ်ချင်း	ဂျူး
208	တိတ်ဆိတ်ခြင်း၏အသံ	ဂျူး
209	ဘယ်ဘက်လက်ထဲမှာ ဝှက်ထားတဲ့ ပုံပြင်	တာရာမင်းဝေ
210	သစ်တစ်ပင်ကောင်း	ဂျူး
211	အဝေးဆုံး	ခင်လေးမူ
212	အိပ်မက်ကြိုးတံတား	ဂျူး
213	ညမီးကျီး	ဂျူး
214	ခါးအောက်ပိုင်း	မင်းလူ

N°	Titre	Auteur
215	ဂိတ်စ– ဂိတ်ဆုံး	မင်းလူ
216	ဆူးတွေနဲ့လူ	နေမျိုး
217	ပုလင်းချစ်သူ	မင်းလူ
218	မီးပုံးတပ်ထားတဲ့အိမ်	ချိုပိန်းနောင်
219	သမိုင်းဝင် ပထဝီစာအုပ်	နေမျိုး
220	သူတို့တွေ အလွန်ခက်တယ်	မင်းလူ
221	မေ့ကျန်ရစ်ခဲ့သော မျက်နှာ	မင်းလူ
222	အချစ်ကိုပြုပြင်မွမ်းမံခြင်း	မင်းလူ
223	အပေါ်သွေး အောက်သွေး	မင်းလူ
224	ကဗျာမှာပဲငိုနေမယ်	သစ္စာနီ
225	ကျည်ဆံ	မနော်ဟရီ
226	ကြေးရုပ်မြို့တော်	နေမျိုး
227	ကွာခြားခဲ့သည်	ကြည်အေး
228	ကွက်လူကြောင်	နိုင်စွမ်း
229	ကျွန်တော့်ချစ်သူသည် စက်ဘီးမစီးတတ်ပါ	ဆူးငှက်
230	ကျွန်တော့်ဆံပင်တွေ ရှည်တုန်းက	သစ္စာနီ
231	ခင်ပြုံးစေးကအပြန်	မောင်ဖေသန်း
232	ခင်ဗျားဟာ ဧည့်သည်တစ်ယောက်ဖြစ်တယ်	သူဝေး
233	ချစ်သူမရှိတဲ့ကောင်မလေး	ဝင်းပပ (မန်းတက္ကသိုလ်)
234	ချည်သားကုတ်အကျီလေးတစ်ထည်	မိချမ်းဝေ
235	ခွင့်လွှတ်ရန်မဟုတ်ပါ မြနှင်းလှိုင်	သင်းရီ
236	ဂမ္ဘီရ	မြတ်သင်း
237	ဂျာနယ်စလစ်	အကြည်တော်
238	ဂြိုလ်ပြာမစ်ရှင်	မောင်ခိုင်လတ်
239	ငိုသံပြတိုက်	နေမျိုး
240	စကားဖယောင်းနဲ့ မနွဲ့ပျောင်းသောမေတ္တာ	စောမဉ္ဇူ (ဖျာပုံ)
241	စီးကရုတ် မီးခိုးထဲက ထွက်လာတဲ့ ကြောင်	မိုးသက်ဟန်
242	ဆောင်း	နေမျိုး
243	ဇိုးသမားဆုံးမနည်းတစ်မျိုး	အရိုး
245	တစ်ထောင့်ကိုးရာကိုးဆယ့်ခြောက်	မနော်ဟရီ
246	တည်ကောင်	ကျောင်းကုန်းညှဉ်မှူး
244	တောကျီးကန်း	နေမျိုး

N°	Titre	Auteur
247	ထဘီဖာ လိုပါသေးရဲ့၊ တဘက်ဖာ လိုပါသေးရဲ့၊ ဘီးနဲ့မှန် လိုပါသေးရဲ့	နေဝင်းမြင့်
248	ထိုအရပ်သို့ရောက်ဖူးသည်	ကောလ်ထန်မောရိယယ်
249	ဒါအကုန်ပဲလား	နေမျိုး
250	ဒိုးဒိုးဓမ္မတိုးဓဝတိုးဝ	သင်းအောင်ပန်း
251	နှင်းညကြယ်များ	သူဝေး
252	နှစ်ပေါင်းများစွာ	သူဝေး
253	နှာခေါင်းနှစ်ပေါက်နဲ့ လမ်းလျှောက်ခြင်း	မြေလတ်မောင်မြင့်သူ
254	နှုတ်ဆက်သူ	သူဝေး
255	ပတ္တမြားမိုး နှင့် ရေအိုးစင်	သစ္စာနီ
256	ပန်းတိုင်သို့ တက္ကစီဖြင့် သွားခြင်း	စုသဲမွန်
257	ပျော်ပွဲစား	သစ္စာနီ
258	ပြန်လည်ခံစားဂန္တ ဝင်ဝတ္ထုများ	အကြည်တော်
259	ပုလွေသမား	အရိုး
260	ပေးကားပေး၏မရသည့်နှယ်	မာန်မြင့်(ကန့်တလူ)
261	ဖြေသိမ့်တေး	မြတ်လေးမောင်
262	ဗန္ဓတ်	သစ္စာနီ
263	ဘဝ၏ တခြားသောအရာ	အောင်သူရ
264	မမသန်း	ဝင်းစည်သူ
265	မျှော်လင့်ခြင်းလရောင်	သွေး(စစ်ကိုင်း)
266	မြေမနိုင်ရေဆိုင်ဝဲ	ညို နှင်းအိမ်
267	မလုံသည်ကား နှလုံးသားတည်း	မောင်လေးအောင်
268	မှင်စာကလေးရောင်နီ	ကြည်အေး
269	မိုး၏ ဇီဝသီချင်း	အောင်သူရ
270	မိုးသည်းညဲ ကဇွန်းမမှာ	မောင်မောင်လှမြင့်
271	ရွေးချယ်မှုပတ်လမ်းတို	အောင်သူရ
272	ရိုးလှသန္တာ	ရှင်ငြိမ်းမယ်
273	ရေနှင့်လုပ်ထားသော ပျော်ရွှင်မှု	သစ္စာနီ
274	ရေမြေကြောင့် ပန်းတို့အရောင်ပြောင်းကြကုန်၏	အေးကိုသက်
275	ရေအိမ်	နေဝင်းမြင့်
276	လမ်းမတော်သားများ	လေးကိုတင်
277	လွတ်ငြိမ်းချမ်းသာခွင့်	ဗညားလှိုင်ဦး
278	လှေကားပေါ်ကလူ	မိုးသက်ဟန်

N°	Titre	Auteur
279	လိပ်ပြာကူးသောည	ညီသော်လှ
280	လူနှင့် ကမ္ဘာပတ်လမ်း	တာရာမင်းဝေ
281	သတိရစရာ လူတစ်ယောက်	ဆောင်းဝင်းလတ်
282	သရအော်သံ	ညီမင်းညို
283	သားများမိ	မိုးငြိမ်းဇ
284	သီးသန့်ကောင်းကင်	သစ္စာနီ
285	သူမနှင့် ကြွေအိုးဖြူကလေးတစ်လုံး	သူဝေး
286	ဟင်းလင်းပွင့်လျက်အနေအထားအတိုင်း ထိုမြို့	မိုးသက်ဟန်
287	အခန်းထဲမှာ သူမ မရှိဘူး	နေမျိုး
288	အချစ်ရူးမလေး	မတင်သိမ့်
289	အချိုဆုံးအချစ်သီချင်းသစ်သို့	မောင်ညိုပြာ
290	အဝေးက ဆိုက်ပရပ်စ်ပင်	မောင်အောင်ဝင်း
291	အဝေးကမြင်ကွင်း	သင်းသင်းသာ
292	အသက်၏သရုပ်ပြသဘော	အောင်သူရ
293	အသံထွက်၏ဘဝ	အောင်သူရ
294	အားနာခြင်းအမှု	အကြည်တော်
295	အာနိုး နှင့် ကြော်ငြာ	အကြည်တော်

C.2 Titres des textes du corpus TED Talks

Tab. C.2 : Textes du corpus TED Talks

N°	Titre	Langue
1	3 rules to spark learning	anglais
2	3 tips to boost your confidence	anglais
3	3 ways to fix a broken news industry	anglais
4	3 ways to plan for the very long term	anglais
5	3 ways to speak English	anglais
6	4 powerful poems about Parkinsons and growing older	anglais
7	5 dangerous things you should let your kids do	anglais
8	5 techniques to speak any language	anglais
9	5 ways to kill your dreams	anglais
10	5 ways to lead in an era of constant change	anglais
11	8 secrets of success	anglais
12	8 traits of successful people	anglais
13	10 top time-saving tech tips	anglais
14	10 ways to have a better conversation	anglais
15	10-year-olds wisdom-if youre unhappy change yourself	anglais
16	12 truths I learned from life and writing	anglais
17	404 the story of a page not found	anglais
18	1000 TEDTalks 6 words	anglais
19	A 12-year-old app developer	anglais
20	A 50-cent microscope that folds like origami	anglais
21	A beatboxing lesson from a father-daughter duo	anglais
22	A better way to talk about love	anglais
23	A birds-eye view of a breathtaking honeymoon	anglais
24	A brief history of banned numbers	anglais
25	A brief history of numerical systems	anglais
26	A burial practice that nourishes the planet	anglais
27	A cello with many voices	anglais
28	A circle of caring	anglais
29	A dance in a hurricane of paper wind and light	anglais

N°	Titre	Langue
30	A dance to honor Mother Earth	anglais
31	A delightful way to teach kids about computers	anglais
32	A father-daughter bond one photo at a time	anglais
33	A few ways to fix a government	anglais
34	A flying camera on a leash	anglais
35	After your final status update	anglais
36	A garden in my apartment	anglais
37	A glimpse of life on the road	anglais
38	A hilarious celebration of lifelong female friendship	anglais
39	A journey through the mind of an artist	anglais
40	A kids response to spam messages	anglais
41	A kinder gentler philosophy of success	anglais
42	A life lesson from a volunteer firefighter	anglais
43	All it takes is 10 mindful minutes	anglais
44	A look inside the brain in real time	anglais
45	Alzheimers is not normal aging and we can cure it	anglais
46	Am I dying ? The honest answer	anglais
47	A modern take on piano violin cello	anglais
48	A moving song from women in prison for life	anglais
49	A multimedia theatrical adventure	anglais
50	A musical escape into a world of light and color	anglais
51	An 11-year-old prodigy performs old-school jazz	anglais
52	An artists unflinching look at racial violence	anglais
53	An art made of trust vulnerability and connection	anglais
54	And now the real news	anglais
55	An electrifying acoustic guitar performance	anglais
56	An engineers vision for tiny forests everywhere	anglais
57	A new kind of music video	anglais
58	A next-generation digital book	anglais
59	An intergalactic guide to using a defibrillator	anglais
60	An Internet without screens might look like this	anglais
61	An intro to TEDx	anglais
62	An ultra-low-cost college degree	anglais

N°	Titre	Langue
63	An underwater art museum teeming with life	anglais
64	A park underneath the hustle and bustle of New York City	anglais
65	A passionate personal case for education	anglais
66	Aphasia The disorder that makes you lose your words	anglais
67	A precise three-word address for every place on earth	anglais
68	Archaeology from space	anglais
69	Architecting the invisible creating a culture of innovation	anglais
70	A realistic vision for world peace	anglais
71	Are we enough	anglais
72	Are you a body with a mind or a mind with a body	anglais
73	Are you a giver or a taker	anglais
74	Are you human	anglais
75	A roadmap to end aging	anglais
76	A robot that eats pollution	anglais
77	Art made of storms	anglais
78	Art that craves your attention	anglais
79	A sci-fi vision of love from a 318-year-old hologram	anglais
80	A simple birth kit for mothers in the developing world	anglais
81	A simple way to break a bad habit	anglais
82	Asking for help is a strength not a weakness	anglais
83	Ask Why	anglais
84	A smart loan for people with no credit history yet	anglais
85	A song for my hero the woman who rowed into a hurricane	anglais
86	A song inspired by the ocean	anglais
87	A story of mixed emoticons	anglais
88	A summer school kids actually want to attend	anglais
89	A supercharged motorcycle design	anglais
90	A talking squawking parrot	anglais
91	A TED speakers worst nightmare	anglais
92	A teen just trying to figure it out	anglais
93	Auf der Suche nach der verlorenen Zeit (In Search of Lost Time)	allemand
94	Augmented reality techno-magic	anglais

N°	Titre	Langue
95	Autism	anglais
96	A visual history of social dance in 25 moves	anglais
97	A warm embrace that saves lives	anglais
98	A window to our health	anglais
99	Awoo	anglais
100	A word game to communicate in any language	anglais
101	A young guitarist meets his hero	anglais
102	A young inventors plan to recycle Styrofoam	anglais
103	A young scientists quest for clean water	anglais
104	A-Z of unusual animals	anglais
105	Balade dans l'univers à la découverte de Laniakea (Strolling in the Universe : discovering Laniakea)	français
106	Basic living needs	anglais
107	Be an opportunity maker	anglais
108	Beautiful new words to describe obscure emotions	anglais
109	Before I die I want to	anglais
110	Between imagination and reality	anglais
111	Body parts on a chip	anglais
112	Brain chemistry lifehacks	anglais
113	Breaking free from technology	anglais
114	Break the silence for suicide attempt survivors	anglais
115	Breeding Change Cats Culture Conservation	anglais
116	Bridging the digital divide	anglais
117	Bring TED to the classroom with TED-Ed Clubs	anglais
118	Build a tower build a team	anglais
119	Building blocks that blink beep and teach	anglais
120	Can a computer write poetry	anglais
121	Can you find the next number in this sequence	anglais
122	Can you really tell if a kid is lying	anglais
123	Capturing memories in video art	anglais
124	Casting a spell on the cello	anglais
125	Check your intuition The birthday problem	anglais
126	Claim your manspace	anglais

N°	Titre	Langue
127	Clonie	anglais
128	Clues to prehistoric times found in blind cavefish	anglais
129	Color blind or color brave	anglais
130	Compassion and the true meaning of empathy	anglais
131	Could a Saturn moon harbor life	anglais
132	Flow - the secret to happiness	anglais
133	Could we cure HIV with lasers	anglais
134	Courage is contagious	anglais
135	Criticism sticks	anglais
136	Crop insurance an idea worth seeding	anglais
137	Dance tiny robots	anglais
138	Dancing with light	anglais
139	Design at the intersection of technology and biology	anglais
140	Does money make you mean	anglais
141	Does the media have a duty of care	anglais
142	Dog Days Are Over	anglais
143	Doing the impossible cutting through fear	anglais
144	Don't build your home, grow it	anglais
145	Don't eat the marshmallow	anglais
146	Don't fear intelligent machines Work with them	anglais
147	Dont fear superintelligent AI	anglais
148	Don't insist on English	anglais
149	Don't like clickbait. Don't click	anglais
150	Don't suffer from your depression in silence	anglais
151	Do our flaws make us human	anglais
152	Do the green thing	anglais
153	Drawing on humor for change	anglais
154	Drawings that show the beauty and fragility of Earth	anglais
155	Easy DIY projects for kid engineers	anglais
156	Eco-friendly drywall	anglais
157	Educating refugee children by using e-learning tools	anglais
158	Eine kokette Arie (A flirtatious aria)	allemand

N°	Titre	Langue
159	Electrical experiments with plants that count and communicate	anglais
160	Empowering women brings economic growth (in Burmese)	birman
161	Et si nous éduquions nos enfants à la joie (What if we educated our children to joy ?)	français
162	Everybody can transcribe TEDx talks	anglais
163	Everyday leadership	anglais
164	Every kid needs a champion	anglais
165	Falling in love is the easy part	anglais
166	Fashion that celebrates African strength and spirit	anglais
167	Fighting with nonviolence	anglais
168	Finding faith when you're born into two religions	anglais
169	Finding the story inside the painting	anglais
170	Finding your place in the world through the pursuit of your passion	anglais
171	Flow the secret to happiness	anglais
172	Flüchtlinge in Deutschland - das Eis mit Humor brechen (Refugees in Germany - Breaking the ice with humour)	allemand
173	Food and Identity Crisis We are what we eat	anglais
174	Forget multitasking try monotasking	anglais
175	Forget shopping Soon you'll download your new clothes	anglais
176	Forget Wi-Fi Meet the new Li-Fi Internet	anglais
177	For these women reading is a daring act	anglais
178	From womb to world – the journey that shapes our life	anglais
179	Fun fierce and fantastical African art	anglais
180	Gandhi's letter	anglais
181	Get comfortable with being uncomfortable	anglais
182	Get ready for hybrid thinking	anglais
183	Getting rid of 1000 things	anglais
184	Getting smart about learning	anglais
185	Get your next eye exam on a smartphone	anglais
186	Give teachers the chance to help me discover my talents	anglais
187	Global warmings theme song Manhattan in January	anglais
188	Glorious visions in animation and performance	anglais

N°	Titre	Langue
189	Go ahead make up new words	anglais
190	Good news in the fight against pancreatic cancer	anglais
191	Got a meeting Take a walk	anglais
192	Governments dont understand cyber warfare We need hackers	anglais
193	Gravity and the human body	anglais
194	Greek mythology monologue	anglais
195	Greeting the world in peace	anglais
196	Grit The power of passion and perseverance	anglais
197	Grow your own clothes	anglais
198	Guided meditation	anglais
199	Happy maps	anglais
200	Helping young people find their genius	anglais
201	Hidden Memories A Glimpse of History through Old Publications	birman
202	Hidden miracles of the natural world	anglais
203	High-altitude wind energy from kites	anglais
204	High School Training Ground	anglais
205	Home is a song Ive always remembered	anglais
206	Homeless animals	anglais
207	How AI can bring on a second Industrial Revolution	anglais
208	How a penny made me feel like a millionaire	anglais
209	How autism freed me to be myself	anglais
210	How blood pressure works	anglais
211	How blue jeans were invented Moments of Vision 10	anglais
212	How books can open your mind	anglais
213	How Braille was invented Moments of Vision 9	anglais
214	How butterflies self-medicate	anglais
215	How can I help ?	anglais
216	How can technology transform the human body	anglais
217	How coffee got quicker Moments of Vision 2	anglais
218	How data from a crisis text line is saving lives	anglais
219	How did Hitler rise to power ?	anglais
220	How digital transparency can expand your experience	anglais

N°	Titre	Langue
221	How do contraceptives work ?	anglais
222	How do drugs affect the brain ?	anglais
223	How does asthma work ?	anglais
224	How does the brain interpret language ?	anglais
225	How does your body know what time it is ?	anglais
226	How does your body process medicine ?	anglais
227	How do scars form ?	anglais
228	How do we measure distances in space ?	anglais
229	How do you know you exist ?	anglais
230	How fake news does real harm	anglais
231	How free is our freedom of the press	anglais
232	How germs travel on planes – and how we can stop them	anglais
233	How high can you count on your fingers ? Spoiler much higher than 10	anglais
234	How I beat stage fright	anglais
235	How I became an entrepreneur at 66	anglais
236	How I brought a river and my city back to life	anglais
237	How I built a jet suit	anglais
238	How I defend the rule of law	anglais
239	How I fail at being disabled	anglais
240	How I fell in love with quasars blazars and our incredible universe	anglais
241	How I found myself through music	anglais
242	How I learned to communicate my inner life with Aspergers	anglais
243	How I learned to read – and trade stocks – in prison	anglais
244	How Im preparing to get Alzheimers	anglais
245	How interpreters juggle two languages at once	anglais
246	How I teach kids to love science	anglais
247	How I turned a deadly plant into a thriving business	anglais
248	How language generates your world and mine	anglais
249	How meditation can reshape our brains	anglais
250	How my dads dementia changed my idea of death and life	anglais
251	How new technology helps blind people explore the world	anglais
252	How optical illusions trick your brain	anglais

N°	Titre	Langue
253	How out-of-body experiences could transform yourself	anglais
254	How playing an instrument benefits your brain	anglais
255	How playing sports benefits your body and your brain	anglais
256	How school makes kids less intelligent	anglais
257	How small are we in the scale of the universe	anglais
258	How smudge-proof lipstick was invented Moments of Vision 6	anglais
259	How super glue was invented Moments of Vision 8	anglais
260	How the Band-Aid was invented Moments of Vision 3	anglais
261	How the bendy straw was invented Moments of Vision 12	anglais
262	How the Internet enables intimacy	anglais
263	How the jump rope got its rhythm	anglais
264	How the news distorts our worldview	anglais
265	How the popsicle was invented Moments of Vision 11	anglais
266	How the rubber glove was invented Moments of Vision 4	anglais
267	How the sandwich was invented Moments of Vision 5	anglais
268	How the stethoscope was invented Moments of Vision 7	anglais
269	How to avoid surveillance with the phone in your pocket	anglais
270	How to be a hero on the web	anglais
271	How to become a millionaire in 3 years	anglais
272	How to build an information time machine	anglais
273	How to buy happiness	anglais
274	How to control someone else's arm with your brain	anglais
275	How to find a TEDx talk to transcribe	anglais
276	How to find meaning when reality hits you	anglais
277	How to gain control of your free time	anglais
278	How to get a new hip	anglais
279	How to get back to work after a career break	anglais
280	How to get better at the things you care about	anglais
281	How to grow a forest in your backyard	anglais
282	How to grow fresh air	anglais
283	How to inspire every child to be a lifelong reader	anglais
284	How to know your life purpose in 5 minutes	anglais

N°	Titre	Langue
285	How to learn From mistakes	anglais
286	How to let altruism be your guide	anglais
287	How to live passionatelyno matter your age	anglais
288	How to make a profit while making a difference	anglais
289	How to make peace Get angry	anglais
290	How to make work-life balance work	anglais
291	How to manage your time more effectively according to machines	anglais
292	How to measure success of international development projects	anglais
293	How to practice effectively - for just about anything	anglais
294	How to raise a black son in America	anglais
295	How to raise successful kids – without over-parenting	anglais
296	How to save the world or at least yourself from bad meetings	anglais
297	How to speak so that people want to listen	anglais
298	How to start a movement	anglais
299	How to stay calm when you know youll be stressed	anglais
300	How to succeed : Get more sleep	anglais
301	How to tie your shoes	anglais
302	How to tutor a billion students	anglais
303	How to use one paper towel	anglais
304	How to visualize one part per million	anglais
305	How to win friends and influence people	anglais
306	How trees talk to each other	anglais
307	How well find life on other planets	anglais
308	How women wage conflict without violence	anglais
309	How yarn bombing grew into a worldwide movement	anglais
310	How YouTube thinks about copyright	anglais
311	Hunting for Peru's lost civilizations – with satellites	anglais
312	Ideas worth dating	anglais
313	If I controlled the Internet	anglais
314	I like to think negatively (Burmese)	anglais
315	Im not your inspiration thank you very much	anglais

N°	Titre	Langue
316	Infinity explained in 3 minutes	anglais
317	Innovation inspiration infrastructure three keys for rethinking education	anglais
318	In praise of macro – yes macro – finance in Africa	anglais
319	In search of language barriers	anglais
320	Inside the mind of a master procrastinator	anglais
321	Introducing TED-Ed Lessons Worth Sharing	anglais
322	Inventing is the easy part Marketing takes work	anglais
323	Is DNA the future of data storage	anglais
324	Is telekinesis real	anglais
325	Is there a real you	anglais
326	It's TED the Musical	anglais
327	It's time for The Talk	anglais
328	It's up to you to make it a perfect day	anglais
329	KAir battery - an electrical Inventory	anglais
330	Keep your goals to yourself	anglais
331	Kiteflyers' Hill	anglais
332	Knowledge fuels success	anglais
333	Know your worth and then ask for it	anglais
334	Language beyond rhythm	anglais
335	La Vie en Rose	anglais
336	Learning From Innovators in Every Sector	anglais
337	Learning to breathe	anglais
338	Learning to code coding to learn	anglais
339	Learn to read Chinese with ease	anglais
340	Lessons from a solar storm chaser	anglais
341	Less stuff more happiness	anglais
342	Let's change math education	anglais
343	Let's change the world one poem at a time	anglais
344	Let's end ageism	anglais
345	Let's not use Mars as a backup planet	anglais
346	Let's revive the Golden Rule	anglais
347	Let's teach for mastery – not test scores	anglais

N°	Titre	Langue
348	Life is your talents discovered	anglais
349	Life lessons through tinkering	anglais
350	Lifesaving scientific tools made of paper	anglais
351	Lifes third act	anglais
352	Living beyond limits	anglais
353	Living bricks : how heritage buildings enhance social fabric	anglais
354	Looking for a job ? Highlight your ability not your experience	anglais
355	Looks aren't everything. Believe me I'm a model	anglais
356	Love Is a Loaded Pistol	anglais
357	Love letters to strangers	anglais
358	Magical houses made of bamboo	anglais
359	Making a terrorist	anglais
360	Making sense of spelling	anglais
361	Maps that show us who we are not just where we are	anglais
362	Mark Stewarts To Whom It May Concern Thank You	anglais
363	Math isnt hard its just a language	anglais
364	Meditation as medicine	anglais
365	Meet a young entrepreneur cartoonist designer activist	anglais
366	Meeting people is the new form of learning	anglais
367	Meet the microscopic life in your home – and on your face	anglais
368	Mentoring and knowledge sharing	anglais
369	Mind-blowing magnified portraits of insects	anglais
370	Mind your language	anglais
371	Mixed reality in higher education	anglais
372	More than words	anglais
373	Mother of Pearl If I Had You	anglais
374	Music is medicine music is sanity	anglais
375	Musiques et langues - un mariage de saveurs	anglais
376	My DNA vending machine	anglais
377	My Fine Reward	anglais
378	My invention that made peace with lions	anglais
379	My library of human imagination	anglais

N°	Titre	Langue
380	My magic moves	anglais
381	my mama BLACK BANANA (Songs)	anglais
382	My mushroom burial suit	anglais
383	My simple invention designed to keep my grandfather safe	anglais
384	My trek to the South Pole	anglais
385	My underwater robot	anglais
386	My wish To launch a new era of openness in business	anglais
387	My year of saying yes to everything	anglais
388	My year reading a book from every country in the world	anglais
389	”Natural Woman” and ”Johnny and Donna”	anglais
390	Nature Beauty Gratitude	anglais
391	New moment new opportunity	anglais
392	No body’s perfect teenage eating disorders	anglais
393	”(Nothing But) Flowers” with string quartet	anglais
394	Obstacle courses teach life lessons	anglais
395	Old books reborn as art	anglais
396	Older people are happier	anglais
397	On being a woman and a diplomat	anglais
398	One child too many	anglais
399	One language one people	anglais
400	One Laptop per Child	anglais
401	One more reason to get a good nights sleep	anglais
402	One of the most difficult words to translate	anglais
403	On tennis love and motherhood	anglais
404	On violin and cello	anglais
405	Opening up to kindness	anglais
406	Open-sourced blueprints for civilization	anglais
407	OTP Learning Series 04 How to translate	anglais
408	OTP Learning Series 08 How to tackle reading-speed issues	anglais
409	OTP Learning Series 09 How to edit titles and descriptions	anglais
410	OTP volunteer talks about why transcribing TEDx talks is important	anglais
411	Our antisocial phone tricks	anglais

N°	Titre	Langue
412	Our campaign to ban plastic bags in Bali	anglais
413	Our human capacity for language insights from signed languages	anglais
414	Our natural sleep cycle is nothing like what we do now	anglais
415	Our responsibility to share	anglais
416	Paper beats plastic. How to rethink environmental folklore	anglais
417	Para entender el autismo no quites la mirada	espangol
418	Peace on Earth	anglais
419	Perché lunico futuro degno di noi dovrà includere tutti (Why the only future worth building includes everyone)	italien
420	Philosophy in prison	anglais
421	Photos from a storm chaser	anglais
422	Photos that changed the world	anglais
423	Photos that give voice to the animal kingdom	anglais
424	Pirates nurses and other rebel designers	anglais
425	Playing Pink Noise on guitar	anglais
426	Please please people… Let's put the awe back in awesome	anglais
427	Poems of war peace women power	anglais
428	Politics and the English language	anglais
429	Portraits that transform people into whatever they want to be	anglais
430	Print your own medicine	anglais
431	Profits not always the point	anglais
432	Programming bacteria to detect cancer and maybe treat it	anglais
433	Psychedelic science	anglais
434	Queen's Bohemian Rhapsody performed in sign language	anglais
435	Rebuilding Myanmar One Child Labourer at a Time	birman
436	Redemption Song	anglais
437	Remember to say thank you	anglais
438	Reporting crisis via texting	anglais
439	Rollercoaster	anglais
440	Science in service to the public good	anglais
441	Seasons of Love	anglais
442	Secondary sugar kills	anglais

N°	Titre	Langue
443	See invisible motion hear silent sounds	anglais
444	See your stories played back at you	anglais
445	Shape-shifting tech will change work as we know it	anglais
446	Should you donate differently	anglais
447	Should you live for your résumé or your eulogy	anglais
448	Simple hacks for life with Parkinson's	anglais
449	Simplifying complexity	anglais
450	Singing What I Want	anglais
451	Six reasons why you should be more trustful	anglais
452	Slow down take care	anglais
453	Songs that bring history to life	anglais
454	Space Oddity	anglais
455	Speak to the heart	anglais
456	Sputnik mania	anglais
457	St James Infirmary Blues	anglais
458	Street art with a message of hope and peace	anglais
459	Stress at school	anglais
460	Success failure and the drive to keep creating	anglais
461	Success is a continuous journey	anglais
462	Symbiosis A surprising tale of species cooperation	anglais
463	Synthetic voices as unique as fingerprints	anglais
464	Taking chances to reach students	anglais
465	Talk about your death while youre still healthy	anglais
466	Talk nerdy to me	anglais
467	Teachers need real feedback	anglais
468	Teach girls bravery not perfection	anglais
469	Teach statistics before calculus	anglais
470	Teach teachers how to create magic	anglais
471	TED-Ed Clubs Celebrating and amplifying student voices around the world	anglais
472	TED-Ed Clubs presents TED-Ed Weekend	anglais
473	TEDs secret to great public speaking	anglais
474	TED Translators Come Join Us	anglais

N°	Titre	Langue
475	TEDxChange Pilot Program	anglais
476	TEDx Introduction	anglais
477	Draw Your Future	anglais
478	Texting that saves lives	anglais
479	The 3 agencies with the power to make or break economies	anglais
480	The 41 work week	anglais
481	The 100000-student classroom	anglais
482	The agony of trying to unsubscribe	anglais
483	The ancestor of language	anglais
484	The antidote to apathy	anglais
485	The Art of Changing Minds	anglais
486	The art of stillness	anglais
487	The balancing act of compassion	anglais
488	The beauty of being a misfit	anglais
489	The beauty of human skin in every color	anglais
490	The beauty of what well never know	anglais
491	The benefits of a bilingual brain	anglais
492	The best computer interface ? Maybe your hands	anglais
493	The best gift I ever survived	anglais
494	The bird song	anglais
495	The birth of Wikipedia	anglais
496	The boost students need to overcome obstacles	anglais
497	The Chinese zodiac explained	anglais
498	The curious evidence of the unlimited brain	anglais
499	The dancer, the singer, the cellist… and a moment of creative magic	anglais
500	The danger of a single story	anglais
501	The danger of hiding who you are	anglais
502	The danger of silence	anglais
503	The day I stood up alone	anglais
504	The demise of guys	anglais
505	The discovery that could rewrite physics	anglais
506	The Dog Song	anglais

N°	Titre	Langue
507	The electric rise and fall of Nikola Tesla	anglais
508	The fascinating secret lives of giant clams	anglais
509	The first 21 days of a bees life	anglais
510	The forgotten art of the zoetrope	anglais
511	The future of good food in China	anglais
512	The future of money	anglais
513	The future of news Virtual reality	anglais
514	The gift of language spoken clearly	anglais
515	The happy secret to better work	anglais
516	The hidden beauty of pollination	anglais
517	The hidden opportunities of the informal economy	anglais
518	The hidden power of smiling	anglais
519	The history of our world in 18 minutes	anglais
520	The history of tea	anglais
521	The jobs well lose to machines – and the ones we wont	anglais
522	The kids are reading lets join them	anglais
523	The killer American diet thats sweeping the planet	anglais
524	The left brain vs right brain myth	anglais
525	The long reach of reason	anglais
526	The lost art of letter-writing	anglais
527	The magic ingredient that brings Pixar movies to life	anglais
528	The magic of Fibonacci numbers	anglais
529	The magic of Khmer classical dance	anglais
530	The magic of words what we speak is what we create	anglais
531	The military case for sharing knowledge	anglais
532	The most important lesson from 83000 brain scans	anglais
533	The most Martian place on Earth	anglais
534	The music wars	anglais
535	The mysterious world of underwater caves	anglais
536	The need for early second language exposure	anglais
537	The next manufacturing revolution is here	anglais
538	The next outbreak. We're not ready	anglais
539	The next step in nanotechnology	anglais

N°	Titre	Langue
540	The nit-picking glory of The New Yorkers Comma Queen	anglais
541	The only wrong you can do is not doing anything	anglais
542	Theory of success	anglais
543	The Panama Papers exposed a huge global problem. What's next ?	anglais
544	The paradox of value	anglais
545	The passing of time caught in a single photo	anglais
546	The philosophy of Stoicism	anglais
547	The playful wonderland behind great inventions	anglais
548	The power of believing that you can improve	anglais
549	The power of curiosity	anglais
550	The power of listening	anglais
551	The Power of Simple Words	anglais
552	The power of song - inspiring the potential for peace	anglais
553	The price of shame	anglais
554	The prison of your mind	anglais
555	The psychology of time	anglais
556	The psychology of your future self	anglais
557	The quantified self	anglais
558	The reality of Luck	anglais
559	The refugees of boom-and-bust	anglais
560	The reporting system that sexual assault survivors want	anglais
561	Theres a better way to die and architecture can help	anglais
562	Theres more to life than being happy	anglais
563	The row-bot that feeds on pollution	anglais
564	The science of spiciness	anglais
565	The search for the true face of Leonardo	anglais
566	The secret to effective nonviolent resistance	anglais
567	The shocking move to criminalize nonviolent protest	anglais
568	The single biggest reason why startups succeed	anglais
569	The sore problem of prosthetic limbs	anglais
570	The sound of silence	anglais
571	The spellbinding art of human anatomy	anglais

N°	Titre	Langue
572	The Stages of the Ages	anglais
573	The sublime beauty of Indian ragas	anglais
574	The Sun continent powered by the Sun	anglais
575	The surprising habits of original thinkers	anglais
576	The surprisingly charming science of your gut	anglais
577	The technology of storytelling	anglais
578	The unexpected beauty of everyday sounds	anglais
579	The US needs paid family leave – for the sake of its future	anglais
580	The Value of Time - The importance of Childhood	anglais
581	The violin and my dark night of the soul	anglais
582	The way we think about work is broken	anglais
583	The weirdness of water could be the answer	anglais
584	The wonder of Zulu wire art	anglais
585	The world's English mania	anglais
586	The worlds most mysterious book	anglais
587	The writer's block	anglais
588	Things I've learned in my life so far	anglais
589	This app knows how you feel – from the look on your face	anglais
590	This app makes it fun to pick up litter	anglais
591	This is what it's like to go undercover in North Korea	anglais
592	This scientist makes ears out of apples	anglais
593	Three Dirty Secrets of Happiness and Health	anglais
594	To hear this music you have to be there. Literally.	anglais
595	Touchscreen	anglais
596	Tough decisions use your heart	anglais
597	Trust in animals	anglais
598	Try something new for 30 days	anglais
599	Turning trash into toys for learning	anglais
600	Two poems about what dogs think probably	anglais
601	Two reasons companies fail – and how to avoid them	anglais
602	Txtng is killing language JK	anglais
603	Underwater astonishments	anglais
604	Vipassana meditation and body sensation	anglais

N°	Titre	Langue
605	Visualizing hidden worlds inside your body	anglais
606	Wait It Out	anglais
607	Want to be an activist ? Start with your toys	anglais
608	Want to be happier ? Stay in the moment	anglais
609	Want to be happy ? Be grateful	anglais
610	Want to be more creative Go for a walk	anglais
611	Waste management	anglais
612	Wearing nothing new	anglais
613	We can reprogram life. How to do it wisely	anglais
614	Weird or just different	anglais
615	What a driverless world could look like	anglais
616	What ants teach us about the brain cancer and the Internet	anglais
617	What are those floaty things in your eye ?	anglais
618	What can we learn from shortcuts	anglais
619	What causes economic bubbles ?	anglais
620	What causes kidney stones ?	anglais
621	What do we do when antibiotics dont work any more	anglais
622	What happens in your brain when you pay attention	anglais
623	What happens when a city runs out of room for its dead	anglais
624	What happens when our computers get smarter than we are	anglais
625	What happens when you have a disease doctors cant diagnose	anglais
626	What happens when you lose everything	anglais
627	What if 3D printing was 100x faster ?	anglais
628	What if age is just a state of mind ?	anglais
629	What I learned from 2000 obituaries	anglais
630	What is depression ?	anglais
631	What is Math About ?	anglais
632	What its like to be Muslim in America	anglais
633	What it takes to be a great leader	anglais
634	What makes a good life Lessons from the longest study on happiness	anglais
635	Whats next in 3D printing	anglais

N°	Titre	Langue
636	What squatter cities can teach us	anglais
637	What's the difference between accuracy and precision ?	anglais
638	What's the value of vitamins	anglais
639	What's wrong with our food system ?	anglais
640	What's your Rubik's cube ?	anglais
641	What teen pregnancy looks like in Latin America	anglais
642	What the discovery of gravitational waves means	anglais
643	What we don't know about mothers milk	anglais
644	What we know and dont know about Ebola	anglais
645	What were missing in the debate about immigration	anglais
646	What will be the next big scientific breakthrough	anglais
647	What will future jobs look like	anglais
648	What will you tell your daughters about 2016	anglais
649	What would happen if you didnt drink water	anglais
650	What would happen if you didn't sleep	anglais
651	What you dont know about marriage	anglais
652	What Youve Got	anglais
653	Where do new words come from	anglais
654	Where do superstitions come from	anglais
655	Where is home	anglais
656	Who am I Think again	anglais
657	Why are we so attached to our things	anglais
658	Why design should include everyone	anglais
659	Why do airlines sell too many tickets	anglais
660	Why do buildings fall in earthquakes	anglais
661	Why does it take so long to grow up today	anglais
662	Why do I make art To build time capsules for my heritage	anglais
663	Why do people get so anxious about math	anglais
664	Why do we cry. The three types of tears	anglais
665	Why do we feel like we need relationships to be happy	anglais
666	Why do we feel nostalgia	anglais
667	Why do we love A philosophical inquiry	anglais
668	Why genetic research must be more diverse	anglais

N°	Titre	Langue
669	Why good leaders make you feel safe	anglais
670	Why Google Glass	anglais
671	Why I believe the mistreatment of women is the number one human rights abuse	anglais
672	Why I keep speaking up even when people mock my accent	anglais
673	Why I live in mortal dread of public speaking	anglais
674	Why I make robots the size of a grain of rice	anglais
675	Why I'm a weekday vegetarian	anglais
676	Why is glass transparent	anglais
677	Why I speak up about living with epilepsy	anglais
678	Why I still have hope for coral reefs	anglais
679	Why is Vermeer's Girl with the Pearl Earring considered a masterpiece	anglais
680	Why is x the unknown	anglais
681	Why I take the piano on the road and in the air	anglais
682	Why lunch ladies are heroes	anglais
683	Why open a school ? To close a prison	anglais
684	Why our screens make us less happy	anglais
685	Why should you read Tolstoy's War and Peace	anglais
686	Why the best hire might not have the perfect resume	anglais
687	Why the metric system matters	anglais
688	Why we laugh	anglais
689	Why we need crazy ideas	anglais
690	Why we should build wooden skyscrapers	anglais
691	Why you should know how much your coworkers get paid	anglais
692	Why you should talk to strangers	anglais
693	Why you think youre right – even if youre wrong	anglais
694	Wireless data from every light bulb	anglais
695	Wisdom from great writers on every year of life	anglais
696	Without Geography We Are Nowhere	anglais
697	Women of Hope	anglais
698	World peace starts with ourselves	anglais
699	Would winning the lottery make you happier	anglais

N°	Titre	Langue
700	Wry photos that turn stereotypes upside down	anglais
701	Yoga and meditation subtle practices for change	anglais
702	You raise me up	anglais
703	Your body is my canvas	anglais
704	Your body language may shape who you are	anglais
705	Your elusive creative genius	anglais
706	Your genes are not your fate	anglais
707	Your smartphone is a civil rights issue	anglais
708	You smell with your body not just your nose	anglais
709	Yup I built a nuclear fusion reactor	anglais
710	От слов к поступкам (From words to deeds)	russe
711	Ծաղրի շղթան (The chain of mocking)	arménien
712	定年後は社会へ恩返しわたしの自由時間の使い方 (Repaying society after retirement — how to use my free time)	japonais
713	思うは招く(Hope invites)	japonais

Glossaire

1SG première personne du singulier 89, 96, 173
2SG deuxième personne du singulier 96
IR irréalis 67
DISPL déplacement espace temps 101
NEG négation 137
NFUT non futur 89, 101
NOMR nominalisateur 136
OBJ complément d'objet direct 96
PL pluriel 115
POSS marqueur de possession 83, 97, 98
PTC particule 136, 137, 178
V verbe 67, 136, 142

ALT Asian Language Treebank 49, 51, 131
ASCII norme informatique de codage de caractères latins de base (lettres non accentuées en majuscule ou minuscule, chiffres, ponctuation) permettant à l'origine de représenter des textes anglais, American Standard Code for Information Interchange (*Code américain normalisé pour l'échange d'information*) 30, 32, 33, 35, 55, 74

BLE Le birman comme langue étrangère 4, 6, 7, 207

caractère unité minimale distinctive d'un système d'écriture. En Unicode, un caractère peut être représenté par des glyphes multiples 14
CECR Cadre européen commun de référence pour les langues 185, 221
Creative Commons organisation non gouvernementale internationale qui fournit gratuitement au public des licences de droits de propriété intellectuelle standardisées 49, 51
CRF Conditional Random Field, *champ aléatoire conditionnel* 88, 89

DFXP Distribution Format Exchange Profile, format XML pour la synchronisation de texte avec d'autres médias comme l'audio et la vidéo, appelé aussi TTML 55

FRM fréquence réduite moyenne, mesure qui combine fréquence et dispersion 201

glyphe forme graphique d'un caractère, affichée ou imprimée, instanciée par une police de caractères donnée 15, 41

hànyǔ pīnyīn 汉语拼音 système officiel de transcription du chinois mandarin standard x, 115

L1 langue maternelle 4, 186, 234

L2 deuxième langue, langue apprise après la langue maternelle, dans un contexte d'usage quotidien de la L2. 4

ligature une combinaison de caractères adjacents se combinant pour former un glyphe unique 31

matra caractère vocalique associé à une voyelle dépendant, autre que la voyelle inhérente d'un caractère consonantique 25

MED Myanmar-English Dictionary မြန်မာ-အင်္ဂလိပ် အဘိဓာန် (Myanmar Language Commission 1993) 7, 117, 118, 124, 125, 130–134, 312, 316

MLC Myanmar Language Commission 3, 50, 78, 81, 147, 234

MLT Myanmar Language Test 8

MNA Myanmar News Agency 156

mojibake texte brouillé ou altéré résultant d'un mauvais décodage 36

NICT National Institute of Information & Communications Technology, Japon 49

NLD National League for Democracy အမျိုးသား ဒီမိုကရေစီ အဖွဲ့ချုပ်. La *Ligue nationale pour la démocratie*, le parti politique de Aung San Suu Kyi 156

Perl langage de programmation informatique 53, 54, 64, 75–77

point de code valeur numérique codant de façon unique un caractère dans le standard Unicode 14, 17, 41, 116

police police de caractères ou police d'écriture. Ensemble de glyphes d'un même style typographique qui permet la représentation visuelle de caractères 15, 41

sandhi modifications dans la prononciation de syllabes liées x

scriptio continua style d'écriture sans marques de séparation entre les mots 10, 85

SVM Support Vector Machine (*machine à vecteurs de support*) 88, 89, 222, 224, 227

TAL Traitement Automatique des Langues i, 45, 100, 232, 281

taliste informaticien(ne) spécialisé(e) dans le domaine du TAL 29, 45

TED Technology, Entertainment and Design 55

Treebank corpus textuel analysé qui comporte des annotations syntaxiques 49

TTML Timed Text Markup Language, norme du W3C pour la synchronisation de texte avec d'autres supports comme l'audio et la vidéo 280

TUFS Tokyo University of Foreign Studies 62, 281

UBLI Center of Usage-Based Linguistic Informatics, TUFS 62

UCSY University of Computer Studies, Yangon, Birmanie 49

URI Uniform Ressource Identifier (*Identifiant de ressource uniforme*) 75, 76

URL Uniform Resource Locator (*localisateur uniforme de ressource*), l'adresse d'une page hypertexte sur Internet 50

W3C World Wide Web Consortium, organisme de standardisation des technologies du World Wide Web 281

Bibliographie

Aksan, Yesim, Mustafa Aksan et Ümit Mersinli (2017). *A frequency Dictionary of Turkish*. Routledge frequency dictionaries. London : Routledge.

Alderson, J Charles (2005). *Diagnosing foreign language proficiency : The interface between learning and assessment*. A&C Black.

– (sept. 2007). « Judging the Frequency of English Words ». In : *Applied Linguistics* 28.3, p. 383-409. DOI : 10.1093/applin/amm024.

Allott, Anna (1985). « Language policy and language planning in Burma ». In : *Papers in Southeast Asian Linguistics No. 9 : Language Policy, Language Planning and Sociolinguistics in Southeast Asia*. Sous la dir. de David Bradley. Camberra.

Allott, Anna, Patricia Herbert et John Okell (1989). « Burma ». In : *South-East Asia Languages and Literatures : a Select Guide*. Sous la dir. de Patricia Herbert et Anthony Milner. ISBN : 978-0824812676.

Amano, Shigeaki et T Kondo (2000). « On the NTT psycholinguistic databases 'lexical properties of Japanese' ». In : *Journal of the Phonetic Society of Japan* 4.2, p. 44-50.

American Library Association, Library of Congress, éd. (2011). *ALA-LC Romanization Tables - Burmese*.

Aroonmanakun, Wirote (2002). « Collocation and Thai Word Segmentation ». In : *Proceedings of the Fifth Symposium on Natural Language Processing and The Fifth Oriental COCOSDA Workshop*.

– (2007). « Thoughts on Word and Sentence Segmentation in Thai ». In : *Proceedings of the Seventh Symposium on Natural Language Processing, Dec 13-15*. Pattaya, Thailand.

Aung Kaung Myat (27 mai 2017). *Facebook is banning the derogatory word, kalar, but not in the way you expect*. URL : https://medium.com/@aungkaungmyat/facebook-is-banning-the-derogatory-word-kalar-but-not-in-the-way-you-expect-1ec1ea7dcb44 (visité le 17/12/2019).

Avdyli, Dr.Sc. Rrezarta et Dr.Sc. Fernando Cuetos (juin 2013). « SUBTLEX- AL : Albanian word frequencies based on film subtitles ». In : *ILIRIA International Review* 3.1, p. 285. DOI : 10.21113/iir.v3i1.112.

Aye Myat Mon et Thandar Thein (jan. 2013). « Myanmar Spell Checker ». In : *International journal of Science and Research (IJSR) India* 2.1.

Aye Nyein Mon, Win Pa Pa, Ye Kyaw Thu et Yoshinori Sagisakaa (nov. 2017). « Developing a speech corpus from web news for Myanmar (Burmese) language ». In : *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE. DOI : 10.1109/icsda.2017.8384451.

Baayen, R. Harald (2001). *Word Frequency Distributions*. Dordrecht Boston London : Kluwer Academic Publishers. DOI : 10.1007/978-94-010-0844-0.

Bailin, Alan et Ann Grafstein (2001). « The linguistic assumptions underlying readability formulae : A critique ». In : *Language & communication* 21.3, p. 285-301.

– (2016). *Readability : Text and Context*. Palgrave Macmillan UK. DOI : 10.1057/9781137388773.

Banerjee, S. et T. Pedersen (fév. 2003). « The Design, Implementation, and Use of the Ngram Statistic Package ». In : *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*. Mexico City, p. 370-381.

Baptista, Jorge, Neuza Costa, Joaquim Guerra, Marcos Zampieri, Maria Cabral et Nuno Mamede (2010). « P-AWL : Academic Word List for Portuguese ». In : *Computational Processing of the Portuguese Language*. Sous la dir. de Thiago Alexandre Salgueiro Pardo, António Branco, Aldebaro Klautau, Renata Vieira et Vera Lúcia Strube de Lima. Berlin, Heidelberg : Springer Berlin Heidelberg, p. 120-123. ISBN : 978-3-642-12320-7.

Baroni, Marco (2008). « Distributions in text ». In : *Corpus Linguistics : An International Handbook*. Sous la dir. d'Anke Lüdeling et Merja Kytö. Mouton de Gruyter.

Behrens, Heike et Stefan Pfänder, éd. (fév. 2016). *Experience Counts : Frequency Effects in Language*. De Gruyter. DOI : 10.1515/9783110346916.

Beinborn, Lisa, Torsten Zesch et Iryna Gurevych (2014). « Readability for foreign language learning : The importance of cognates ». In : *ITL-International Journal of Applied Linguistics* 165.2, p. 136-162.

Bellassen, Joël (2008). *Méthode d'Initiation à la Langue et à l'Ecriture chinoises*. COMPAGNIE. ISBN : 9782950413567.

– (2010). « La didactique du chinois et la malédiction de Babel. Émergence, dynamique et structuration d'une discipline ». In : *Études chinoises* 1.1, p. 27-44. DOI : 10.3406/etchi.2010.965.

Berment, Vincent (2004). « Méthodes pour informatiser les langues et les groupes de langues « peu dotées » ». PhD Thesis. Université Joseph-Fourier-Grenoble I.

– (2014). « Some thoughts on how to address commercially unprofitable languages and language pairs. » In : *WSSANLP 2014*. Dublin.

Bernot, Denise, éd. (1988-1992). *Dictionnaire birman-français, vol. 12-15*. T. 12-15. Paris : SELAF / Peeters.

– éd. (1978-1988). *Dictionnaire birman-français, vol. 1-11*. T. 1-11. Paris : SELAF.

– (1971). « L'épithète en birman. Contribution à l'étude des langues sans catégorie adjectivale ». In : *La Linguistique* 7.1, p. 41-53. ISSN : 0075966X, 21010234.

– (1980). « Le Prédicat En Birman Parlé ». In : *Ase8*.

– (1983). « Y a-t-il des catégories adjectivales et adverbiales en birman ? » In : *Cahiers de l'Asie du Sud-Est 13-14*, p. 67-78.

Bernot, Denise, Marie-Hélène Cardinaud et Marie Yin Yin Myint (2001). *Manuel de birman, volume 2 : Grammaire birmane*. Langues et Mondes. L'Asiathèque. ISBN : 978-2911053764.

– (2010). *Manuel de birman : Langue de Myanmar Volume 1*. Langues et Mondes. L'Asiathèque. ISBN : 978-2915255928.

Bernot, Denise, Cristina Cramerotti et Marie Yin Yin Myint, éd. (1998). *Dictionnaire français-birman*. Paris : Langues et Mondes L'Asiathèque.

BGN/PCGN (1970). *Romanization of Burmese BGN/PCGN 1970 Agreement*.

Bhattacharya, Nilhilesh (1965). « Some Statistical Studies of the Bangla Language ». PhD. Kolkata : Indian Statistical Institute.

Biber, Douglas (1993). « Representativeness in Corpus Design ». In : *Literary and Linguistic Computing* 8.4, p. 243-257. DOI : 10.1093/llc/8.4.243.

Biber, Douglas, Randi Reppen, Erin Schnur et Romy Ghanem (2016). « On the (non) utility of Juilland's D to measure lexical dispersion in large corpora ». In : *International Journal of Corpus Linguistics* 21.4, p. 439-464.

Boada, Roger, Marc Guasch, Juan Haro, Josep Demestre et Pilar Ferré (mar. 2019). « SUBTLEX-CAT : Subtitle word frequencies and contextual diversity for Catalan ». In : *Behavior Research Methods* 52.1, p. 360-375. DOI : 10.3758/s13428-019-01233-1.

Bortolini, U., C. Tagliavini et A. Zampolli, éd. (1971). *Lessico di frequenza della lingua italiana contemporanea*. Milano : Garzanti, IBM Italia.

Bowker, Lynne et Jennifer Pearson (2002). *Working with Specialized Language : A Practical Guide to Using Corpora*. Applied linguistics/modern languages/ELT. Routledge. ISBN : 9780415236997. URL : https://books.google.fr/books?id=nx4iO6boNzIC.

Brac de La Perrière, Bénédicte (1999). « Le nom personnel birman : Son choix et ses usages ». In : *D'un nom à l'autre en Asie du Sud-Est. Approches ethnologiques*. Sous la dir. de Josiane Massard-Vincent. Sous la dir. de Simonne Pauwels. Karthala. Paris.

Bradley, David (2009). « Burma, Thailand, Cambodia, Laos, Vietnam ». In : *Sociolinguistics Around the World*. Sous la dir. de Martin J. Ball. London : Routledge, p. 96-105.

– (2011). « Changes in Burmese Phonology and Orthography ». In : *Keynote, Southeast Asian Linguistics Society 21*. Kasetsart U., Bangkok, Thailand. URL : https://www.academia.edu/1559757/Changes_in_Burmese_Phonology_and_Orthography.

Braune, Fabienne et Alexander Fraser (2010). « Improved Unsupervised Sentence Alignment for Symmetrical and Asymmetrical Parallel Corpora ». In : *Proceedings of the 23rd International Conference on Computational Linguistics : Posters*. COLING '10. Beijing, China : Association for Computational Linguistics, p. 81-89.

Brezina, Vaclav (sept. 2018). *Statistics in Corpus Linguistics*. Cambridge University Press. DOI : 10.1017/9781316410899.

Brown, Charles Barrett, Shane Milton Lanning, Wesley Moore Carr et Committee on modern languages (1945). *A Graded Word-book of Brazilian Portuguese*. New York : F. S. Crofts & co.

Browne, C., B. Culligan et J. Phillips (2013). *The New General Service List*. URL : http://www.newgeneralservicelist.org.

Brysbaert, Marc, Matthias Buchmeier, Markus Conrad, Arthur M Jacobs, Jens Bölte et Andrea Böhl (2011). « The word frequency effect : a review of recent developments and implications for the choice of frequency estimates in German. » In : *Experimental psychology* 58.5, p. 412.

Brysbaert, Marc et Kevin Diependaele (oct. 2012). « Dealing with zero word frequencies : A review of the existing rules of thumb and a suggestion for an evidence-based choice ». In : *Behavior Research Methods* 45.2, p. 422-430. DOI : 10.3758/s13428-012-0270-5.

Brysbaert, Marc et Boris New (nov. 2009). « Moving beyond Kučera and Francis : A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English ». In : *Behavior Research Methods* 41.4, p. 977-990. DOI : 10.3758/brm.41.4.977.

Buchanan, A.M. (1927). *A Graded Spanish Word Book*. Toronto : The University of Toronto press.

Buck, Christian, Kenneth Heafield et Bas van Ooyen (mai 2014). « N-gram Counts and Language Models from the Common Crawl ». In : *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland : European Language Resources Association (ELRA), p. 3579-3584. URL : http://www.lrec-conf.org/proceedings/lrec2014/pdf/1097_Paper.pdf.

Buckwalter, Tim et Dilworth Parkinson (juil. 2014). *A Frequency Dictionary of Arabic*. Routledge. DOI : 10.4324/9780203883280.

Buerki, Andreas (2017a). « Computational workshop : N-Gram processor ». In : *Research Symposium on Methods and Applications of Computational (and other) Identification of Formulaicity*. Research Symposium on Methods and Applications of Computational (and other) Identification of Formulaicity. Cardiff University. Cardiff.

– (2017b). « Frequency Consolidation Among Word N-Grams ». In : *Computational and Corpus-Based Phraseology*. Sous la dir. de Ruslan Mitkov. Cham : Springer International Publishing, p. 432-446. ISBN : 978-3-319-69805-2.

Burch, Brent, Jesse Egbert et Douglas Biber (juin 2017). « Measuring and interpreting lexical dispersion in corpus linguistics ». In : *Journal of Research Design*

and Statistics in Linguistics and Communication Science 3.2, p. 189-216. DOI : 10.1558/jrds.33066.

Burke, Sean M. (2002). *Perl and LWP*. O Reilly. URL : http://lwp.interglacial.com/.

Burnard, Lou (1995). *Users Reference Guide, British National Corpus, Version 1.0*. Oxford : Oxford University Computing Services.

Cai, Qing et Marc Brysbaert (2010). « SUBTLEX-CH : Chinese word and character frequencies based on film subtitles ». In : *PloS one* 5.6, e10729.

Carey, Felix (1814). *A Grammar of the Burman Language to Which is Added Roots From Which the Language is Derived*. Serampore : Mission Press.

Carrell, Patricia L (1987). « Readability in ESL ». In : 4.1.

Ćavar, Damir, Lwin Moe, Hai Hu et Kenneth Steimel (2016). « Preliminary results from the Free Linguistic Environment project ». In : *Proceedings of the Joint 2016 Conference on Head-driven Phrase Structure Grammar and Lexical Functional Grammar, Polish Academy of Sciences, Warsaw, Poland*. Sous la dir. de Doug Arnold, Miriam Butt, Berthold Crysmann, Tracy Holloway King et Stefan Müller. Stanford, CA : CSLI Publications, p. 161-181.

Center for Southeast Asian Studies, Northern Illinois University (2017). *SEASite Burmese (Myanmar). The Language, Literature and Culture of the Union of Myanmar (Burma)*. URL : http://www.seasite.niu.edu/burmese/ (visité le 17/04/2017).

Čermák, František et Michal Křen (mar. 2011). *A Frequency Dictionary of Czech*. Routledge. DOI : 10.4324/9780203805978.

Chall, Jeanne Sternlicht (1958). *Readability : An Appraisal Of Research And Application*. Columbus : Ohio State University. ISBN : 978-1258289126.

Chang, Charles Bond (2009). « English Loanword Adaptation In Burmese ». In : *Journal of the Southeast Asian Linguistics Society* 1, p. 77-94. URL : https://core.ac.uk/download/pdf/42547669.pdf.

Chen, Heqin (1922). « Yutiwen Yingyong Zihui [Caractères utilisés en chinois vernaculaire'] ». In : *XinJiaoyu [Nouvelle Education]* 5.5, p. 987-995.

Chen, Peng-Jen, Jiajun Shen, Matt Le, Vishrav Chaudhary, Ahmed El-Kishky, Guillaume Wenzek, Myle Ott et Marc'Aurelio Ranzato (2019). *Facebook AI's WAT19 Myanmar-English Translation Task Submission*. arXiv : 1910.06848 [cs.CL].

Chinnak, Sommai (2015). **การเรียนการสอนและผู้เชี่ยวชาญภาษาพม่าในประเทศ** *(Learning & Teaching and Experts of Burmese Language in Thailand).*

Christodouloupoulos, Christos et Mark Steedman (nov. 2014). « A massively parallel corpus : the Bible in 100 languages ». In : *Language Resources and Evaluation* 49.2, p. 375-395. DOI : 10.1007/s10579-014-9287-y.

Cobb, T. et M. Horst (2004). « Is there room for an AWL in French ? » In : *Vocabulary in a Second Language : Selection, acquisition, and testing*. Sous la dir. de P. Bogaards et B. Laufer. Amsterdam : John Benjamins, p. 15-38.

Coltheart, Max (1981). « The MRC psycholinguistic database ». In : *The Quarterly Journal of Experimental Psychology Section A* 33.4, p. 497-505.

Comité des manuels sur le contenu des programmes d'études 2016-2017 (2015). *Grammaire du birman, tome 1, chapitre 3, 7e année* မြန်မာသဒ္ဒါ အတွဲ (၁)၊ အခန်း (၃) - သတ္တမတန်း။.

Conseil de l'Europe (2021). *Cadre européen commun de référencepour les langues : Apprendre, enseigner, évaluer*. Sous la dir. de Division des politiques éducatives, Service de l'éducation, Conseil de l'Europe. Strasbourg : Éditions du Conseil de l'Europe. URL : https://rm.coe.int/cadre-europeen-commun-de-reference-pour-les-langues-apprendre-enseigne/1680a4e270.

Coxhead, Averil (2000). « A New Academic Word List ». In : *TESOL Quarterly* 34.2, p. 213-238.

Crossley, Scott A, Jerry Greenfield et Danielle S McNamara (2008). « Assessing text readability using cognitively based indices ». In : *Tesol Quarterly* 42.3, p. 475-493.

Cuetos, Fernando, Maria Glez-Nosti, Analía Barbón et Marc Brysbaert (2012). « SUBTLEX-ESP : Spanish word frequencies based on film subtitles ». In : *Psicológica* 33.2, p. 133-143.

Cunningham, Nance (2007). « Why Compile a New English–Burmese and Burmese–English Dictionary ? » In : *Bulletin of the Burma Studies Group* 80. Sous la dir. de Ward Keeler.

Dale, D et J Chall (1948). « Formula for Predicting Readability ». In : *Educational research bulletin* 27, p. 11-20.

Dale, Edgar et Donald Henry Reichert (1957). *Bibliography of Vocabulary Studies*. Columbus, Ohio : Bureau of Educational Research, Ohio State University. 174 p. (Visité le 26/02/2019).

Das, Sreerupa et Rajkumar Roychoudhury (2004). « Testing Level of Readability in Bangla Novels of Bankim Chandra Chattopadhyay with Respect to the Density of Polysyllabic Words. » In : *Indian Journal of Linguistics* 22, p. 41-51.

– (2006). « Readability modelling and comparison of one and two parametric fit : A case study in Bangla ». In : *Journal of Quantitative Linguistics* 13.01, p. 17-34.

Dash, Niladri Sekhar (2007). « Frequency-based analysis of words and morphemes in Bengali text corpus ». In : *Indian Journal of Linguistics* 25.26, p. 223-253. (Visité le 26/02/2019).

Daud, Nuraihan Mat, Haslina Hassan et Normaziah Abdul Aziz (2013). « A corpus-based readability formula for estimate of arabic texts reading difficulty ». In : *World Applied Sciences Journal* 21, p. 168-173.

Davies, Mark (déc. 2005). *A Frequency Dictionary of Spanish*. Routledge. DOI : 10.4324/9780203415009.

Davies, Mark et Dee Gardner (août 2013). *A Frequency Dictionary of Contemporary American English*. Routledge. DOI : 10.4324/9780203880883.

Davies, Mark et Ana Maria Raposo Preto-Bay (nov. 2007). *A Frequency Dictionary of Portuguese*. Routledge. DOI : 10.4324/9780203937631.

de Malézieux, Guillaume, Amélie Bosc et Vincent Berment (2014). « RBMT as an alternative to SMT for under-resourced languages ». In : *Proceedings of the Fifth Workshop on South and Southeast Asian Natural Language Processing*. Dublin, Ireland : Association for Computational Linguistics et Dublin City University, p. 50-54. DOI : 10.3115/v1/W14-5507.

DeFrancis, John (1996). *ABC Chinese-English dictionary*. Richmond : Curzon. ISBN : 0700705112.

Dehaene, Stanislas, Florent Meyniel, Catherine Wacongne, Liping Wang et Christophe Pallier (oct. 2015). « The Neural Representation of Sequences : From Transition Probabilities to Algebraic Patterns and Linguistic Trees ». In : *Neuron* 88.1, p. 2-19. DOI : 10.1016/j.neuron.2015.09.019.

Department of Population (2015). *The 2014 Myanmar Population and Housing Census. The Union Report CEnsus Report Volume 2*. Rapp. tech. Office No. 48 Nay Pyi Taw : Ministry of Immigration et Population.

DeRocher, J.E. (1973). « The Counting of Words : A Review of the History ». In : *Techniques and Theory of Word Counts with Annotated Bibliography*. New York : Syracuse University Research Corpp.

Dewey, Godfrey (1923). *Relative Frequency Of English Speech Sounds*. London : Harvard University Press, p. 187.

Dim Lam Cing et Khin Mar Soe (avr. 2020). « Improving accuracy of Part-of-Speech (POS) tagging using hidden markov model and morphological analysis for Myanmar Language ». In : *International Journal of Electrical and Computer Engineering (IJECE)* 10.2, p. 2023. DOI : 10.11591/ijece.v10i2.pp2023-2030.

Dimitropoulou, Maria, Jon Andoni Duñabeitia, Alberto Avilés, José Corral et Manuel Carreiras (2010). « Subtitle-based word frequencies as the best estimate of reading behavior : The case of Greek ». In : *Frontiers in psychology* 1, p. 218.

Ding, Chenchen, Hnin Thu Zar Aye, Win Pa Pa, Khin Thandar Nwet, Khin Mar Soe, Masao Utiyama et Eiichiro Sumita (mai 2019). « Towards Burmese (Myanmar) Morphological Analysis : Syllable-based Tokenization and Part-of-speech Tagging ». In : *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 19.1, 5 :1-5 :34. ISSN : 2375-4699. DOI : 10.1145/3325885. URL : http://doi.acm.org/10.1145/3325885.

Ding, Chenchen, Hnin Thu Zar Aye, Masao Utiyama, Win Pa Pa et Eiichiro Sumita (2016). *Tokenization and Part-of-Speech AnnotationGuidelines for Myanmar (Burmese) Version 0.2*. URL : https://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/Myanmar-annotation-guideline.pdf.

Ding, Chenchen, Masao Utiyama et Eiichiro Sumita (déc. 2018). « NOVA : A Feasible and Flexible Annotation System for Joint Tokenization and Part-of-Speech Tagging ». In : *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 18.2, 17 :1-17 :18. ISSN : 2375-4699. DOI : 10.1145/3276773. URL : http://doi.acm.org/10.1145/3276773.

Ding, Chenchen, Win Pa Pa, Masao Utiyama et Eiichiro Sumita (2018). « Burmese (Myanmar) Name Romanization : A Sub-syllabic Segmentation Scheme for Statistical Solutions ». In : *Communications in Computer and Information Science*. Springer Singapore, p. 191-202. DOI : 10.1007/978-981-10-8438-6_16.

Ding, Chenchen, Ye Kyaw Thu, Masao Utiyama et Eiichiro Sumita (mai 2016). « Word Segmentation for Burmese (Myanmar) ». In : *ACM Transactions on*

Asian and Low-Resource Language Information Processing 15.4, p. 1-10. DOI : 10.1145/2846095.

Dixon, Robert MW et Alexandra Y Aikhenvald (2002). « Word : a typological framework ». In : *Word : A cross-linguistic typology*. Sous la dir. d'Alexandra Y Aikhenvald et Robert MW Dixon, p. 1-41.

Doctor Tin Aung, Min Kyaw, Nat Nwe et Hla Thamein (p. d.). *Manuel de traduction pour tous* ဘာသာပြန် စာပေစာတမ်းများ, ပြည်သူ့လက်စွဲစာစဉ်. Yangon : Sarpay Beikman.

Dryer, Matthew S. et the WALS author team (2011). *Language page for Burmese, The World Atlas of Language Structures Online*. Sous la dir. de Martin Dryer Matthew S. Haspelmath. URL : http://wals.info/languoid/lect/wals_code_brm%202018-05-28 (visité le 18/04/2016).

DuBay, William H. (2004). *The Principles of Readability*. Costa Mesa, California, US : Impact Information.

Dunning, Ted (mar. 1993). « Accurate Methods for the Statistics of Surprise and Coincidence ». In : *Comput. Linguist.* 19.1, p. 61-74. ISSN : 0891-2017.

Dupertuis, Noemi-Tiina (déc. 2022). « ビルマ語資料の検索・書誌作成のポイント [Points clés pour la recherche et la bibliographie de documents birmans] ». In : アジア情報室通報 *Bulletin of the Asian Resources Room.* À paraître. 20.4. ISSN : 1348-2149.

Duroiselle, Charles (1913). « Burmese Philology ». In : *Journal of the Burma Research Society* 3, p. 103-46.

Dutilleul, Pierre (2018). *Les chiffres de l'édition. Rapport statistique du SNE 2017-2018, France et International*. Rapp. tech. Syndicat national de l'édition.

Eberhard, David M., Gary F. Simons et Charles D. Fennig, éd. (2019). *Ethnologue : Languages of the World : Twenty-second edition*. SIL International. URL : http://www.ethnologue.com.

EF Education First (2021). *EF EPI (EF English Proficiency Index)*. Rapp. tech. EF Education First. URL : https://www.ef.com/wwen/epi/.

Egreteau, Renaud (2012). « Burma in Diaspora : A Preliminary Research Note on the Politics of Burmese Diasporic Communities in Asia ». In : *journal of Current Southeast Asian Affairs* 31.2, p. 115-147.

Engel, E. F. (1931). « The Use of a Standardized Vocabulary in Beginning German ». In : *The Modern Language Journal* 15.4, p. 281-291. DOI : 10.1111/j.1540-4781.1931.tb06495.x.

Esche, Annemarie (2005). « The experience of writing the first German-Myanmar Dictionary ». In : *Studies in Burmese linguistics*. Sous la dir. de Justin Watkins. Canberra : The Australian National University, p. 1-25.

Esplà-Gomis, Miquel, Mikel L Forcada, Gema Ramírez-Sánchez et Hieu Hoang (2019). « ParaCrawl : Web-scale parallel corpora for the languages of the EU ». In : *Proceedings of Machine Translation Summit XVII Volume 2 : Translator, Project and User Tracks*, p. 118-119.

Flesch, Rudolph (1948). « A New Readability Yardstick. » In : *Journal of Applied Psychology* 32.3, p. 221.

François, Thomas (2011). « Les apports du traitement automatique du langage à la lisibilité du français langue étrangère ». PhD Thesis. Université Catholique de Louvain.

– (nov. 2015). « When readability meets computational linguistics : a new paradigm in readability ». In : *Revue française de linguistique appliquée* Vol. XX.2, p. 79-97. DOI : 10.3917/rfla.202.0079.

Garreta, Raul et Guillermo Moncecchi (2013). *Learning scikit-learn : machine learning in python*. Packt Publishing Ltd.

Gass, Susan M. et Larry Selinker (2000). *Second Language Acquisition : An Introductory Course*. Routledge. ISBN : 978-0805835281.

Ghadirian, Sina (2002). « Providing controlled exposure to target vocabulary through the screening and arranging of texts ». In : *Language Learning and Technology* 6.1, p. 147-164.

Goddard, Cliff (2011). « The lexical semantics of language (with special reference to words) ». In : *Language Sciences* 33.1, p. 40-57.

Government of the Union of Myanmar (2002). « Government of the Union of Myanmar Notification 5/89, 18th June, 1989 ». In : *Eighth United Nations Conference on the Standardization of Geographical Names, Berlin, 27 August – 5 September 2002*. URL : https://unstats.un.org/unsd/geoinfo/UNGEGN/docs/8th-uncsgn-docs/inf/8th_UNCSGN_econf.94_INF.75.pdf.

Goyal, Naman, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán

et Angela Fan (2021). « The FLORES-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation ». In :

Green, Antony D. (1995). « The prosodic structure of Burmese : A constraint-based approach ». In : *Working Papers of the Cornell Phonetics Laboratory* 10, p. 67-96.

Green, Antony Dubach (2005). « Word, foot, and syllable structure in Burmese ». In : *Studies in Burmese linguistics*. Sous la dir. de Justin Watkins. Canberra : The Australian National University, p. 1-25.

Grefenstette, Gregory (2010). « Estimating the Number of Concepts ». In : *A Way with Words : Recent Advances in Lexical Theory and Analysis : A Festschrift for Patrick Hanks*. Menha Publishers. URL : https://hal.inria.fr/hal-01081033.

Gries, Stefan Th (2008). « Dispersions and adjusted frequencies in corpora ». In : *International journal of corpus linguistics* 13.4, p. 403-437.

Gries, Stefan Th. et John Newman (2014). « Creating and using corpora ». In : *Research Methods in Linguistics*. Cambridge University Press. ISBN : 1107014336.

Gunning, Robert (1952). « Technique of Clear Writing ». In :

Hammarström, Harald, Sebastian Bank, Robert Forkel et Martin Haspelmath (20 déc. 2021). *Glottolog 4.5*. URL : http://glottolog.org (visité le 28/05/2018).

Haspelmath, Martin (2017). « The indeterminacy of word segmentation and the nature of morphology and syntax ». In : *Folia linguistica* 51.s1000, p. 31-80.

Haygood, James Douglas (1951). *Le vocabulaire fondamental du français : étude pratique sur l'enseignement des langues vivantes*. Genève : Librairie Droz.

Heiden, Serge, Jean-Philippe Magué et Bénédicte Pincemin (juin 2010). « TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement ». In : *10th International Conference on the Statistical Analysis of Textual Data - JADT 2010*. Sous la dir. de Sergio Bolasco, Isabella Chiari et Luca Giuliano. T. 2. 3. Rome, Italy : Edizioni Universitarie di Lettere Economia Diritto, p. 1021-1032. URL : https://halshs.archives-ouvertes.fr/halshs-00549779.

Heitz, Thomas (2006). « Modélisation du prétraitement des textes ». In : *JADT'06 (International Conference on Statistical Analysis of Textual Data)*. T. 1, p. 499-506.

Henmon, V. A. C., éd. (1924). *A French Word Book Based on a Count of 400,000 Running Words*. Madison : University of Wisconsin.

Henry, Georges (1975). *Comment mesurer la lisibilité*. Sous la dir. de Labor-Nathan. Paris.

Hildebrandt, Kristine A. (2015). « The prosodic word ». In : *The Oxford Handbook of the Word*.

Hla Hla Htay et Kavi Narayana Murthy (2008). « Myanmar Word Segmentation using Syllable level Longest Matching ». In : *Proceedings of the IJCNLP-2008 Workshop on Asian Language Resources, Jan 11-12*. Hyderabad, India.

Hla Thamein (2000). *Myanmar Proverbs in Myanmar and English*. Yangon : Tuiṅ Laṅ Cā pe.

Hlaing, Tin Htay et Yoshiki Mikami (juil. 2014). « Automatic syllable segmentation of Myanmar texts using finite state transducer ». In : *International Journal on Advances in ICT for Emerging Regions (ICTer)* 6.2. DOI : 10.4038/icter.v6i2.7150.

Hlaing, Zar Zar, Ye Kyaw Thu, Myat Myo Nwe Wai, Thepchai Supnithi et Ponrudee Netisopakul (2020). « Myanmar POS Resource Extension Effects on Automatic Tagging Methods ». In : *2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*. IEEE, p. 1-6.

Hnin Tun, San San (2006). « Discourse Marking in Burmese and English : A Corpus-Based Approach ». PhD. Université de Nottingham.

– (2013). « La grammaire du discours en birman parlé : Fonctions des particules énonciatives dans la grammaire du birman parlé ». PhD. Université Sorbonne Nouvelle, Paris 3.

Hnin Tun, San San et Patrick McCormick (août 2015). *Colloquial Burmese. The Complete Course for Beginners*. Routledge. DOI : 10.4324/9780203123867.

Hopple, Paulette (2007). « Burmese Particles as Boundary Marking Units of Text ». In : *"Particle Party" Workshop* (Payap University). Chiang Mai.

Hopple, Paulette Mary (2003). « The structure of nominalization in Burmese ». PhD Thesis. University of Texas at Arlington.

Hosken, Martin (2012). *Representing Myanmar in Unicode Details and Examples Version 4*. URL : https://www.unicode.org/notes/tn11/UTN11_4.pdf.

Hu, M. et I.S.P. Nation (2000). « Vocabulary density and reading comprehension ». In : *Reading in a Foreign Language* 13.1, p. 403-430.

Huang, Hung-Tzu et Hsien-Chin Liou (2007). « Vocabulary Learning in an Automated Graded Reading Program ». In : *Language Learning and Technology* 11.3, p. 64-82.

Islam, Zahurul, Alexander Mehler et Rashedur Rahman (2012). « Text Readability Classification of Textbooks of a Low-Resource Language ». In : *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*. Bali, Indonesia : Faculty of Computer Science, Universitas Indonesia, p. 545-553.

ISO/IEC (2017). *International Standard ISO/IEC 10646 : Information technology — Universal Coded Character Set (UCS)*. Rapp. tech. Version Fifth edition 2017-12. ISO International Organization for Standardization. URL : https://standards.iso.org/ittf/PubliclyAvailableStandards/c069119_ISO_IEC_10646_2017.zip (visité le 11/07/2019).

Jacobs, Mike, Andrew Glass et Peter Constable (17 juin 2018). *Creating and Supporting OpenType Fonts for Myanmar Script*. Sous la dir. de Microsoft Typography. URL : https://docs.microsoft.com/en-au/typography/script-development/myanmar (visité le 03/03/2019).

Jenny, Mathias (2015). « Foreign influence in the Burmese language ». In : *International Conference on Burma/Myanmar Studies, 24-26 July 2015*. Chiang Mai.

Jenny, Mathias et San San Hnin Tun (fév. 2017). *Burmese : A Comprehensive Grammar*. Routledge. DOI : 10.4324/9781315651194.

Johansson Kokkinakis, Sofie, Emma Sköldberg, Birgit Henriksen, Kari Kinn et Janne Bondi Johannesse (2012). « Developing Academic Word Lists for Swedish, Norwegian and Danish – a joint research project Fjeld ». In : *Proceedings of the 15th EURALEX International Congress 7 11 August, University of Osl*. Sous la dir. de R.V. Torjusenn et J.M. Torjusen. Oslo, p. 563-569.

Jones, Randall et Erwin Tschirner (2015). *A Frequency Dictionary of German*. Routledge. DOI : 10.4324/9780203883037.

Judson, Adoniram, éd. (1826). *A Dictionary of the Burman language with Explanations in English*. Culcutta : American Baptist Mission Press.

– (1842). *Grammatical Notices of the Burmese Language*. Maulmain : American Baptist Mission Press.
– éd. (1852). *Burmese and English, A Dictionary*. Maulmain : American Baptist Mission Press.

Juilland, Alphonse G., Dorothy R. Brodin, Catherine Davidovitch, Mary Ann Ignatius, Ileana Juilland et Lilian Szklarczyk (1970). *Frequency Dictionary of French Words*. Romance languages and their structures. Mouton. ISBN : 9789027915535.

Kahane, Sylvain (2008). « Les unités minimales de la syntaxe et de la sémantique : le cas du français ». In : *Congrès Mondial de Linguistique Française 2008*. EDP Sciences. DOI : 10.1051/cmlf08106.

Kawaguchi, Yuji (2007). « Foundations of center of Usage-based Linguistic Informatics (UBLI) ». In : *Corpus-Based Perspectives in Linguistics*. John Benjamins Publishing Company, p. 3-28. DOI : 10.1075/ubli.6.02kaw.

Keeler, Ward et Allen Lyan (2021). *Burmese : a cultural approach*. Hong Kong : Hong Kong University Press. ISBN : 9789888528400.

Keniston, Hayward, éd. (1941). *A Standard List of Spanish Words and Idioms*. Boston : D.C. Heath and Company.

Keuleers, Emmanuel, Marc Brysbaert et Boris New (2010). « SUBTLEX-NL : A new measure for Dutch word frequency based on film subtitles ». In : *Behavior research methods* 42.3, p. 643-650.

Khin Khin Aye (sept. 2020). « Myanmar English ». In : *The Handbook of Asian Englishes*. Sous la dir. de Kingsley Bolton, Werner Botha et Andy Kirkpatrick. John Wiley & Sons, Inc., p. 355-371. DOI : 10.1002/9781118791882.ch15.

Khin Myo Chit (1984). *A Wonderland of Burmese Legends*. Bangkok : The Tamarind Press.

Khin War War Htike, Ye Kyaw Thu, Zuping Zhang, Win Pa Pa, Yoshinori Sagisaka et Naoto Iwahashi (2017). « Comparison of Six POS Tagging Methods on 10K Sentences Myanmar Language (Burmese) POS Tagged Corpus ». In : 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2017) (17 avr. 2017). Budapest, Hungary.

Kincaid, JP, RP Fishburne Jr, RL Rogers et BS Chissom (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel*. Research Branch Report

8-75, Millington, TN : Naval Technical Training, U.S. Naval Air Station, Memphis, TN.

El-Kishky, Ahmed, Vishrav Chaudhary, Francisco Guzmán et Philipp Koehn (2020). « CCAligned : A Massive Collection of Cross-Lingual Web-Document Pairs ». In : *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.480.

Kyaw Htet Minn et Khin Mar Soe (2019). « Myanmar Word Stemming and Part-of-Speech Tagging using Rule Based Approach ». In : *National Journal of Parallel and Soft Computing*.

Lafon, Pierre (1980). « Sur la variabilité de la fréquence des formes dans un corpus ». In : *Mots* 1.1, p. 127-165. DOI : 10.3406/mots.1980.1008.

Lammerts, D. Christian et Arlo Griffiths (2016). *Standardized and simplified systems for the transliteration of Old Burmese, Burmese, Pali, and Sanskrit written in the Burmese script, Version 2.2*. Rapp. tech.

Latter, Thomas (1845). *A Grammar of the Language of Burmah*. Calcutta : Thacker et Company.

Laufer, Batia (1992). « How much lexis is necessary for reading comprehension ? » In : *Vocabulary and applied linguistics*. Springer, p. 126-132.

Laufer, Batia et Geke C Ravenhorst-Kalovski (2010). « Lexical threshold revisited : Lexical text coverage, learners' vocabulary size and reading comprehension ». In :

Lebart, Ludovic et André Salem (1994). *Statistique textuelle*. Paris : Dunod.

Leech, Geoffrey (2011). « Frequency, corpora and language learning ». In : *A taste for corpora : In honour of Sylviane Granger*, p. 7-32.

Lewis-Wong, Jennifer et Satenik Mkhitaryan (juil. 2016). « Pratique de la lecture en thaï et hindi en L2 : classification automatique de textes par progression lexicale ». In : *JEP-TALN-RECITAL 2016*. Actes de la conférence conjointe JEP-TALN-RECITAL 2016, volume 09 : ELTAL. Paris, France.

Li, Xingshan, Klinton Bicknell, Pingping Liu, Wei Wei et Keith Rayner (2014). « Reading is fundamentally similar across disparate writing systems : A systematic characterization of how words and characters influence eye movements in Chinese reading. » In : *Journal of Experimental Psychology : General* 143.2, p. 895-913. DOI : 10.1037/a0033580.

Liao, Han-Teng (jan. 2017). « Encoding for access ». In : *ACM SIGCAS Computers and Society* 46.4, p. 18-24. DOI : 10.1145/3040489.3040493.

Lively, Bertha A. et Sidney Leavitt Pressey (1923). *A Method for Measuring the "Vocabulary Burden" of Textbooks*.

Lo Bianco, Joseph (29 jan. 2016). *Building a National Language Policy for Myanmar : A Brief Progress Report*. UNICEF.

Lonsdale, Deryle et Yvon Le Bras (mar. 2009). *A Frequency Dictionary of French*. Routledge. DOI : 10.4324/9780203883044.

Lu, Xiaofei (8 juil. 2014). *Computational Methods for Corpus Annotation and Analysis*. Springer-Verlag GmbH. 186 p. ISBN : 9401786453.

Luzoe (1996). *Myanmar Newspaper Reader*. Kensington : Dunwoody Press. 308 p.

Lwin Moe (14 oct. 2016). *burmeseunicode*. URL : https://github.com/lwinmoe/burmeseunicode.

Lyssenko, Nicolas et Delphine Weulersse (1986). *Méthode programmée du chinois moderne*. Sous la dir. de Lyssenko. Paris.

Mallick, Bhakti Prasad, Nikhilesh Bhattacharya, Subhas Chandra Kundu et Mina Dawn (1998). *Phonemic and Morphemic Frequency in the Bengali Language*. Kolkata : Asiatic Society.

Mandera, Paweł, Emmanuel Keuleers, Zofia Wodniecka et Marc Brysbaert (juin 2014). « Subtlex-pl : subtitle-based word frequency estimates for Polish ». In : *Behavior Research Methods* 47.2, p. 471-483. DOI : 10.3758/s13428-014-0489-4.

Marie, Benjamin, Hour Kaing, Aye Myat Mon, Chenchen Ding, Atsushi Fujita, Masao Utiyama et Eiichiro Sumita (nov. 2019). « Supervised and Unsupervised Machine Translation for Myanmar-English and Khmer-English ». In : *Proceedings of the 6th Workshop on Asian Translation*. Hong Kong, China : Association for Computational Linguistics, p. 68-75. DOI : 10.18653/v1/D19-5206.

Maung Maung Pye (1952). *Tales of Burma*. Sous la dir. de Macmillan et Ltd Co. Calcutta.

McEnery, Tony et Andrew Hardie (2012). *Corpus linguistics : Method, theory and practice*. Cambridge University Press.

McLaughlin, G Harry (1969). « SMOG grading-a new readability formula ». In : *Journal of reading* 12.8, p. 639-646.

Meara, Paul (1980). « Vocabulary Acquisition : A Neglected Aspect of Language Learning ». In : *Language Teaching* 13.3-4, p. 221. DOI : 10.1017/s0261444800008879.

Mi Mi Khaing (1958). « Burmese names : A guide ». In : *The Atlantic. February*.

Milton, J (2010). « The development of vocabulary breadth across the CEFR levels ». In : *Second Language Acquisition and Testing in Europe*. Sous la dir. d'I. Bartning, Martin M. et I. Vedder. Online : Eurosla, p. 211-232.

Milton, James et Nicola Hopkins (2006). « Comparing Phonological and Orthographic Vocabulary Size : Do Vocabulary Tests Underestimate the Knowledge of Some Learners ». In : *The Canadian Modern Language Review / La revue canadienne des langues vivantes* 63.1, p. 127-147. DOI : 10.1353/cml.2006.0048.

Minn Latt (1959). « A contribution towards the identification of the word and the parts of speech in modern Burmese ». English. In : *Archiv orientální : Quarterly journal of African and Asian studies* 27.2, p. 318-335.

Moore, Bruce (1997). *The Australian concise Oxford dictionary of current English*. Melbourne ; New York : Oxford University Press.

Muller, Charles (1973). *Initiation au méthodes de la statistique linguistique*. Classiques Hachette.

Müller, Charles (1977). *Principes et méthodes de statistique lexicale*. Paris : Hachette université.

Myanmar Language Commission (1993). *Myanmar-English Dictionary* မြန်မာ-အင်္ဂလိပ် အဘိဓာန်. Union of Myanmar : Department of the Myanmar Language Commission, Ministry of Education.

– (2003). *Burmese Orthography* မြန်မာစာလုံးပေါင်းသတ်ပုံကျမ်း.

– (2005). *Burmese Grammar* မြန်မာသဒ္ဒါ. Yangon.

– (7 août 2008). *Myanmar Dictionary* မြန်မာအဘိဓာန်. Union of Myanmar : Department of the Myanmar Language Commission, Ministry of Education.

Myint Soe (1999). « A Grammar of Burmese ». Thèse de doct. University of Oregon.

Myint, Cynthia (2011a). « A hybrid approach for part-of-speech tagging of Burmese texts ». In : *2011 International Conference on Computer and Management (CAMAN)*. IEEE, p. 1-4.

– (2011b). « A Part of Speech Tagger for Myanmar Text ». Thèse de doct. MERAL Portal.

Nagy, William E, Irene-Anna N Diakidoy et Richard C Anderson (1993). « The acquisition of morphology : Learning the contribution of suffixes to the meanings of derivatives ». In : *Journal of reading Behavior* 25.2, p. 155-170.

Nakazawa, Toshiaki, Chenchen Ding, Raj Dabre, Anoop Kunchukuttan, Nobushige Doi, Yusuke Oda, Ondřej Bojar, Shantipriya Parida, Isao Goto et Hidaya Mino, éd. (nov. 2019). *Proceedings of the 6th Workshop on Asian Translation*. Hong Kong, China : Association for Computational Linguistics.

Nation, I. S. P. (oct. 2001). *Learning Vocabulary in Another Language*. Cambridge University Press. DOI : 10.1017/cbo9781139858656.

– (2006). « How Large a Vocabulary is Needed For Reading and Listening ? » In : *Canadian Modern Language Review* 63.1, p. 59-82. DOI : 10.3138/cmlr.63.1.59.

Nation, Paul (1993). « Using dictionaries to estimate vocabulary size : Essential, but rarely followed, procedures ». In : *Language testing* 10.1, p. 27-40.

Nation, Paul et Paul Meara (2010). « Vocabulary ». In : *An Introduction to applied linguistics*. Sous la dir. de Norbert Schmitt. London : Hodder Education.

Navigli, R. et S. Ponzetto (2012). « BabelNet : The Automatic Construction Evaluation and Application of a Wide-Coverage Multilingual Semantic Network ». In : *Artificial Intelligence* 193, p. 217-250.

Ne Win San, နေဝင်းစံ (2012). « Peut-on mourir d'une faute d'orthgraphe ? သတ်ပုံမှားရင် သေတတ်ပါသလား ». mya. In : VV., AA. *La langue birmane, règles et régulations* စည်းနဲ့ ကမ်းနဲ့ မြန်မာစာ. Yangon, Myanmar : ရွှေပြည်တန်စာပေ, p. 255.

New, Boris, Marc Brysbaert, Jean Veronis et Christophe Pallier (sept. 2007). « The use of film subtitles to estimate word frequencies ». In : *Applied Psycholinguistics* 28.4, p. 661-677. DOI : 10.1017/s014271640707035x.

Nishi, Yoshio (mar. 1998). « The Orthographic Standardization of Burmese : Linguistic and Sociolinguistic Speculations ». In : *Bulletin of the National Museum of Ethnology* 22.4, p. 975-999.

Nivre, Joakim et al. (2017). *Universal Dependencies 2.1*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. URL : http://universaldependencies.org/.

Niyomtham, Wirat, Oranuch Niyomtham, Sunantha Thesuk et Sandar Aye (2017). **คู่มือการสอนภาษาพม่าแนวปริชาน** *[Burmese Language Teaching Guide] Myan-*

mar Flower Grammar. Myanmar Language Learning Development Center, Naresuan University Faculty of Humanities. ISBN : 9786164260528.

Nomoto, Hiroki, Kenji Okano, David Moeljadi et Hideo Sawada (2018). « TUFS Asian Language Parallel Corpus (TALPCo) ». In : *Proceedings of the Twenty-Fourth Annual Meeting of the Association for Natural Language Processing*, p. 436-439.

Nusbaum, H. C., D. Pisoni et C. Davis (1984). *Sizing up the hoosier mental lexicon : Measuring the familiarity of 20,000 words*. Research on speech perception, progress report 10. Indiana University, p. 357-376.

Nyein Thwet Thwet Aung et Ni Lar Thein (2011). « Word sense disambiguation system for Myanmar word in support of Myanmar-English machine translation ». In : *SICE Annual Conference 2011*, p. 2835-2840.

Office of the Superintendent (1908). *Tables for the Transliteration of Burmese into English, 1907*. Rangoon, Burma : Government Printing.

Oka, Shigenori (2002). 音で引く・タイ語検索Book *[Ouvrage de référence du thaï avec recherche phonétique]*. Ōsaka : Boisu. ISBN : 9784434026690.

Okano, Kenji, Nang Mya Kay Khaing, Thant Sin Hmwe et Sandar Myint (2016). *MLT Myanmar Language Test*. URL : http://www.mlt-myanmar.com/index.php/about-mlt/1017-2/.

Okell, John (1969). *A Reference Grammar of Colloquial Burmese*. London : Oxford University Press.

– (1971). *A guide to the romanization of Burmese*. James G. London : Luzac [for] The Royal Asiatic Society of Great Britain et Ireland.

– (2014). *Burmese by Ear : or Essential Myanmar*. Pansodan Books. ISBN : 978-0987925329. URL : https://www.soas.ac.uk/bbe/.

– (mar. 2018). *A Corpus of Modern Burmese*. DOI : 10.5281/zenodo.1202324.

Okell, John et Ann Allott (2017). *Burmese/Myanmar Dictionary of Grammatical Forms*. Richmond : Curzon Press.

Okell, John, U Saw Tun et Daw Khin Mya Swe (2010a). *Burmese : An Introduction to the Literary Style*. Southeast Asia Publications, Northern Illinois Uni. ISBN : 978-1877979446.

– (2010b). *Burmese : An Introduction to the Script*. Southeast Asia Publications, Northern Illinois University. ISBN : 978-1877979439.

– (2010c). *Burmese : An Introduction to the Spoken Language, Book 1*. Northern Illinois University. ISBN : 978-1877979415.

Ortiz Suárez, Pedro Javier, Benoît Sagot et Laurent Romary (2019). « Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures ». en. In : sous la dir. de Piotr Bański, Adrien Barbaresi, Hanno Biber, Evelyn Breiteneder, Simon Clematide, Marc Kupietz, Harald Lüngen et Caroline Iliadi. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019. Mannheim : Leibniz-Institut für Deutsche Sprache, p. 9-16. DOI : 10.14618/ids-pub-9021.

Pann Yu Mon, Chew Yew Choong et Yoshiki Mikami (mar. 2011). « Language Specific Crawler for Myanmar Web pages ». In : *IJCSI International journal of Computer Science Issues* 8.2.

Pawley, A et F Hodgetts Syder (1983). « Two puzzles for linguistic theory : Nativelike selection and nativelike fluency ». In : *Language and communication*, p. 191-225.

Pearson, P David (1974). « The effects of grammatical complexity on children's comprehension, recall, and conception of certain semantic relations ». In : *Reading Research Quarterly*, p. 155-192.

Pedregosa, F. et al. (2011). « Scikit-learn : Machine Learning in Python ». In : *Journal of Machine Learning Research* 12, p. 2825-2830.

Pellard, Thomas (2018). *Formation LaTeX INALCO Doctorat*. Paris, France. URL : https://hal.archives-ouvertes.fr/cel-01527916 (visité le 12/06/2019).

Pham, Hien, Benjamin V. Tucker et R. Harald Baayen (mar. 2019). « Constructing two vietnamese corpora and building a lexical database ». In : *Language Resources and Evaluation*. DOI : 10.1007/s10579-019-09451-x.

Pitler, Emily et Ani Nenkova (oct. 2008). « Revisiting Readability : A Unified Framework for Predicting Text Quality ». In : *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii : Association for Computational Linguistics, p. 186-195.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li et Peter J. Liu (2019). « Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer ». In : *arXiv e-prints*. arXiv : 1910.10683.

Richaudeau, François (1976). « Faut-il brûler les formules de lisibilité ? » In : *Communication et langages* 30.1, p. 6-19. DOI : 10.3406/colan.1976.4293.

Ringbom, Hakan (1983). « Borrowing and Lexical Transfer ». In : *Applied Linguistics* 4.3, p. 207-212. DOI : 10.1093/applin/4.3.207.

Riza, Hammam et al. (oct. 2016). « Introduction of the Asian Language Treebank ». In : *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE. DOI : 10.1109/icsda.2016.7918974.

Rogers, James Martin, Frank E. Daulton, Ian B.MacLean et Gordon A. Reid (2015). « Is native speaker intuition reliable forhigh-frequency context creation ? » In : *Journal of Inquiry and Research* 102, p. 57-69.

Roy, S.N. et J. Roy (1946). *Vocabulary Tests No. 2. Easy Bengali Vocabulary*. Unpublished Monograph. Kolkata : Indian Statistical Institute.

Ryan, John Charles (2017). « From Padauk to Hyacinth ». In : *Southeast Asian Ecocriticism : Theories, Practices, Prospects*, p. 179.

Sato, Satoshi (2014). « Text readability and word distribution in Japanese ». In : *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, p. 2811-2815.

Savický, Petr et Jaroslava Hlavácová (déc. 2002). « Measures of Word Commonness ». In : *Journal of Quantitative Linguistics* 9.3, p. 215-231. DOI : 10.1076/jqul.9.3.215.14124.

Schmid, Helmut (1994). « Probabilistic Part-of-Speech Tagging Using Decision Trees ». In : *Proceedings of International Conference on New Methods in Language Processing*.

Schmitt, Norbert (18 août 2010). *Researching Vocabulary. A Vocabulary Research Manual*. Springer-Verlag GmbH.

Schmitt, Norbert et Bruce Dunham (oct. 1999). « Exploring native and non-native intuitions of word frequency ». In : *Second Language Research* 15.4, p. 389-411. DOI : 10.1191/026765899669633186.

Sharoff, Serge, Elena Umanskaya et James Wilson (avr. 2014). *A Frequency Dictionary of Russian*. Routledge. DOI : 10.4324/9781315852157.

Silberztein, Max (2015). *La formalisation des langues : l'approche de NooJ*. ISTE Group.

Soares, Ana Paula, João Machado, Ana Costa, Álvaro Iriarte, Alberto Simões, José João de Almeida, Montserrat Comesaña et Manuel Perea (2015). « On the advantages of word frequency and contextual diversity measures extracted from subtitles : The case of Portuguese ». In : *Quarterly Journal of Experimental Psychology* 68.4, p. 680-696.

Soe Lai Phyue (2020). « Development of the lexico-conceptual knowledge resource for Myanmar NLP applications ». In : *Scientific Journal of Pure and Applied Sciences* 9.3, p. 915-925.

Sorell, C. Joseph (avr. 2014). *Word Frequencies*. Sous la dir. de John R Taylor. Oxford University Press. DOI : 10.1093/oxfordhb/9780199641604.013.005.

Southeast Asian Languages Library (2006). Sous la dir. de The University of Wisconsin-Madison Center for Southeast Asian Studies et Center for Research in Computational Linguistics. URL : http://sealang.net/burmese/.

Spaulding, Seth (déc. 1956). « A Spanish Readability Formula ». In : *The Modern Language Journal* 40.8, p. 433-441. DOI : 10.1111/j.1540-4781.1956.tb02145.x.

St John, Richard Fleming St. Andrew (1894). *A Burmese reader : being an easy introduction to the written language and companion to Judson's grammar ; for the use of civil service students and others who wish to acquire the language quickly and thoroughly*. Oxford : The Clarendon press.

– (1911). *Burmese self-taught. in Burmese and Roman characters, with phonetic pronunciation (Thimm's system)*. London : E. Marlborough.

Stecklow, Steve (15 août 2018). « Why Facebook is losing the war on hate speech in Myanmar ». In : sous la dir. de Reutors. URL : https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/ (visité le 04/05/2020).

Stewart, John Alexander (1955). *Manual of Colloquial Burmese*. London : Luzac et Company.

Su Su Yee, Sann, Chenchen Ding, Khin Soe, Masao Utiyama et Eiichiro Sumita (fév. 2019). « Modifying NOVA-annotated Myanmar Data to Universal Part-of-Speech Tagset ». In :

Tanaka-Ishii, Kumiko et Hiroshi Terada (2018). « Word Familiarity and Frequency ». In : *CoRR* abs/1806.03431. arXiv : 1806.03431.

Tanaka-Ishii, Kumiko, Satoshi Tezuka et Hiroshi Terada (juin 2010). « Sorting texts by readability ». In : *Computational Linguistics* 36, p. 203-227. DOI : 10.1162/coli.09-036-R2-08-050.

Taw Sein Ko (1898). *Elementary hand-book of the Burmese language*. Rangoon : Printed by the Superintendent, Government Printing.

Than Than Win (1998). « Burmese-English accent : Description, causes, and consequences ». Thèse de doct. Northern Illinois University.

Thorndike, Edward L. et Irving Lorge, éd. (1944). *The Teacher's Word Book of 30,000 Words*. New York : Bureau of Publications, Teachers' College, Columbia University.

Thorndike, Edward L., éd. (1921). *The Teacher's Word Book*. New York City : Teachers' College, Columbia University.

Tiberius, Carole et Tanneke Schoonheim (déc. 2013). *A Frequency Dictionary of Dutch*. Routledge. DOI : 10.4324/9781315857480.

Tiedemann, Jörg (mai 2012). « Parallel Data, Tools and Interfaces in OPUS ». In : *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. Istanbul, Turkey : European Language Resources Association (ELRA), p. 2214-2218.

Tin Htay Hlaing et Yoshiki Mikami (2 mai 2011). « Conference on Human Language Technology for Development ». In : *Collation Weight Design for Myanmar Unicode Texts*. Alexandria, Egypt.

Tognini-Bonelli, Elena (2001). *Corpus linguistics at work*. Amsterdam Philadelphia : J. Benjamins. ISBN : 9789027222763.

Tono, Yukio, Makoto Yamazaki et Kikuo Maekawa (déc. 2013). *A Frequency Dictionary of Japanese*. Routledge. DOI : 10.4324/9781315823287.

Tournier, Jean et Nicole Tournier (2017). *Dictionnaire de lexicologie française*. Paris : Ellipses. ISBN : 9782340021785.

Trask, R. L. (2004). « What is a word ? » In : *University of Sussex Working Papers in Linguistics*.

Tuldava, Juhan (1993). « Measuring text difficulty ». In : *Glottometrika* 14, p. 69-81.

Tun Thura Thet, Jin-Cheon Na et Wunna Ko Ko (2008). « Word segmentation for the Myanmar language ». In : *J. Inform. Sci.* 34.5, p. 688-704.

U Po Hla (2008). *Revised Myanmar Orthographical Vocabulary*. Yangon : Seikku Cho Cho.

UNGEGN (2013). *Report on the Current Status of United Nations Romanization Systems for Geographical Names - Burmese, Version 4.0*. Report. United

Nations Group of Experts on Geographical Names Working Group on Romanization Systems. URL : http://www.eki.ee/wgrs/rom2_my.pdf.

Unicode Consortium (2019a). « Chapter 16, Southeast Asia ». In : *The Unicode Standard, Version 12.0.0*. Mountain View, CA : The Unicode Consortium. ISBN : 978-1-936213-22-1. URL : http://www.unicode.org/versions/Unicode13.0.0/ch16.pdf.

– (2019b). « Chapter 4, Character Properties ». In : *The Unicode Standard, Version 12.0.0*. Mountain View, CA : The Unicode Consortium. ISBN : 978-1-936213-22-1. URL : http://www.unicode.org/versions/Unicode12.0.0/.

– (2019c). *Frequently Asked Questions, Myanmar Scripts and Languages*. URL : https://www.unicode.org/faq/myanmar.html.

– (2019d). *The Unicode Standard, Version 12.1.0*. URL : http://www.unicode.org/versions/Unicode12.0.0/.

– (2019e). *Unicode 12.1 Character Code Chart Myanmar*. URL : http://www.unicode.org/charts/PDF/U1000.pdf.

Utiyama, Masao et Eiichiro Sumita (2015). « Open collaboration for developing and using Asian Language Treebank (ALT) ». In : ASEAN IVO Forum.

Van Heuven, Walter JB, Pawel Mandera, Emmanuel Keuleers et Marc Brysbaert (2014). « SUBTLEX-UK : A new and improved word frequency database for British English ». In : *Quarterly journal of experimental psychology* 67.6, p. 1176-1190.

Vander Beke, George E. (1932). *French Word Book*. Macmillian. 188 p.

Vittrant, Alice (2004). « La modalité et ses corrélats en birman : dans une perspective comparative ». Thèse de doctorat dirigée par Mazaudon, Martine Linguistique générale Paris 8 2004. PhD Thesis.

– (2005). « Comment constituer son corpus d'étude : exemple d'une enquête linguistique sur le birman vernaculaire ». In : *Actes du Colloque Recueil des données en Sciences du langage et constitution de corpus : données, méthodologie, outillage*.

– (juin 2019). « Burmese ». In : *The Mainland Southeast Asia Linguistic Area*. Sous la dir. d'Alice Vittrant et Justin Watkins. De Gruyter Mouton, p. 56-130. DOI : 10.1515/9783110401981-003.

Wang, Karen Ming-Tzu et Paul Nation (2004). « Word meaning in academic English : Homography in the academic word list ». In : *Applied linguistics* 25.3, p. 291-314.

Wang, Rui, Haipeng Sun, Kehai Chen, Chenchen Ding, Masao Utiyama et Eiichiro Sumita (nov. 2019). « English-Myanmar Supervised and Unsupervised NMT : NICT's Machine Translation Systems at WAT-2019 ». In : *Proceedings of the 6th Workshop on Asian Translation*. Hong Kong, China : Association for Computational Linguistics, p. 90-93. DOI : 10.18653/v1/D19-5209.

West, M. (1953). *A general service list of English words*. London : Longman, Green & Co.

Wheatley, Julian K. (1982). *Burmese : a Grammatical Sketch*.

– (2018). « Burmese ». In : *The World's Major Languages*. Sous la dir. de Bernard Comrie. Routledge.

Wierzbicka, Anna (1996). *Semantics : Primes and universals : Primes and universals*. Oxford University Press, UK.

Wilkins, David Arthur (1972). *Linguistics in language teaching*. Edward Arnold.

Win Pa Pa et Ni Lar Thein (2008). « Myanmar word Segmentation using Hybrid Approach ». In : *Proc of ICCA*, p. 16-170.

Win Win Thant, Tin Myat Htwe et Ni Lar Thein (sept. 2011). « Statistical Function Tagging and Grammatical Relations of Myanmar Sentences ». In : *Computer Science Conference Proceedings*. Academy & Industry Research Collaboration Center (AIRCC). DOI : 10.5121/csit.2011.1319.

Winskel, Heather, Prakash Padakannaya et Aparna Pandey (2011). « Eye movements and reading in the alphasyllabic scripts of South and Southeast Asia ». In : *South and Southeast Asian Psycholinguistics*. Sous la dir. d'Heather Winskel et Prakash Padakannaya. Cambridge University Press, p. 315-326. DOI : 10.1017/cbo9781139084642.035.

Wray, Alison (mar. 2014). *Why Are We So Sure We Know What a Word Is ?* Sous la dir. de John R Taylor. Oxford University Press. DOI : 10.1093/oxfordhb/9780199641604.013.032.

Xiao, Richard, Paul Rayson et Tony McEnery (sept. 2015). *A Frequency Dictionary of Mandarin Chinese*. Routledge. DOI : 10.4324/9780203883075.

Xinhuanet (7 déc. 2017). *More than 200 foreign students learn Myanmar language this year*. URL : http://www.xinhuanet.com//english/2017-12/07/c_136807948.htm (visité le 10/04/2018).

Xu, Jiajin (2015). « Corpus-based Chinese studies : A historical review from the 1920s to the present ». In : *Chinese Language and Discourse* 6.2, p. 218-244.

Yadana Aung (2020). *Advancing in Burmese : A Drill Book for Intermediate to Advanced Learners*. Yangon : Katha Bridge Publishing.

Ye Kyaw Thu, Andrew Finch, Yoshinori Sagisaka et Eiichiro Sumita (2013). « A Study of Myanmar Word Segmentation Schemes for Statistical Machine Translation ». In : *11th International Conference on Computer Applications (ICCA 2013)* (26 fév. 2013). Yangon, Myanmar, p. 167-179.

Ye Kyaw Thu, Andrew Finch, Eiichiro Sumita et Yoshinori Sagisaka (2014). « Integrating dictionaries into an unsupervised model for Myanmar word segmentation ». In : *Proceedings of WSSANL*, p. 20-27.

Yêkháun, Mînn Latt (1966). *Modernization of Burmese*. English. Oriental Institute in Academia, Publishing House of the Czechoslovak Academy of Sciences Prague, p. 349.

Yin May Oo, A Theeraphol, Chen Fang Li, Pasindu De Silva, Supheakmungkol Sarin, Knot Pipatsrisawat, Martin Jansche, Oddur Kjartansson et Alexander Gutkin (2020). « Burmese Speech Corpus, Finite-State Text Normalization and Pronunciation Grammars with an Application to Text-to-Speech ». In : *Proc. 12th Language Resources and Evaluation Conference (LREC 2020)*. 11–16 May, Marseille, France, p. 6328-6339.

Yin, Binyong, Li Leyi et Jin Huishu (29 juin 2012). *Basic Rules of the Chinese Phonetic Alphabet Orthography* 汉语拼音正词法基本规则 *Hànyǔ Pīnyīn Zhèngcífǎ Jīběn Guīzé – GB/T 16159–2012*. Standard. AQSIQ, SAC.

Zin Maung Maung et Y Mikami (2008). « A Rule-based Syllable Segmentation of Myanmar Text ». In : *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, January 2008*. January 2008. Hyderabad, India, p. 51-58.

Zipf, George Kingsley (1945). « The meaning-frequency relationship of words ». In : *The Journal of general psychology* 33.2, p. 251-256.

– (1949). *Human behavior and the principle of least effort : An introduction to human ecology*. Addison-Wesley Press.

Liste des tableaux

Table des figures

Table des matières

Jennifer Lewis-Wong

Enjeux et méthodes pour la création de corpus en langues peu dotées. Application à la classification de textes pour l'apprentissage du birman.

Résumé

Trouver du matériel de lecture adapté aux apprenants de langues peu enseignées est un problème courant, tant pour les apprenants que pour les enseignants. Le traitement automatique offre des méthodes prometteuses pour faciliter ce processus. Comme leur mise en œuvre nécessite des corpus d'entraînement spécifiques à la langue, et que ces langues sont également peu dotées, la qualité des corpus est encore plus importante. Il nous a semblé nécessaire de considérer les particularités de la langue et de l'informatisation de son système d'écriture et le contexte d'utilisation du corpus, les études en linguistique et en lexicographie, les aspects culturels et même la tradition d'enseignement, car les apprenants sont probablement davantage influencés par les ressources existantes lorsqu'elles sont peu nombreuses. Cette thèse porte sur une méthode d'évaluation lexicale de textes pour le birman langue étrangère. D'abord la création de deux types de corpus : des textes authentiques et des ressources didactiques, ce dernier renseignant comment segmenter en unités minimales d'analyse ou « mots », prétraitement nécessaire car le birman ne les délimite pas par des espaces. Nous prenons également en compte les aspects culturels et la fréquence conjointe des syllabes dans l'entraînement d'un outil de segmentation. Les textes authentiques sont utilisés pour créer une liste de fréquences lexicales, utilisant la méthode de la fréquence réduite moyenne pour tenir compte de la dispersion. Cette liste est utilisée pour entraîner une SVM afin de classer les textes par difficulté croissante, méthode purement lexicale et prometteuse pour les langues peu dotées.

Mots-clés Birman, Apprentissage des langues étrangères, Langues peu dotées, Création de corpus, Fréquence lexicale, Lisibilité

Abstract

Finding reading material suitable for learners of less commonly taught languages is a common issue, both for learners and teachers. Natural language processing offers promising methods to facilitate the selection process. Since their implementation requires language specific training corpora, and such languages are also less well-resourced, corpus quality is even more important. We have found it necessary to take in to account not only the particularities of the language and how its writing system is computerised, but also the context of how the corpus is to be used, considering aspects such as orthography, studies in linguistics and lexicography, cultural aspects and even the teaching tradition, as students are probably more influenced by existing resources when they are scarce. This thesis looks at the application of a method for text evaluation for learners of Burmese as a foreign language. We detail the creation of two types of corpora : authentic texts and didactic resources, using the second type to inform how the authentic texts are segmented into minimal units of analysis or "words", a necessary pretreatment as Burmese does not delimit words with spaces. We also take into account cultural aspects and ngram syllable frequency in training a dictionary-based segmentation tool. The authentic text corpora are then used to create a general lexical frequency list, using the averaged reduced frequency method to account for dispersion. This list is then used to create a support vector machine to order texts by increasing difficulty using solely lexical data, a method that is promising for less-resourced languages.

Keywords Burmese, Foreign language learning, Under-resourced languages, Corpus creation, Lexical frequency, Readability