Institut National des Langues et Civilisations Orientales

Département Textes, Informatique, Multilinguisme

La lexicographie bilingue en traduction automatique d'une langue peu dotée : une chaîne opératoire pour l'amharique

MASTER

TRAITEMENT AUTOMATIQUE DES LANGUES

Parcours :

Recherche et développement

par

Malik MARMONIER

Directeur de mémoire :

Damien Nouvel

Co-directrice :

Delombera Negga

Année universitaire 2021/2022

TABLE DES MATIÈRES

Remerciements **5**

Résumé **7**

Liste des figures **8**

Liste des tableaux **11**

Introduction **13**

I Enjeux conceptuels et état de l'art **17**

1 La Traduction automatique comme domaine, tâche et document **19**
1.1 Définitions (lexicographique et historiographique) 19
1.2 Tokenization 22
1.3 Architectures / Modèles 25
1.4 Évaluation 30

2 Les Langues peu dotées, l'amharique et la lexicographie bilingue **35**
2.1 Définitions (conventionnelle et opérationnelle) 35
2.2 Amharique 37
2.3 Lexicographie bilingue 48

II Expérimentations / chaîne opératoire **53**

3 Données fondamentales **55**
3.1 Introduction 55
3.2 Wolf Leslau et l'English-Amharic Context Dictionary 56
3.3 Images scannées/photographiées 59

4 Protocole **61**
4.1 Introduction 61
4.2 Segmentation et reconnaissance optique des caractères 61
4.3 Extraction d'un corpus de segments alignés 71
4.4 Entraînement de modèles et augmentation des données par rétrotraduction 77

5 Résultats et discussion **85**
5.1 Introduction 85
5.2 Résultats 86
5.3 Discussion 89

Conclusion générale **93**

Bibliographie **95**

Remerciements

J'ai l'agréable devoir de remercier M. Damien Nouvel d'avoir accepté de diriger le présent mémoire, et pour l'excellence et l'acuité de son conseil ainsi que son immense générosité intellectuelle qui m'a profondément touché et inspiré.

Je tiens également à remercier Mme Delombera Negga d'avoir accepté de co-diriger le présent mémoire, et pour son infinie gentillesse ainsi que l'inestimable concours de son enseignement à ma compréhension des enjeux liés à la traduction de l'amharique.

Je n'aurais garde d'oublier, dans mes remerciements, M. Demeke Asres Ayele, mon tuteur de stage à l'Université d'Addis-Abeba, qui m'a témoigné tant de bienveillance et d'égards ; qu'il trouve ici l'expression de ma plus profonde gratitude !

Le présent mémoire a mis en jeu de très nombreux aspects du TAL — outre la traduction automatique — et je ne m'acquitte que faiblement de ma dette à leur égard en remerciant chaleureusement ici tous les enseignants, passés et présents, des licences de Traitement Automatique des Langues / Traitement Numérique Multilingue de l'INALCO que j'ai suivies avec un immense plaisir, ainsi que ceux du Master TAL de l'INALCO et, par-delà, du programme "pluriTAL", pour la qualité de leur enseignement et leur dévouement à leurs étudiants.

Je voudrais enfin remercier l'INALCO qui m'a permis de tant apprendre sur le monde.

Résumé

La traduction automatique, en tant que champ de recherche scientifique et de développement technique, témoignait – au tournant des années 2020 – d'un intérêt croissant pour les applications de son objet d'étude aux langues dites "peu dotées", langues pour lesquelles les données nécessaires au développement de systèmes de traduction automatique neuronaux n'existent qu'en quantités infimes, les approches communément employées pour pallier ce déficit consistant à tenter d'obtenir ces données depuis le Web où les langues peu dotées souffrent pourtant d'un défaut notable de représentation. La présente étude s'est proposé d'extraire les données nécessaires à la compilation d'un corpus parallèle au départ de ressources alternatives – les seules images des pages de documents relevant de la lexicographie bilingue – et d'étudier, en les comparant, l'impact de telles données sur les performances de systèmes de traduction automatique, à la faveur d'une suite d'expériences appliquées à la paire de langue amharique-anglais.

Mots-clefs : traduction automatique, langue peu dotée, lexicographie bilingue, corpus parallèle, amharique

Liste des figures

1.1 Exemple de tokenisation BPE (a) et BPE-dropout multiples (b) ; les tirets noirs représentent les fusions possibles, les tirets verts les fusions effectives et les tirets rouges les fusions omises aléatoirement, ([Provilkov et al., 2020]) . . . 24
1.2 "X is all you need", Vinay Uday Prabhu . . . 26
1.3 Architecture Transformer . . . 27
1.4 "Multi-head Attention" . . . 28
1.5 "Scaled Dot-Product Attention" . . . 29
1.6 Visualisation du mécanisme de l'auto-attention permise par la librairie BertViz . . . 30
1.7 Principale implémentation du modèle d'évaluation COMET . . . 33
2.1 Version anglaise d'un formulaire de recensement éthiopien ; le huitième champ concerne la langue des individus recensés . . . 38
2.2 Stèle votive sabéenne présentant une inscription en écriture sudarabique, Musée du Louvre . . . 41
2.3 Alphasyllabaire amharique, Wikipedia . . . 42
2.4 Systèmes de chiffres alphabétiques grec, copte et éthiopien ; Wikipedia . . . 43
2.5 Exemples de décompositions selon la norme NFKD des caractères composés du standard Unicode, Wikipedia . . . 43
2.6 Table des caractères redondants de l'amharique contemporain, in [Teshome and Besacier, 2012] . . . 45
2.7 Métadonnées du corpus CC-100 ayant servi à l'entraînement du modèle XLM-R ([Conneau et al., 2020], p. 11) . . . 47
2.8 Quantité de données (à l'échelle logarithmique) issues de Wikipedia (barres oranges) et de CommonCrawl (barres bleues) ayant servi à l'entraînement du modèle XLM-R ([Conneau et al., 2020]) . . . 48
2.9 Liste de mots sur tablette d'argile assyrienne, British Museum . . . 48
2.10 "Dictionnaire bilingue" anglais-allemand, au sens usuel de ce terme en TA comme domaine . . . 50
3.1 Wolf Leslau en 2004, lors de la 32ème Conférence Nord Américaine de Linguistique Afro-Asiatique, Pete Unseth . . . 56
3.2 Exemple d'une page de l'English-Amharic Context Dictionary, [Leslau, 1973] 57
3.3 "ale [une sorte de bière]", exemple de *realia* glosé entre crochets . . . 58
3.4 Système de reprographie pour la photographie d'ouvrages fragiles en possession de la BULAC, BULAC . . . 59
3.5 "PDF-A", aperçu de la première page . . . 60
3.6 "PDF-B", aperçu de la première page . . . 60
4.1 Logo d'eScriptorium . . . 62
4.2 Sous-fichiers PDF importés dans l'interface d'eScriptorium reprenant l'ordre alphabétique de la macrostructure du dictionnaire . . . 63
4.3 Méthodes d'import des données brutes dans l'interface d'eScriptorium . . . 63

4.4 Mauvaise segmentation des textes en colonnes du modèle par défaut d'eScriptorium . . . 64
4.5 Interface pour la définition d'une ontologie des zones de textes d'eScriptorium . . . 64
4.6 Image PNG segmentée en zones de textes d'une page de dictionnaire (jaune = colonnes, bleu = en-tête) . . . 65
4.7 Interface pour la définition d'une ontologie des lignes de textes d'eScriptorium (cadre rouge) . . . 65
4.8 Image PNG segmentée en zones de textes et en lignes d'une page de dictionnaire, les rectangles de couleur à la gauche des lignes marquent leur type . . . 66
4.9 Structure par défaut du modèle de segmentation Kraken. . . . 68
4.10 Interface d'eScriptorium pour l'entrée manuelle des transcriptions . . . 68
4.11 Structure du modèle OCR *ad hoc* . . . 69
4.12 Données synthétiques produites manuellement au moyen d'un éditeur graphique sur la base de séquences de caractères amhariques générées aléatoirement . . . 70
4.13 Interface d'eScriptorium avec : l'image PNG brute d'une page de dictionnaire (à la gauche de l'écran), sa segmentation en zones de textes et en lignes (au centre) et sa transcription manuelle/automatique (à la droite de l'écran) . . . 71
4.14 Interface d'eScriptorium pour l'exportation des données de colonnes transcrites au format XML Alto . . . 71
4.15 Aperçu des données transcrites au format XML Alto . . . 72
4.16 Reconstitution des entrées et des sous-entrées du dictionnaire en dépit des sauts de ligne, de colonnes et de pages, au départ d'un document XML . . . 72
4.17 Logo de l'application Doccano . . . 72
4.18 Premier exemple d'annotation des entrées et sous-entrées au moyen de l'interface de l'application Doccano . . . 73
4.19 Second exemple d'annotation des entrées et sous-entrées au moyen de l'interface de l'application Doccano . . . 73
4.20 Aperçu d'un fichier JSON d'entrées annotées exporté depuis Doccano . . . 74
4.21 Aperçu d'un fichier CSV pour l'entraînement d'un modèle Bi-LSTM d'annotation automatique . . . 75
4.22 "Stats" d'un modèle d'annotation automatique des exemples anglais et de leurs traductions amhariques au sein des entrées du dictionnaire transcrit 75
4.23 Aperçu des segments anglais extraits d'un ouvrage relevant de la lexicographie bilingue . . . 76
4.24 Aperçu des segments amhariques extraits d'un ouvrage relevant de la lexicographie bilingue . . . 76
4.25 Trois corpus d'entraînement . . . 77
4.26 Logo de Fairseq . . . 78
4.27 Logo de COMET . . . 80
4.28 Trois modèles de TA comme tâche amharique → anglais représentés et nommés en fonction de leurs données d'entraînement NON augmentées . . . 82
4.29 Trois modèles de TA comme tâche anglais → amharique représentés et nommés en fonction de leurs données d'entraînement NON augmentées et visant à la production de données augmentées par rétrotraduction . . . 82

4.30 Trois modèles de TA comme tâche amharique → anglais représentés et nommés en fonction de leurs données d'entraînement augmentées 83

Liste des algorithmes

1 Phase d'apprentissage de l'algorithme BPE 23

Liste des tableaux

2.1 Décomposition du mot graphique *tawqaläcäw* en caractères et en morphèmes 44
4.1 Illustration du formalisme visant à l'abstraction du texte des entrées en vue de l'entraînement d'un modèle d'annotation automatique. De haut en bas : le texte de départ de la sous-entrée, la première phase de l'algorithme, la seconde phase (finale) de l'algorithme, les étiquettes au format BIO . . . 74
4.2 Nombre des segments et des mots graphiques compris dans chacun des corpus constitués pour l'entraînement de modèles de TA distincts 77
4.3 Nombre des segments et des mots graphiques compris dans chacun des corpus de validation et d'évaluation accompagné du détail de leurs sources. 78
4.4 Nombre des segments et des mots graphiques compris dans chacun des corpus augmentés G+, B+ et D+ 81
5.1 Scores BLEU résultant de l'évaluation automatique des modèles Transformer entraînés sur des données NON augmentées dans la direction amharique → anglais ; la présence d'une astérisque à la droite des valeurs-p, présentées entre parenthèses, indique que les performances du modèle considéré diffèrent significativement de celles de la baseline (Modèle B) . . 86
5.2 Résultats d'évaluation automatique (chrF et TER) des modèles entraînés sur des données NON augmentées dans la direction amharique → anglais ; (N.B. : Le score TER est un taux d'erreur et gagne, par conséquent, à être minimisé) 86
5.3 Résultats d'évaluation COMET des modèles Transformer amharique → anglais entraînés sur des données NON augmentées 87
5.4 Scores BLEU résultant de l'évaluation automatique des modèles anglais → amharique entraînés dans l'objectif de générer des données retraduites ; (on notera que les scores BLEU sont incommensurables d'une paire de langues à l'autre et d'une direction de traduction à l'autre) 87
5.5 Scores chrF et taux d'erreur TER résultant de l'évaluation automatique des modèles anglais → amharique entraînés dans l'objectif de générer des données retraduites 88
5.6 Résultats d'évaluation COMET des modèles Transformer anglais → amharique entraînés sur des données NON augmentées dans l'objectif de générer des données retraduites 88

5.7 Scores BLEU résultant de l'évaluation automatique des modèles Transformer amharique → anglais entraînés sur des données augmentées par rétrotraduction 88
5.8 Scores chrF et taux d'erreur TER, résultant de l'évaluation automatique des modèles amharique → anglais entraînés sur des données augmentées . 89
5.9 Résultats d'évaluation COMET des modèles Transformer amharique → anglais entraînés sur des données augmentées 89
5.10 Sources des segments et des mots graphiques du jeu de données d'évaluation des modèles. 90
5.11 Part des données d'entraînement des modèles issues de Gezmu et al. (en nombre de segments) 91
5.12 Ratios types/tokens comparés du corpus de segments parallèles issu d'un dictionnaire et d'un corpus compilé au départ du Web (moyenne calculée sur la base de 10 000 échantillons aléatoires de 100 segments de langue anglaise) 92

INTRODUCTION

Présentation Générale

La traduction automatique est l'application la plus ancienne, par temps de paix, des technologies de l'information à des données langagières et constitue, à ce titre, le berceau du traitement automatique des langues ([Cori, 2020], p. 88). Née au sortir de la Seconde Guerre Mondiale et inspirée à un influent bailleur de fonds de la recherche américaine par les progrès fulgurants de la cryptanalyse survenus au cours de ce conflit, la traduction automatique en tant que domaine de recherche scientifique et de développement technique est chargée d'une passionnante histoire.

Objet d'un vif — et naïf — engouement dès l'aube des années 1950 — la première démonstration publique d'un système de traduction automatique fut organisée dès 1954 et célébrée comme un succès — la traduction automatique devait pâtir, dès la seconde moitié des années 1960, de cet excès d'enthousiasme. Un rapport publié en 1966 par l'Automatic Language Processing Advisory Committee (ALPAC), notamment, faillit bien reléguer ce champ de recherche aux oubliettes de l'Histoire, du moins aux États-Unis, son lieu d'origine, où les études en traduction automatique disparurent presque corps et bien, pendant près de vingt ans, suite à la parution de ce fameux rapport.

Il fallut attendre le tournant des années 1980 pour que d'ambitieux projets industriels et scientifiques européens (Eurotra) et japonais (Ordinateurs de cinquième génération), faisant la part belle à la traduction automatique, n'achèvent de convaincre les bailleurs de fonds de la recherche américaine de la nécessité de reconquérir le terrain perdu. Les travaux de recherche qui s'ensuivirent devaient donner lieu à une révolution sans précédent en traduction automatique, à un changement radical du paradigme technique de cette discipline se concrétisant en la marginalisation progressive des approches dites "par règles" — jusqu'alors dominantes et qui reposaient sur le développement de parseurs et de transducteurs pour l'analyse de phrases source et la génération de leurs traductions cible — en faveur d'approches fondées sur l'analyse statistique de données textuelles bilingues, approches inspirées par des avancées contemporaines en reconnaissance automatique de la parole.

La traduction automatique, désormais "statistique", forte du généreux mécénat du complexe militaro-industriel américain de l'après-11 septembre, connut dès lors un fulgurant essor et fut, à titre d'exemple, la clef du succès du système de traduction automatique le plus connu du grand public : Google Translate.

Mais c'est un autre changement de paradigme, intervenu au milieu des années 2010, suite à l'adoption de techniques issues de l'apprentissage profond, qui devait offrir à la traduction automatique, désormais neuronale, ses plus grands succès d'estime.

Au crépuscule des années 2010, il se trouvait ainsi des chercheurs pour affirmer que la parité homme/machine avait été atteinte en traduction automatique ([Wu et al., 2016, Hassan et al., 2018]), et un nouvel algorithme appelé "Transfor-

mer" ([Vaswani et al., 2017]), fruit des travaux de recherche menés au sein des laboratoires de l'entreprise Google, était en passe de révolutionner, par delà la traduction automatique, le traitement automatique des langues dans son ensemble, voire même d'autres pans entiers de la recherche en intelligence artificielle.

À l'aube des années 2020, la traduction automatique, forte de ces succès, semble s'être mise en quête de nouveaux défis, et avoir jeté son dévolu sur l'application de son objet d'étude à des langues dites "peu dotées" ("*low-resource languages*"), tels le swahili, le mongol ou l'amharique, soit à des langues pour lesquelles les données nécessaires au développement de systèmes de traduction automatique n'existent qu'en quantités infimes.

Cette prééminence des travaux relatifs aux langues peu dotées au sein de la recherche contemporaine en traduction automatique trouve sans doute sa plus vive illustration dans la publication quasi-simultanée, en 2021, de pas moins de trois rapports ("*surveys*") indépendants exclusivement consacrés à la traduction automatique des langues peu dotées ([Ranathunga et al., 2021, Wang et al., 2021, Haddow et al., 2021]), un sujet qui bénéficie par ailleurs, désormais, de son propre cycle de conférences-ateliers annuelles — LoResMT, dont la cinquième édition se tenait, en septembre 2022, en Corée du Sud [1] — mais dont la popularité s'étend bien au-delà du seul cadre de cette manifestation : les organisateurs de la quinzième conférence généraliste de l'Association for Machine Translation in the Americas (AMTA) estimaient ainsi, à titre d'exemple, en septembre 2022, que les problématiques liées à la traduction automatique des langues peu dotées avaient constitué le sujet de recherche le plus traité cette année-là, représentant à lui-seul près du tiers des travaux soumis par les chercheurs [2].

Or cet engouement pour les questions liées aux applications de la traduction automatique aux langues dites "peu dotées" semblait également donner lieu, en ce début de décennie 2020, à une curieuse escalade verbale, s'incarnant dans l'émergence progressive de travaux portant sur des langues qualifiées de "très peu dotées" ("*very low-resource*") ([Baruah et al., 2020]), de "vraiment peu dotées" ("*truly low-resource*") ([Bustamante et al., 2020]), puis d'"extrêmement peu dotées" ("*extremely low-resource*") ([Tars et al., 2021]), la palme de cette surenchère revenant sans conteste à Google AI, un laboratoire de recherche dont certains des chercheurs évoquaient, en mai 2022, au titre d'un article de blog [3] la traduction automatique de langues "zéro-dotées" ("*zero-resource*"), parmi lesquelles figureraient, selon leurs dires, bon nombre de langues africaines, tels le lingala, le bambara, l'oromo ou le tigrigna ([Caswell and Bapna, 2022]).

Le point de départ de la présente étude est le sentiment de circonspection interloquée que ne manque pas de susciter chez l'étudiant en langues orientales la lecture d'un tel article. Ces langues désignées comme "zéro-dotées" sont, en effet, loin d'être sans ressources : de nombreuses grammaires et moults dictionnaires, riches en exemples traduits, ont été produits pour ces idiomes par d'éminents savants, comme suffirait à en attester une rapide consultation du catalogue en ligne de la Bibliothèque Universitaire des Langues et Civilisations (BULAC), familière des étudiants de l'INALCO.

Le présent mémoire se propose d'explorer la possibilité d'extraire les ressources nécessaires au développement d'un système de traduction automatique depuis de tels

1. https://sites.google.com/view/loresmt/home
2. Cf. https://aclanthology.org/2022.amta-research.pdf, page V.
3. https://ai.googleblog.com/2022/05/24-new-languages-google-translate.html

documents, notamment au départ d'ouvrages relevant de la lexicographie bilingue des langues orientales, d'étudier les mécanismes possibles d'une telle extraction et l'impact des données obtenues, en aval, sur les performances d'un système de traduction automatique, en prenant appui sur l'exemple de la langue amharique (Éthiopie).

Ce mémoire voudrait par ailleurs offrir aux chercheurs en langues orientales et aux TAListes novices une "recette" possible à suivre pour le développement d'un système de traduction automatique "au départ de rien" ("*from scratch*") ou presque ! Cette ambition explique la mention, au sous-titre de ce mémoire, du terme de "chaîne opératoire", emprunté à l'anthropologie française des techniques ([Coupaye, 2022]), et qui désigne dans ces études l'ensemble des opérations mises en jeu dans la production d'artefacts (telles des poteries, par exemple), allant du repérage d'une source de matières premières (prospection d'un puits d'argile), à l'extraction des-dites matières premières (argile meuble), à la fabrication des artefacts eux-mêmes (façonnage, cuisson) et à leur usage (comme récipients ou offrandes). De manière similaire, le protocole expérimental présenté en seconde partie de ce mémoire voudrait permettre à quiconque le souhaiterait de mettre au point un système de traduction automatique, au départ de simples sources documentaires (dictionnaires bilingues imprimés) d'où serait extraite la matière première (un corpus parallèle de segments de textes informatiques) permettant le développement d'un système de traduction automatique (par entraînement d'un modèle neuronal) dont les performances d'usage seraient évaluées (par des métriques automatiques).

Plan de lecture

Ce mémoire s'organise en deux parties principales :

— Une première partie consiste en une présentation des enjeux conceptuels et de l'état de l'art de la problématique abordée et porte notamment sur les notions de traduction automatique, de langue peu dotée et de lexicographie bilingue, et sur les enjeux proprement liés à la traduction automatique de l'amharique.

— Une suite d'expériences en forme de chaîne opératoire est présentée dans la seconde partie de ce mémoire, appliquée à la paire de langues amharique-anglais.

Première partie

Enjeux conceptuels et état de l'art

CHAPITRE 1

LA TRADUCTION AUTOMATIQUE COMME DOMAINE, TÂCHE ET DOCUMENT

Sommaire

1.1 Définitions (lexicographique et historiographique) **19**
1.1.1 Définition lexicographique 19
1.1.2 Définition "historiographique" 21
1.2 Tokenization **22**
1.3 Architectures / Modèles **25**
1.4 Évaluation **30**
1.4.1 BLEU 32
1.4.2 COMET 33

1.1 Définitions (lexicographique et historiographique)

Les ouvrages dédiés à la traduction automatique (TA) sont avares de définitions relatives à leur objet d'étude. Ainsi, aucun des deux manuels universitaires de référence ([Koehn, 2010, Koehn, 2020]) consacrés à cet unique thème ne semble offrir au lecteur la moindre analyse de la signification même du terme de "traduction automatique" ("*machine translation*"), et les deux ouvrages de vulgarisation publiés, sur ce même sujet, par Thierry Poibeau, en anglais ([Poibeau, 2017]) et en français ([Poibeau, 2019]), semblent présenter la même lacune, suivis en cela par nombre de thèses de doctorat dédiées à la TA.

Or si le terme de "traduction automatique" peut, sans doute, être légitimement considéré comme transparent, il n'est toutefois pas dénué d'ambiguïtés, comme tend à le démontrer une rapide analyse d'ordre lexicographique de ses occurrences au sein des résultats du moteur de recherche Google, et vaut ainsi d'être défini.

1.1.1 Définition lexicographique

Le terme de "traduction automatique" est, en effet, susceptible de revêtir au moins trois sens distincts selon le contexte de ses occurrences [1], désignant ainsi un domaine de recherche scientifique et de développement technique dans les exemples suivants :

— "SYSTRAN : 50 ans d'innovation en Traduction Automatique"

1. Occurrences relevées, à l'été 2022, parmi les résultats du moteur de recherche Google pour les mots-clefs "traduction automatique".

— "les avancées en traduction automatique pourraient au contraire permettre de progresser vers une collaboration plus efficace entre humain et machine"
— "On l'appelle « décodeur » parce qu'en traduction automatique, les langues sont considérées comme des séries de codes qu'il faut décoder."
— "On s'accorde généralement pour attribuer la paternité de la recherche en traduction automatique à Warren Weaver"
— "Le problème des noms propres est sans doute l'un des plus difficiles à résoudre en traduction automatique"
— "Par la suite, Y. Bar-Hellal [sic] a organisé, en 1962, la première conférence en traduction automatique."
— "Le Laboratoire d'Informatique de l'Université du Maine (LIUM) recrute un maître de conférence en traduction automatique statistique."
— "il est aussi chercheur en traduction automatique au Mit [sic], au research laboratory of electronics, et adhère à l'idée de la TA comme technologie de guerre."

Alors même que d'autres occurrences de ce terme semblent rapprocher la notion de "traduction automatique" d'un procédé ou bien d'une tâche :

— "Dans tous les cas, il ne faut jamais considérer le résultat d'une traduction automatique comme un produit fini."
— "WPML propose une traduction automatique optimisée par Microsoft, Google ou DeepL."
— "Une traduction automatique de qualité en 26 langues, c'est possible !"
— "Les logiciels de traduction automatique statistique produisent une traduction de qualité satisfaisante lorsque des corpus suffisamment volumineux et pertinents sont disponibles."
— "il est important d'utiliser des outils de traduction automatique fiables et efficaces."
— "Le logiciel de traduction automatique analyse le texte source et crée une représentation transitoire"
— "Google, Microsoft, Amazon et Systran ont récemment mis au point des logiciels de traduction automatique d'une qualité sans précédent"
— "Si les systèmes de traduction automatique permettent d'obtenir des résultats à première vue corrects, rien ne vaut la traduction humaine"

Quand, dans d'autres contextes encore, ce même terme fait indéniablement référence à des documents :

— "Post-édition d'une traduction automatique : choses à savoir et outils à utiliser."
— "Vos traductions automatiques disponibles immédiatement. Il arrive que l'on ait besoin d'une traduction automatique dans des situations parfois imprévues."
— "des phrases (...) accompagnées ou non d'une traduction ou d'une traduction automatique"
— "Une traduction automatique qui a été éditée par un traducteur humain, peut être considérablement meilleur marché qu'une traduction humaine traditionnelle."
— "[L]a solution la plus courante consiste simplement à comparer une traduction automatique avec des traductions manuelles de référence au moyen de séquences contiguës de n mots appelées « n-grammes »."
— "L'utilisateur ne prend pas en compte les erreurs car il sait qu'il s'agit d'une

traduction automatique dont vous n'êtes pas responsable."
— "Évidemment, il s'agit d'une traduction automatique par Google Translate et la traduction n'est donc pas toujours parfaite"

On s'attachera par conséquent — et par acquit de conscience — à désambiguïser, autant que faire se pourra, dans le restant de cette étude, les notions de TA comme domaine, comme tâche et comme document, en en retenant les définitions suivantes, semi-circulaires :

— la TA comme tâche sera définie, d'après Dorothy Kenny (in [Baker and Saldanha, 2019], p. 306) comme "la conversion automatique d'un texte d'une langue naturelle vers une autre" ;
— la TA comme domaine, comme l'étude et l'application des principes, des méthodes et des techniques visant au développement de systèmes capables d'effectuer la conversion automatique d'un texte d'une langue naturelle vers une autre ;
— et la TA comme document, comme le ou les texte(s) résultant de la conversion automatique de textes d'une langue naturelle vers une autre.

Si, toutefois, les définitions d'ordre "lexicographique" de la notion de "traduction automatique" sont rares sous la plume des chercheurs, cela s'explique sans doute, en grande partie, par le fait que la TA comme domaine tend à privilégier pour elle-même des approches définitionnelles relevant d'un tout autre ordre, qui pourrait être qualifié "d'historiographique", ainsi qu'en témoigne, à titre d'exemple, le choix du chercheur Thierry Poibeau, mentionné précédemment, d'organiser chacun de ses deux ouvrages consacrés à ce domaine selon un plan chronologique plutôt que thématique.

1.1.2 Définition "historiographique"

Il convient de souligner, en effet, que — d'une manière remarquable dans le contexte notoirement anhistorique du traitement automatique des langues ([Hirst, 2013]) — la TA comme domaine a fait l'objet d'un exemplaire travail d'historicisation, et ce sous l'impulsion notable de W. John Hutchins, bibliothécaire universitaire de formation et auteur de brillants travaux dédiés à l'Histoire de ce champ d'étude. [2]

Cette historicisation tend à s'articuler autour de pivots consistant en des changements, abruptes ou progressifs, du paradigme technique dominant de la TA comme tâche, à l'exemple du passage, au tournant des années 1990, d'approches dites "par règles" à celles dites "statistiques" — [Way, 2010] offre une intéressante analyse rétrospective des ressorts humains et des tensions qui accompagnèrent cette évolution. La dernière en date de ces révolutions techniques fut marquée par le basculement du domaine vers des approches dites "neuronales" de la TA comme tâche, lequel basculement — pour paraphraser Ernest Hemingway — se produisit "graduellement, puis soudainement" !

En effet, si l'idée d'appliquer des "réseaux de neurones" ("*neural networks*") à la TA comme tâche semble attestée dès la fin des années 1980 et tout au long des années 1990 ([Koehn, 2020], p. 39), la mise en œuvre concrète de telles approches

2. Des travaux à nouveau consultables en ligne (https://mt-archive.net/), à l'été 2022, au sein d'une archive maintenue par l'EAMT (European Association for Machine Translation), suite au décès de l'auteur en janvier 2021.

n'a toutefois progressé qu'à pas lents. Ces réseaux de neurones n'ont ainsi été appliqués, tout d'abord, qu'à des aspects périphériques de la TA comme tâche (telle l'évaluation automatique des traductions générées), puis plus centraux (tels les modèles de langue sur lesquels prenaient appui les décodeurs de la TA statistique) ; et il fallut attendre les années 2010 pour qu'aient lieu les premières tentatives d'implémentations purement neuronales de la TA comme tâche, au moyen de réseaux de neurones convolutifs ("*convolutional neural networks*") et récurrents ("*recurrent neural networks*"), quoique les résultats médiocres obtenus sur la base de ces approches n'offrissent un répit à la TA statistique. C'est finalement l'introduction par [Bahdanau et al., 2016] du mécanisme de l'attention — une forme d'alignement pondéré entre les représentations des mots source et cible en traduction automatique — qui fut le véritable déclencheur d'une ruée vers ce nouveau paradigme neuronal qui devait redéfinir, de fond en comble, l'état de l'art du domaine.

La suite de ce chapitre est consacrée à la description de trois des aspects majeurs de cet état de l'art de la TA neuronale : la tokenisation, les architectures de ses modèles, et l'évaluation.

1.2 Tokenization

Les textes électroniques, du "point de vue" des logiciels qui en assurent le traitement, ne consistent qu'en des séquences de valeurs binaires. Ces suites de 0 et de 1 représentent les glyphes des systèmes d'écriture des cultures du monde, en vertu de conventions arbitraires établies pour l'encodage des caractères électroniques, tels ISO-8859-1, MacRoman, Windows5210, ou encore les diverses implémentations du standard Unicode.

Dans le monde poético-fantastique d'une nouvelle de Jorge Luis Borges, la TA comme tâche pourrait être réduite à la simple consultation d'une table d'indices ("*lookup table*"), de dimensions infinies, laquelle associerait, à chaque séquence possible de valeurs binaires encodant les textes d'une langue, une autre séquence de valeurs binaires encodant un texte parfaitement équivalent, de par son sens, sa forme et sa fonction, écrit dans une autre langue. L'unité de traduction sur laquelle opérerait un tel système de "TA magique" serait dès lors le texte lui-même, tout d'un bloc.

Dans le monde réel, néanmoins, la TA comme tâche, pour être applicable à une part substantielle de l'infinité des énoncés possibles des langues, se doit de mettre à profit les propriétés analogiques et génératives du langage humain, et doit dès lors segmenter ces énoncés en sous-unités, traditionnellement appelées "tokens" en traitement automatique des langues (TAL) ; le processus de segmentation d'un texte électronique en tokens étant lui-même désigné du nom de "tokenisation" ("*tokenization*").

La tokenisation, en TAL, est communément accomplie à deux échelles distinctes. À un niveau dit "macroscopique" ([Mielke et al., 2021]) les textes électroniques sont ordinairement segmentés en phrases, sur la base de conventions typographiques propres à chaque langue[3] ; et à un niveau dit "microscopique" ([Ibid.]) ces phrases

3. Une tâche non-triviale s'il en est, dans la mesure où certain marqueurs de fin de phrases sont susceptibles d'être ambigus, tel, par exemple, le point dans l'énoncé anglais suivant : "Mr. Fogg was on his way to the U.S.A.".

sont à leur tour segmentées en sous-unités dont la définition tend à recouper celle des mots-formes de la linguistique.

Une évolution remarquable de l'état de l'art du TAL contemporain, et de la TA comme tâche, tient à l'émergence d'approches nouvelles de la tokenisation "microscopique", consistant à segmenter, de manière automatique, les textes électroniques en sous-unités de dimensions inférieures à celles du mot-forme, permettant ainsi aux systèmes développés de capturer des informations susceptibles de représenter la morphologie ou la phonologie des signifiants des énoncés, et d'être mieux à même, du fait de ces économies d'échelle, de traiter les discours humains dans leurs infinies variations. Cette approche de la tokenisation "en deçà du mot" ou "en sous-mots" ("*sub-word tokenization*") est à mettre au crédit du "recyclage" par [Sennrich et al., 2016b] d'un algorithme de compression des données informatiques développé par [Gage, 1994] appelé "*byte-pair encoding*" ("encodage par paires d'octets") ou algorithme BPE. [4]

Algorithme 1: Phase d'apprentissage de l'algorithme BPE

Function BPE (*un corpus d'entraînement* C, *un nombre de fusions* k) **:**
  $V \leftarrow$ *l'ensemble des caractères dans* C
  for $i = 1$ **to** k **do**
    $\mathrm{t}_{gauche}, \mathrm{t}_{droit} \leftarrow$ *la paire de tokens adjacents la plus fréquente dans* C
    $\mathrm{t}_{nouveau} \leftarrow <\mathrm{t}_{gauche}, t_{droit}>$
    $V \leftarrow V + \mathrm{t}_{nouveau}$
    $C = C$ *au sein duquel la séquence* $<\mathrm{t}_{gauche}, \mathrm{t}_{droit}>$ *a été remplacée par* $\mathrm{t}_{nouveau}$
  end for
  return V *//vocabulaire ordonné*

Cet algorithme BPE — dans son implémentation originelle en TAL — procède en deux phases.

Dans un premier temps, l'algorithme instancie un vocabulaire comprenant l'ensemble des caractères présents au sein d'un corpus d'entraînement, additionné d'un caractère supplémentaire devant permettre la représentation des espaces [5] ; ce corpus d'entraînement est alors parcouru, de manière itérative, par l'algorithme afin d'y fusionner ("*merge*") la paire de symboles adjacents la plus fréquente pour lui substituer un token unique. Le pseudo-code de l'Algorithme 1 représente cette première phase dite d'apprentissage. La seconde phase de cet algorithme – celle de la tokenisation à proprement parler – consiste en l'application des opérations de fusion apprises lors de la constitution du vocabulaire des tokens, dans l'ordre exact de leur apprentissage, donnant ainsi à l'algorithme un caractère déterministe et glouton ("*greedy*").

Cette implémentation "primordiale" de l'algorithme BPE se heurte néanmoins au problème suivant : des sous-mots fréquents, susceptibles de représenter des propriétés sémantiques utiles des mots-formes qu'ils composent (tels par exemples des

4. Il convient de noter que contrairement à ce que le nom de l'algorithme pourrait laisser entendre, les algorithmes BPE n'opèrent généralement pas, dans le contexte du TAL, à l'échelle des octets mais bien à celle des points de code des caractères, deux notions nettement distinctes dès lors que les textes à traiter présentent des caractères non ASCII. Non sans confusion, cependant, il existe également des approches de la tokenisation BPE bel et bien fondées sur l'octet et qui sont généralement désignées du nom de Byte-level BPE ou BBPE (cf. [Wang et al., 2019]).

5. L'implémentation originelle de cet algorithme utilisait à cet effet l'interpoint <·>.

affixes morphologiques), tendent à se trouver systématiquement subsumés dans de plus vastes tokens, au cours de la seconde phase de l'algorithme BPE, privant potentiellement les opérations postérieures à la tokenisation — telles celles de la TA comme tâche — d'informations précieuses. Ainsi, à titre d'exemple, l'important suffixe "-ed" de la langue anglaise, marqueur du prétérit des verbes réguliers, tend — pour peu que les données d'entraînement parcourue par l'algorithme BPE au cours de sa phase d'apprentissage soient de dimensions substantielles — à n'être jamais tokenisé comme tel, mais intégré à d'autre séquences de caractères du fait même de sa haute fréquence : "ted", "med", "shed", etc.

Plusieurs solutions à ce problème ont été proposées qui s'emploient, fondamentalement, à tempérer le caractère déterministe et glouton de l'algorithme BPE. L'une de ces solutions, proposée par [Kudo and Richardson, 2018], met en jeu un système complexe de modèles de langues visant à classer de multiples tokenisations possibles d'un même segment sur la base de mesures d'entropie, dans l'objectif d'apprendre des règles de tokenisation optimales. Une autre approche — plus simple — appelée "BPE-dropout" et proposée par [Provilkov et al., 2020], ne consiste qu'en une altération mineure de l'approche originelle : dans cette configuration, l'algorithme, au cours de sa seconde phase, celle de la tokenisation "à proprement parler", applique au texte à tokeniser les opérations de fusion des tokens adjacents apprises en première phase, mais avec une probabilité inférieure à 1, laissant ainsi une "chance" aux bigrammes fréquents d'être représentés par un token indépendant.

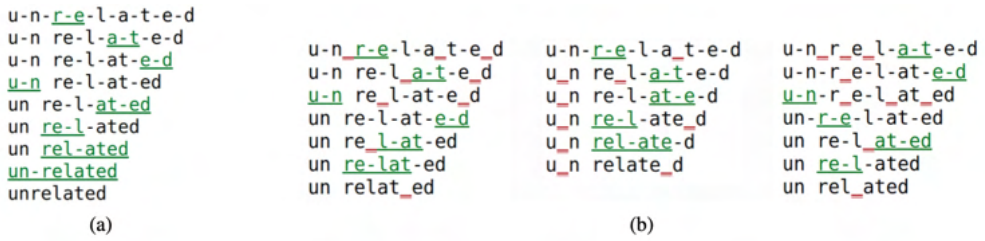


FIGURE 1.1 – Exemple de tokenisation BPE (a) et BPE-dropout multiples (b) ; les tirets noirs représentent les fusions possibles, les tirets verts les fusions effectives et les tirets rouges les fusions omises aléatoirement, ([Provilkov et al., 2020])

À l'été 2022, la direction dans laquelle pourrait évoluer l'état de l'art de la tokenisation est incertaine. Une tendance, néanmoins, semble se dessiner qui pourrait conduire à la disparition pure et simple de cet aspect du TAL. Ainsi, une phase dite de pré-tokenisation, consistant à détacher les ponctuations des caractères auxquels elles sont accolées, semble en passe de tomber en désuétude, du fait des nouvelles implémentations de l'algorithme BPE, telles celles offertes par la librairie SentencePiece [6], et il n'est pas inimaginable qu'à moyen terme la tokenisation, ou du moins la microtokenisation, suive le même chemin, dans un monde où les tokens seraient, par défaut, les caractères isolés des textes informatiques, une approche aux antipodes de celle magico-borgessienne évoquée précédemment, mais non moins remarquable ! Dans l'état de l'art actuel de la TA comme tâche, néanmoins, l'architecture dominante, appelée Transformer ([Vaswani et al., 2017]), et les modèles qui l'incarnent sont impropres au traitement de textes alphabétiques tokenisés en caractères [7], garantissant un sursis aux algorithmes de tokenisation précédemment évoqués. La sec-

6. https://github.com/google/sentencepiece
7. Du fait du degré de complexité quadratique du mécanisme de l'auto-attention qui les sous-tend.

tion suivante de cette étude est précisément dévolue à la présentation de cette architecture dite "Transformer".

1.3 Architectures / Modèles

La TA neuronale se réduit, en tant que tâche et en termes concrets, à une longue séquence d'opérations mathématiques effectuées par du matériel informatique, et peut être conceptualisée, dès lors, à un haut degré d'abstraction, comme l'application d'une très longue fonction composée — comportant potentiellement plusieurs millions de constantes ou de paramètres — à une variable X, consistant en une phrase source tokenisée, donnant un résultat Y : les tokens d'une traduction en langue cible de la phrase d'entrée. De telles fonctions peuvent être agencées de très nombreuses façons inventives, mais sont communément rassemblées, sur la base de similarités structurelles, en des catégories désignées du nom d'architectures dans le contexte de l'apprentissage profond.[8] Ces architectures, entités abstraites et théoriques, trouvent une incarnation concrète, matérielle et logicielle, dans des artefacts généralement désignés du nom de "modèles".

Les modèles sur lesquelles s'appuie la TA neuronale relèvent de la catégorie générale des architectures dites "encodeur-décodeur" ("*encoder-decoder architectures*")[9], lesquelles architectures sont fondamentalement structurées, à un haut niveau d'abstraction, de la façon suivante : un premier module, appelé "encodeur" ("*encoder module*"), reçoit en entrée un segment de texte tokenisé et génère au départ de ces données une représentation complexe de cette séquence de tokens, pour en capturer les propriétés sémantiques ; cette représentation est subséquemment transmise au second module de cette architecture — le décodeur ("*decoder module*")— lequel procède alors à la génération "auto-régressive" ("*autoregressive*"), token par token, de la traduction du segment d'entrée.

Les architectures encodeur-décodeur dominantes, en TA neuronale, ont longtemps été celles permises par les réseaux de neurones récurrents, notamment dans leurs variantes GRU ([Cho et al., 2014]) et LSTM ([Sutskever et al., 2014]) bidirectionnelles, associées à des mécanisme d'attention ([Bahdanau et al., 2016]) visant à calculer une forme d'alignement pondéré entre les tokens d'entrée et de sortie au sein des modèles. L'une des limitations de ces architectures tenait à leur traitement strictement séquentiel des données d'entrées, token par token, nécessitant de longs temps d'entraînement, ne tirant qu'un faible parti des processeurs graphiques, artefacts au cœur de l'apprentissage profond, et débouchant sur des modèles aux performances médiocres, notamment en traduction des segments longs. Des efforts visant à résoudre ces difficultés furent entrepris qui prirent appui sur des réseaux de neurones convolutifs ([Gehring et al., 2017]), mais ne rencontrèrent qu'un succès limité, pour des raisons liées aux faiblesses intrinsèques de ces approches en TAL (cf. [Hinton, 2017]).

Il fallut attendre l'année 2017 pour qu'un article soit publié qui devait révolutionner l'état de l'art de la TA neuronale et du TAL général. Ayant pour titre "Attention is all you need" ([Vaswani et al., 2017]), ce "papier", avec près de 46000 références à l'été 2022, est devenu l'un des plus cités par les chercheurs en intelligence artificielle, donnant même naissance à un "mème" de publications scientifiques dont la figure 1.2 ne documente qu'une partie.

8. Bien que le terme d'algorithme soit parfois également employé dans ce sens, par certains auteurs.
9. Le terme de "séquence à séquence" ("*sequence-to-sequence architecture*") est également couramment employé.

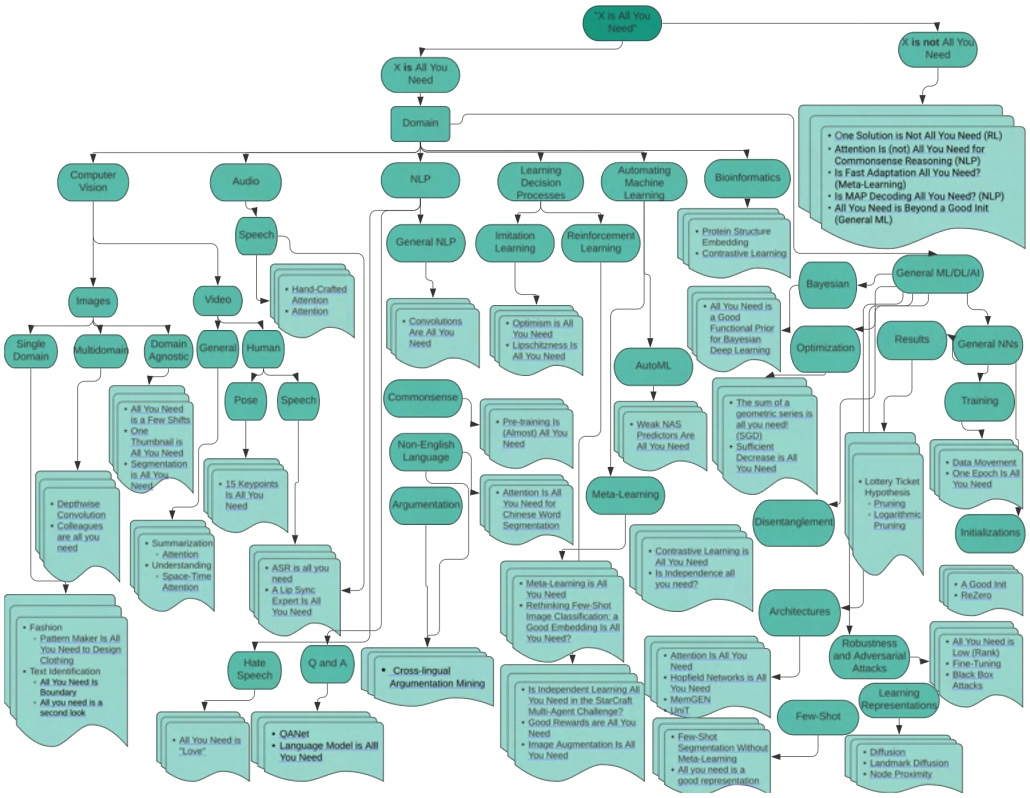


FIGURE 1.2 – "X is all you need", Vinay Uday Prabhu

Le reste de cette sous-section est consacré à la description — à un niveau moyen d'abstraction — de cette architecture, baptisée "Transformer" par ses inventeurs, et de ses modèles, consistant en un commentaire de la désormais célèbre figure 1.3, issue de l'article original.

Un modèle Transformer à l'entraînement est instancié avec pour paramètres des valeurs aléatoires, et son encodeur (la partie gauche de la figure 1.3) reçoit en entrée un segment de texte tokenisé, dont les tokens ont été remplacés par des indices numériques définis lors d'une phase de prétraitement des données d'apprentissage. L'implémentation originelle des modèles Transformer semble avoir reposé sur des plongements de mots ("*input embeddings*" sur la figure 1.3) pré-entraînés, soit sur une table d'indices associant chaque token à un vecteur de valeurs représentant certains aspects de ses attributs grammaticaux et sémantiques appris, préalablement, par analyse automatique des propriétés distributionnelles desdits tokens au sein d'un corpus d'apprentissage. Néanmoins, la majorité des implémentations actuelles de l'architecture Transformer (particulièrement celles basées sur Torch) semblent prendre appui sur des plongements de mots dynamiques, autrement dit sur une couche de neurones associant les indices des tokens d'entrée à un vecteur de paramètres initialisés aléatoirement et modifiés, au fil de l'entraînement du modèle, afin de capturer les attributs pertinents des tokens du point de vue de la TA comme tâche.

La force de l'architecture Transformer découle de la capacité impartie à son encodeur de traiter l'ensemble des tokens du segment d'entrée en parallèle, simultanément, plutôt qu'en séquence. Un effet collatéral d'une telle approche est, toutefois, que les informations relatives à la position des tokens dans la phrase d'entrée sont ainsi implicitement perdues et doivent donc être "réinjectées" explicitement à l'inté-

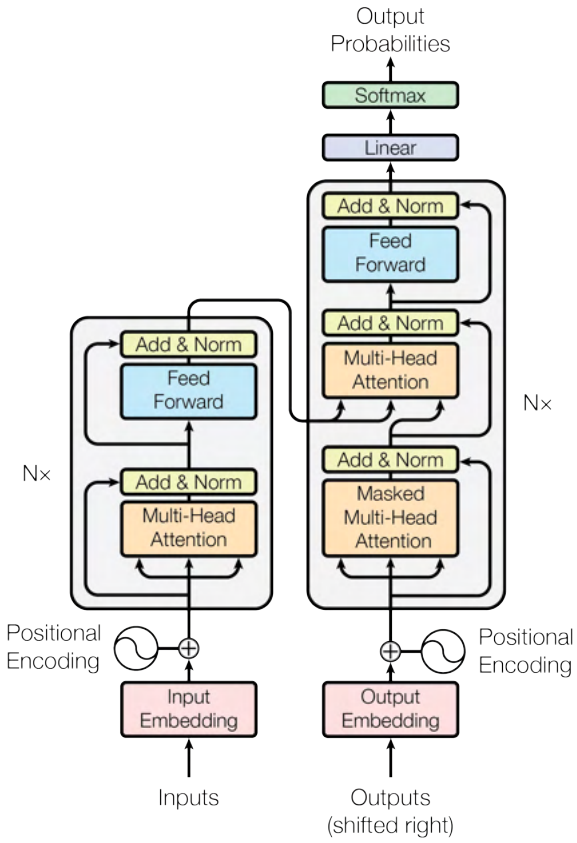


FIGURE 1.3 – Architecture Transformer

rieur des modèles. Cette fonction est assurée par une couche dite de "plongements positionnels" ("*positional embeddings*"), partie intégrante de cette architecture. Si l'implémentation originelle des modèles Transformer prenait appui sur des fonctions sinusoïdales afin de générer les valeurs statiques de ces plongements positionnels, la plupart de leurs implémentations actuelles semblent recourir à une couche de neurones additionnelle, associant les positions possibles des tokens du segment d'entrée à une matrice de vecteurs initialisée aléatoirement et dont les valeurs sont ajustées, au fil de l'entraînement, afin de permettre au modèle de capturer des informations relatives au rôle joué par chacune de ces positions au sein des segments de phrases. [10] Ces plongements positionnels sont additionnés, élément par élément, aux plongements de mots, et la structure de données résultante (désignée du nom d'"embedding" pour le restant de cette explication) est enfin proprement transmise à l'encodeur du modèle.

A un très haut niveau d'abstraction, les opérations effectuées par l'encodeur se résument à modifier la structure de données résultant de ces embeddings, de façon à injecter dans la représentation de chacun de leurs tokens des informations relatives à leur contexte.

A un niveau modéré d'abstraction, ces modifications sont effectuées par une série de sous-modules appelés blocs d'encodage ("*encoder blocks*", représentés par un encart gris sur la gauche de la figure 1.3), de la façon suivante : trois structures de données vectorielles sont instanciées (avec des valeurs aléatoires, au début de la phase

10. Néanmoins, la nature exacte de ce qu'apprennent, dans les faits, ces plongements positionnels dynamiques est l'objet de subtiles controverses, cf. [Wang and Chen, 2020].

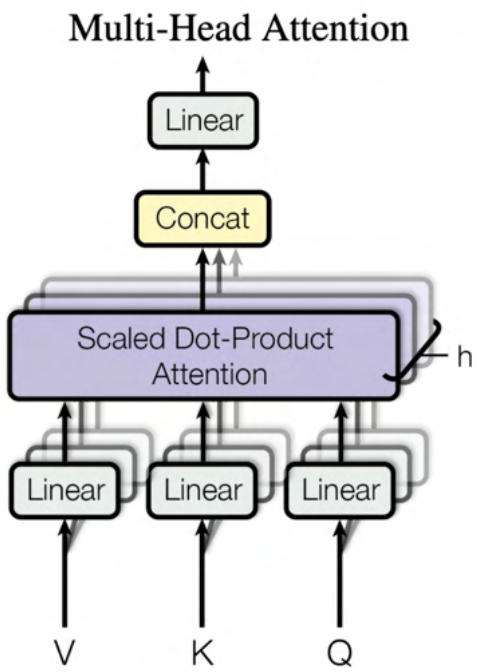


FIGURE 1.4 – "Multi-head Attention"

d'apprentissage du modèle) pour chacun des tokens du segment d'entrée. Le produit vectoriel ("*cross-multiplication*") de chacune de ces matrices et des embeddings des tokens d'entrée est subséquemment calculé, générant trois matrices, conceptualisées par les inventeurs de cette architecture comme une base de donnée associant une requête ("*query*"), à une clef ("*key*") et à une valeur ("*value*"). Les produits scalaires ("*dot products*") de la matrice de requête d'un token donné et des embeddings de chacun des tokens du segment d'entrée sont ensuite calculés (d'où la complexité quadratique de cet algorithme précédemment évoquée). Ces produits scalaires reflètent le degré d'affinité mutuelle des tokens (cf. figure 1.6), dans le contexte du segment d'entrée, et sont la base du mécanisme dit d'auto-attention ("*self attention*") au coeur de l'architecture Transformer. Ces scores d'attention sont ensuite rassemblés en un vecteur unique pour chaque token, et normalisés au moyen d'un facteur d'échelle ("*scaling factor*") et d'une fonction exponentielle normalisée ("*softmax*") afin de s'additionner à un. Le produit scalaire du vecteur résultant et de la matrice de valeur de chacun des tokens est ensuite calculé, générant un embedding modifié incorporant un degré dense d'informations relatives au contexte de chaque token. Cette séquence d'opération, baptisée "*scaled dot-product attention*" par les inventeurs de cette architecture, est schématiquement représentée figure 1.5.

Le sous-module réalisant cette complexe séquence d'opérations est appelé "tête d'attention" ("*attention head*"), dans l'architecture Transformer. Le module d'attention dite "multi-tête" ("*multi-head attention*", détaillé figure 1.4) consiste en 8 sous-modules parallèles (du moins dans l'implémentation originelle de cette architecture), effectuant ainsi simultanément ces mêmes opérations, au départ de paramètres différents — initialisés aléatoirement au début de la phase d'entraînement du modèle — permettant ainsi à chaque "tête d'attention" de développer une "vision" différente des tokens en contexte. Les données produites par les têtes d'attention sont ensuite concaténées, normalisées et ajoutées à une version non traitée des données d'entrée

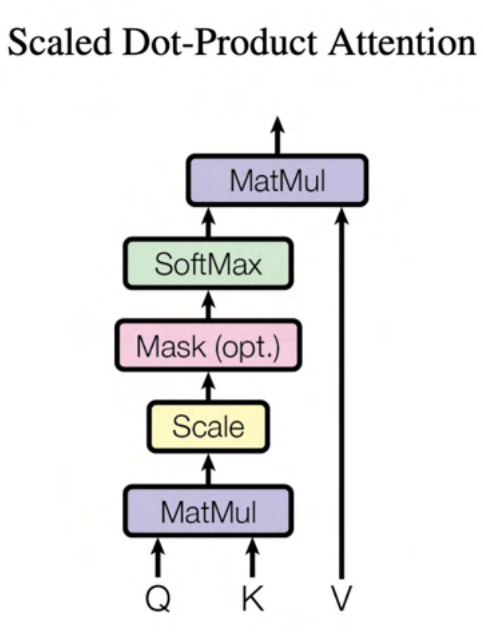


FIGURE 1.5 – "Scaled Dot-Product Attention"

grâce à un saut de connexion ("*skipped connection*"); la structure de données résultante est subséquemment transmise à un réseau de neurones "feed-forward" d'un type particulier [11]. L'ultime opération réalisée par les blocs d'encodage consiste à normaliser les données des couches feed-forward, tout en procédant à une addition par saut de connexion des données d'entrée.

L'encodeur présenté dans l'implémentation originelle de l'architecture Transformer se composait d'une séquence de six blocs d'encodage. Une fois passé le dernier bloc d'encodage, les données de sortie de l'encodeur sont ainsi finalement transmises au second module de l'architecture Transfomer — le décodeur — accompagnées d'un token spécial (généralement <sos> pour "*start of sentence*") lui servant de signal de départ.

La structure du décodeur du Transformer (représenté dans l'encart gris à la droite de la figure 1.3) est similaire à celle de l'encodeur. Il dispose d'une couche de plongements de mot pour les tokens de langue cible ainsi que d'une couche de plongements positionnels, mais possède deux modules d'attention au lieu d'un par bloc de décodage ("*decoder bloc*"). Le premier sous-module d'attention, appelé "*masked multi-head attention*", se comporte d'une manière similaire à celle des sous-modules équivalents de l'encodeur, à ceci près qu'au cours de l'entraînement du modèle ce sous-module réalise une opération de masquage/démasquage des tokens de la traduction de référence, servant deux fonctions : d'une part, empêcher le décodeur de "tricher" durant son entraînement et, d'autre part, assister son apprentissage de la traduction en lui soumettant, systématiquement, le token précédent attendu lors de sa génération du token cible courant, une approche appelée "*teacher forcing*" et qui n'est pas propre à cette architecture.

L'autre module d'attention, appelé "*encoder-decoder attention*", effectue une opération similaire à celle des sous-modules de l'encodeur, mais en allant chercher

11. Le rôle exact de ces couches "feed-forward", qui comprennent près des deux tiers des paramètres entraînables des modèles Transformer, est l'objet de controverses (cf. [Geva et al., 2021]).

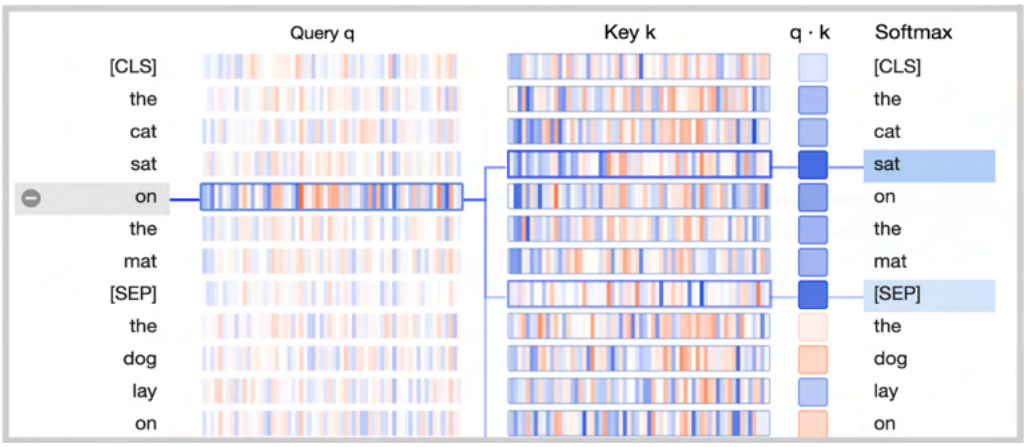


FIGURE 1.6 – Visualisation du mécanisme de l'auto-attention permise par la librairie BertViz

les matrices de clefs et de valeurs dans les données produites par l'encodeur, et les matrices de requête dans ses données propres, au sein des embeddings des précédents tokens cible. En bout de ces sous-modules, le décodeur génère une distribution de probabilités sur le vocabulaire des tokens cible ; le modèle en cours d'entraînement calcule alors une marge d'erreur, sur la base de l'entropie croisée, entre la distribution produite par le décodeur et celle attendue pour les tokens du segment cible de référence, et chacun des paramètres du modèle est alors modifié, en proportion de sa contribution à cette erreur, afin de la réduire. Le modèle entraîné retient, quant à lui, à chaque étape de la génération du décodeur, les tokens cible ayant les probabilités les plus élevées d'être les bons, construisant de multiples hypothèses de traduction finalement classées, sur la base de l'algorithme de recherche par faisceau ("*beam search*"), et soumises à l'utilisateur.

Les architectures de la TA neuronale sont ainsi éminemment complexes, et la direction que pourraient prendre leur évolution future est un mystère plus opaque encore que le détail de leur fonctionnement. Le caractère "auto-régressif" du décodeur Transformer, néanmoins, lequel génère donc les tokens cible un par un, en séquence, semblait mécontenter, au tournant des années 2020, certains chercheurs en quête d'une alternative à cette approche. Le développement de modèles Transformer dits non-autorégressifs ([Huang et al., 2022]) dont l'avantage résiderait dans leurs capacités d'inférence plus rapides, constitue peut-être l'horizon de cette architecture, dans un contexte où le TAL dans son ensemble paraît vouloir faire plus grand cas d'enjeux autres que les seules taux d'erreurs des modèles, enjeux tels, donc, que les temps d'inférence de ces modèles, mais aussi tels que le coût environnemental de leur entraînement ([Ethayarajh and Jurafsky, 2021]). La dernière section de ce chapitre portera précisément sur les modalités d'évaluation de la TA comme tâche ou document.

1.4 Évaluation

Le degré d'importance qu'accordent aux divers types de publications scientifiques les disciplines académiques tend à varier selon les champs d'étude ([Swales, 2004]). Les monographies, à titre d'exemple, règnent ainsi en maître sur les sciences hu-

maines, alors que les articles de revues scientifiques sont l'alpha et l'oméga des sciences sociales, et la TA comme domaine, au XXIe siècle, accorde une importance sans cesse croissante aux conférences et à leurs actes, au détriment d'autres formes de communication scientifique, comme en atteste la disparition récente, en 2021, de la revue *Machine Translation*, au terme de 35 années de publication trimestrielle.

Or les plus importantes de ces conférences (WMT, IWSLT), tendent à s'articuler autour de "tâches partagées" ("*shared tasks*") donnant lieu à des campagnes d'évaluation compétitives visant à permettre la comparaison systématique de solutions concurrentes à un problème donné de la TA comme tâche. Le caractère central de ces formes de "benchmarking" de la recherche scientifique dans le contexte de la TA contemporaine résulte, à n'en pas douter, de l'influence décisive exercée sur ce domaine par les pratiques de financement d'agences liées au complexe militaro-industriel étatsunien, depuis le début des années 1990, au premier rang desquelles la DARPA (Defence Advanced Research Projects Agency).

Cette agence, en effet, a joué un rôle de premier plan dans l'Histoire contemporaine de ce champ d'étude, au travers, notamment, du financement de multiples programmes de recherche, tels que la DARPA MT Initiative (1990-1995), TIDES (2000-2004), GALE (2005-2010) ou encore BOLT (2011-2016), relevant de la TA. Un programme tel que GALE (Global Autonomous Language Exploitation), à titre d'exemple, associait ainsi plusieurs acteurs privés concurrents sous contrat (IBM, SRI, Raytheon BBN), lesquels travaillaient indépendamment [12] au développement de systèmes d'extraction d'information, de traduction et de résumé automatiques de textes chinois et arabes, et se devaient, pour ne pas être exclues de ce programme et continuer à recevoir des financements de la DARPA, d'honorer des objectifs réguliers de performances ("*performance milestones*") définis quantitativement ([Anderson, 2006]).

Les pratiques d'évaluation représentent ainsi un aspect central de la TA comme domaine, voire même un sous-domaine à part entière voué à proposer, chaque année, de nouveaux protocoles d'évaluation, lesquels protocoles — comble de l'ironie — sont eux-mêmes l'objet de campagnes d'évaluation compétitives [13] !

L'état de l'art de cet aspect majeur de la recherche en TA est par conséquent complexe et foisonnant. Une description, nécessairement schématique, pourrait toutefois en être donnée sur la base des trois axes suivants : un axe manuel/automatique permettant de quantifier la mesure dans laquelle les protocoles d'évaluation mis en jeu reposent sur le jugement humain ou sur des algorithmes informatiques ; un axe intrinsèque/extrinsèque, distinguant, d'une part, les protocoles reposant sur l'évaluation directe de la TA comme document, et, d'autre part, les protocoles reposant sur l'évaluation du résultat de tâches annexes effectuées, en aval de la traduction, sur la base de documents traduits par les systèmes à évaluer, telles que des réponses à des questions de compréhension portant sur un texte résultant de la TA comme tâche ; et un troisième axe dit "boîte noire"/"boîte de verre" ("*blackbox*"/"*glassbox*"), mesurant le degré d'accès aux données internes des modèles de TA lors de l'évaluation. [14]

Le mode d'évaluation par défaut de la TA comme tâche était, jusqu'au tournant du troisième millénaire, manuel et intrinsèque : des annotateurs humains, bilingues ou — plus fréquemment — monolingues, étaient invités à juger de la qualité des

12. Quoiqu'en collaboration avec le Linguistic Data Consortium.
13. L'une des dernières en date ayant eu lieu en 2021, cf. https://aclanthology.org/2021.wmt-1.2/.
14. Les approches dites "boîte de verre" sont, par exemple, utiles pour l'estimation de la qualité des traductions offertes par les systèmes interactifs de traduction assistée par ordinateur (TAO).

traductions d'un système sur la base de critères spécifiques, telles la fluidité et l'adéquation, ou bien à classer plusieurs traductions d'un même segment source, produites par divers systèmes, de la meilleure à la moins bonne, en les comparant au segment source et/ou à un étalon-or. Une description remarquable d'un tel protocole figure dans la dixième annexe du célèbre rapport de l'ALPAC précédemment mentionné [15]. Ce mode d'évaluation, communément appelé "DA" (pour "Direct Assessment") dans la littérature contemporaine, présente l'inconvénient d'être très coûteux, tant en argent qu'en temps. Ce coût, ainsi que des controverses relatives à la fiabilité des annotateurs humains ([King, 1996]), incitèrent les chercheurs du domaine et leurs bailleurs de fonds à développer, au tournant des années 2000, de nouveaux protocoles d'évaluation entièrement automatiques. Le plus influent de ces algorithmes, appelé BLEU pour "BiLingual Evaluation Understudy" ([Papineni et al., 2002]), fut également l'un des plus précoces.

1.4.1 BLEU

Développée au sein des laboratoires d'IBM et financée, en partie, par la DARPA, cette métrique automatique s'appuie fondamentalement sur le calcul de la précision des n-grammes de tokens "plafonnés" ("*clipped*") – c'est-à-dire du ratio des n-grammes de tokens corrects dans l'hypothèse de traduction à évaluer au regard d'une ou de plusieurs traductions de référence, "plafonnés" au nombre maximum d'occurrences desdits n-grammes au sein de ces traductions de référence – sur la base du nombre total de n-grammes représentés dans la traduction à évaluer, une mesure appelée "précision n-gramme modifiée" ("*modified n-gram precision*") par ses concepteurs et représentée par le facteur p_n sur la formule 1.1 d'après [Papineni et al., 2002].

$$\text{Score BLEU} = \text{PB} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{1.1}$$

Les n-grammes pris en compte par BLEU vont traditionnellement de l'unigramme au quadrigramme ($N = 4$), et cette métrique prend appui sur la moyenne géométrique de ces "précisions modifiées" [16] pour chacun de ces n-grammes (cf. Formule 1.1) permettant ainsi la prise en compte d'information en partie liées à l'ordre de ces n-grammes.

$$\text{PB} = \begin{cases} 1 & \text{si } c > r \\ \exp\left(1 - \frac{r}{c}\right) & \text{si } c \leq r \end{cases} \tag{1.2}$$

Si la métrique BLEU repose, fondamentalement, sur la précision, une pénalité de brièveté (PB) exponentielle frappant le segment à évaluer tend à la rééquilibrer *a minima* vers le rappel (cf. formule 1.2, d'après [Papineni et al., 2002] ; c désignant le nombre de mots du segment candidat et r le nombre de mots du segment de référence).

15. https://nap.nationalacademies.org/read/9547/chapter/27

16. En effet, pour des raisons qui découlent du fait que les précisions modifiées des n-grammes décroissent de manière exponentielle à mesure que des n-grammes d'ordre supérieur sont considérés, la moyenne géométrique est calculée, au moyen de la moyenne logarithmique avec la méthode dite des poids uniformes ($w_n = \frac{1}{N}$), au lieu de la moyenne arithmétique dans cette formule. cf. https://math.stackexchange.com/questions/2028549/logarithmic-average-with-uniform-weights-geometric-mean.

Pour autant, les scores BLEU ne tiennent aucun compte des paraphrases correctes de l'étalon-or ou de phénomènes sémantiques relevant de la synonymie — un défaut substantiel pour l'évaluation des traductions — ni non plus de l'importance relative des n-grammes potentiellement omis ou ajoutés au sein des hypothèses de traduction (telle l'absence d'une négation), valant à cette métrique des critiques pléthoriques ([Koehn, 2020], p. 55). D'autres reproches adressés aux scores BLEU, lesquels consistent en des valeurs allant de 0 à 1 fréquemment ramenées à des pourcentages, ont également porté sur leur caractère jugé peu interprétable et non uniforme ainsi que sur la nécessité inhérente à leur utilisation de disposer d'importants volumes de données d'évaluations (en effet, du fait de la Formule 1.1, toutes les hypothèses de traduction ne partageant aucun 4-gramme, à titre d'exemple, avec la traduction de référence seront évaluées à zéro par BLEU), et nombre de métriques alternatives ont été proposées pour tenter de remédier à ces faiblesses (WER, PER, NIST, ChrF, Meteor, etc.). L'une des plus récentes de ces nouvelles approches de l'évaluation de la TA comme tâche et document est appelée COMET ([Rei et al., 2020]), et n'ambitionne rien moins que d'offrir à la TA neuronale l'évaluation — également neuronale — qu'elle mériterait selon ses concepteurs.

1.4.2 COMET

À un niveau d'abstraction élevé, l'algorithme d'évaluation de COMET repose, fondamentalement, sur l'encodeur d'un modèle Transformer entraîné à "deviner" un token masqué au sein d'un segment d'entrée. Plusieurs de ces modèles existent qui ont été développés et offerts au public des chercheurs et des ingénieurs par des entreprises du numérique telles qu'Alphabet/Google et Facebook/Meta. Celui de ces modèles sur lequel repose le système d'évaluation COMET est appelé XLM-R ([Conneau et al., 2020]) et fut entraîné par Facebook AI, en 2019, sur plus de deux téraoctets de données textuelles en cent langues issues du Web.

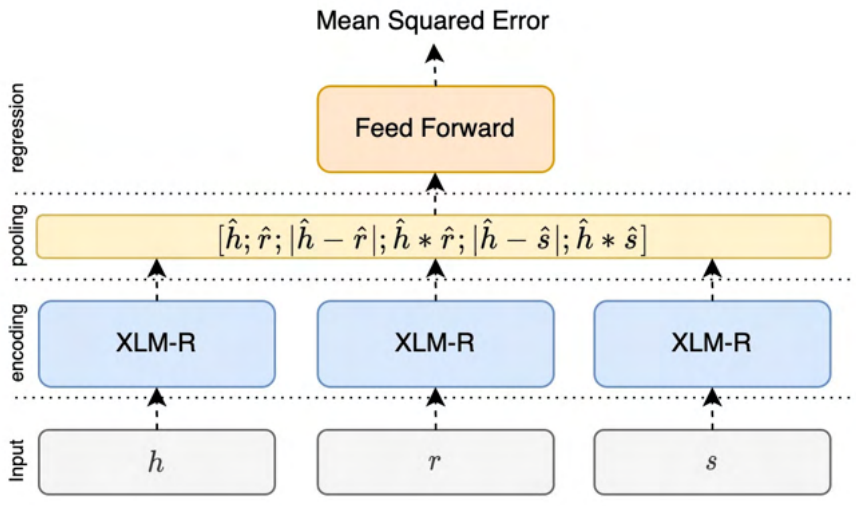


FIGURE 1.7 – Principale implémentation du modèle d'évaluation COMET

Dans le cadre du système COMET, ce modèle se trouve converti en un module

intermédiaire chargé de générer une représentation uniforme des embeddings des tokens des trois segments de phrase mis en jeu par l'opération d'évaluation de la TA comme tâche : un segment source ("s" sur la figure 1.7), une traduction de référence ("étalon-or") de ce segment source ("r" sur la figure 1.7) et une hypothèse de traduction produite par le système de TA candidat à l'évaluation ("h" sur la figure 1.7). Quatre opérations sont ensuite réalisées sur ces embeddings générés par XLM-R : le produit vectoriel de l'embedding du segment source et de la traduction candidate est calculé, de même que le produit vectoriel des embeddings de l'étalon-or et de la traduction candidate ; et la distance de Manhattan entre les embeddings du segment source et ceux de l'hypothèse candidate est calculée. Les quatre vecteurs résultant de ces opérations sont alors concaténés à ceux de l'hypothèse candidate et de l'étalon-or – les auteurs justifiant leur choix de ne pas ajouter à cette structure de données l'embedding du segment source par des considérations liés à la recherche de gains de performance – et la structure de données ainsi obtenue est transmise à des couches de neurones feed-forward entraînées à générer, par régression, les scores décernés à des hypothèses de traduction par des évaluateurs humains dans le cadre des campagnes d'évaluation passées de la conférence WMT. [17]

L'évaluation de la TA comme tâche est sans doute l'aspect de l'état de l'art de la TA comme domaine dont la trajectoire avenir est la plus évidente. Si des voix s'élèvent, ainsi que mentionné précédemmment, dans le contexte général du TAL et de l'IA ("intelligence artificielle"), pour dénoncer l' "obsession du classement" ("*leaderboard mentality*") qui prévaut dans ces domaines ([Ethayarajh and Jurafsky, 2021, Church and Kordoni, 2022]), l'évaluation est un aspect par trop essentiel de la TA comme domaine et des mécanismes de son financement pour disparaître à brève échéance. Ainsi la métrique BLEU, objet de perpétuelles critiques (cf. [Mathur et al., 2020], à titre d'exemple), fêtait en 2022 ses 20 ans — un âge canonique s'il en est pour un artefact du TAL ! — et la recherche portant sur l'amélioration des métriques d'évaluation neuronales, du type de COMET, ne montrait, cette même année, aucun signe d'essoufflement ([Rei et al., 2022]) et se trouvait même en passe d'être étendue à de nouvelles approches indirectes/extrinsèques et automatiques de l'évaluation de la TA comme tâche ([Krubiński et al., 2021]).

On retiendra, cependant, que si ces métriques neuronales sont prometteuses, leur recours à des modèles de langues nécessitant d'être entraînés sur d'importants volumes de données textuelles est susceptible de limiter leur application aux langues dites "peu dotées" — dont il sera question au chapitre suivant de ce mémoire — langues peu dotées qui font pourtant l'objet d'un intérêt grandissant de la recherche en TAL, garantissant peut-être encore quelques longues années de vie au score BLEU !

17. Ces scores consistent en une note allant de 0 à 100 assignée par les évaluateurs humains aux traduction candidates sur la base de critères de fluidité et de fidélité, principalement, convertis en une cote Z ("*z-score*") pour chaque annotateur, même si les modalités exactes de mise en œuvre de ces campagnes de "DA" ont eu tendance à changer d'une année sur l'autre (cf. [Knowles, 2021]). On notera par ailleurs que d'autres variantes de COMET existent qui excluent l'étalon-or, transformant la stricte évaluation de la TA comme document en une estimation de sa qualité ("*quality estimation*"), ou bien incluent l'étalon-or mais génèrent un classement par paire des traductions candidates au lieu d'un score, cf. https://unbabel.github.io/COMET/html/models.html.

CHAPITRE 2

LES LANGUES PEU DOTÉES, L'AMHARIQUE ET LA LEXICOGRAPHIE BILINGUE

Sommaire

2.1 Définitions (conventionnelle et opérationnelle) **35**
2.2 Amharique **37**
2.2.1 Données démographiques et sociolinguistiques 38
2.2.2 Contacts et traits pertinents en typologie linguistique 39
2.2.3 Système d'écriture 40
2.2.4 Données parallèles 44
2.3 Lexicographie bilingue **48**
2.3.1 Nomenclature des dictionnaires bilingues 49
2.3.2 La lexicographie bilingue et la TA comme tâche 50

2.1 Définitions (conventionnelle et opérationnelle)

Le terme de langue a pour signifié un concept vague s'il en est, à telle enseigne que d'éminents linguistes et philosophes, tels Noam Chomsky ([Chomsky, 1986]) ou Donald Davidson ([Davidson, 1986]), ont entrepris d'en réfuter la validité même. En effet, les critères objectifs sur la base desquels des discours prononcés, écrits, signés ou pensés pourraient se voir assignés à des catégories abstraites appelées "français", "anglais", "croate", "serbe" ou "amharique" souffrent d'un défaut notoire de validité universelle. Ainsi, à titre d'exemple, le critère de l'intelligibilité mutuelle — l'un des critères objectifs les plus couramment invoqués pour la définition des langues — n'est pas applicable à des environnements linguistiques majeurs, tel celui de la Scandinavie où des idiolectes mutuellement intelligibles sont néanmoins regroupés en des langues distinctes, tels le danois, le norvégien ou le suédois ([Chambers and Trudgill, 1998], p. 3), alors même, à l'inverse, qu'une langue telle que le kurde désigne des idiolectes mutuellement inintelligibles ([Kreyenbroek and Sperl, 1992], p. 35). Les définitions intensionnelles de la notion de "langue" fondées sur le critère d'intelligibilité mutuelle échouent ainsi à rendre compte de l'extension de ce terme dans ses usages, même savants.

Dans la mesure, néanmoins, où des aspects cruciaux des systèmes modernes d'archivage et d'échange d'information reposent, en dernière analyse, sur la notion de langue — si vague soit-elle — les sociétés humaines contemporaines se sont vues

contraintes de dépasser cette aporie et d'établir, à défaut d'une définition intensionnelle fiable et universelle de la notion de langue, du moins une défintion extensionnelle ferme de ce terme. En ce début de XXIe siècle, il n'existe ainsi pas moins de trois normes internationales visant à l'accomplissement d'un tel objectif.

La première de ces normes, ISO 639-1, confiée à la charge du Centre international d'information sur la terminologie (Infoterm), basé à Vienne, consistait, en 2022, en une liste de 183 langues, à chacune desquelles étant attribué un code alphabétique à deux caractères (tels "en" pour l'anglais ou "fr" pour le français), cette première liste étant elle-même le sous-ensemble d'une deuxième norme, appelée ISO 639-2, laquelle, maintenue par la Bibliothèque du Congrès des États-Unis, recensait, en 2022, près de 487 langues et familles de langues se voyant assigner des codes alphabétiques à trois caractères (tels "afa" pour langues afro-asiatiques, "eng" pour l'anglais, "fra" ou "fre" pour le français [1]). La dernière de ces normes, ISO 639-3, confiée à l'organisation protestante SIL International, répertoriait enfin, en 2022, près de 7 893 langues, également désignées par un code alphabétique à trois caractères.

Ces définitions extensionnelles de la notion de langue ne sont toutefois pas gravées dans le marbre et ces listes sont susceptibles d'être étendues à de nouveaux idiomes, selon des modalités propres à chaque standard. [2]

Les critères d'inclusion de la norme ISO 639-1 sont ainsi les plus rigoureux, exigeant qu'il existe un corpus étendu de documents (textes spécialisés, tels des manuels scolaires ou universitaires, des ouvrages techniques, des revues d'experts, etc.) rédigés dans des registres savants de la langue candidate, ainsi qu'un certain nombre d'ouvrages terminologiques (dictionnaires techniques, glossaires spécialisés, etc.) pour divers domaines. Une autorité compétente (telle qu'un ministère, une académie ou une institution culturelle) doit en outre exprimer son soutien explicite à l'inclusion dans la norme de la langue candidate, laquelle doit par ailleurs répondre à divers critères d'ordres démographique et sociolinguistique (nombre de locuteurs, statut officiel dans un ou plusieurs pays, etc.).

La norme ISO 639-2 repose, quant à elle, sur une définition opérationnelle moins contraignante de la notion de langue, requérant que 50 documents, seulement, rédigés dans la langue candidate, soient versés au dossier des idiomes candidats, accompagnés d'une présentation de la littérature de ces langues, orale ou écrite, de la preuve du soutien d'une autorité politique régionale ou nationale à leur candidature, voire éventuellement d'un statut officiel dont elles jouiraient au sein d'États, ou encore de leur rôle dans l'enseignement scolaire ; les langues candidates échouant à remplir ces conditions sont néanmoins susceptibles d'être désignées par les codes collectifs assignés aux familles de langues. [3]

Enfin, toutes les "langues utilisées par un groupe d'individus pour servir à la communication humaine depuis une période de temps substantielle" [4] sont susceptibles de rejoindre la liste des langues reconnues par la norme ISO 639-3, pour peu que "les motifs avancés [par ailleurs soient] jugés suffisants » ! [5]

Quoi qu'il en soit, les définitions extensionnelles de la notion de langue que constituent ces listes serviront de définition opérationnelle à ce terme pour le restant de

1. Certaines langues sont, en effet, susceptibles de disposer de plusieurs code ISO 639-2 alternatifs, conséquence de la non-uniformité des pratiques anciennes de gestion d'archives.
2. Cf. https://www.loc.gov/standards/iso639-2/faq.html13
3. Ibid.
4. https://iso639-3.sil.org/about/faq
5. https://iso639-3.sil.org/code$_changes/change_managementGuidelines$

cette étude, les langues étant dès lors envisagées comme des catégories d'idiolectes admises par convention.

Une définition de la notion de "langue peu dotée" dans le contexte de la TA comme domaine est dès lors aisée, n'étant fonction que de la nature des données requises pour le développement des systèmes de l'état de l'art de la TA comme tâche.

Ces données consistent principalement, depuis le tournant des années 1990 et le passage des approches "par règles" aux approches statistiques du domaine, en des corpus dits "parallèles", sommes de bitextes en deux langues — soit de textes de départ et de leur traduction — ordinairement alignés phrase à phrase, les "bonnes pratiques" de la TA comme domaine tendant à imposer d'organiser de tels corpus sous la forme de deux fichiers texte ayant pour extension le code ISO 639-1 de la langue de leurs segments, lesquels segments se correspondant ligne à ligne au sein de chacun des fichiers [6] .

Une langue peu dotée dans le contexte de la TA comme domaine est dès lors une langue faiblement représentée au sein des corpus parallèles mis à la disposition des chercheurs. La quantification, en des termes précis, de cette notion de "faible représentation" — basée, par exemple, sur l'établissement d'un seuil maximal de segments parallèles ou de mots — est une problématique que la littérature dédiée à la TA des langues peu dotées tend à négliger.

Si certains des premiers travaux entrepris sur ce thème semblent avoir fixé la limite d'inclusion d'une langue dans la catégorie des langues peu dotées à un million de paires de segments parallèles (cf. [Ranathunga et al., 2021]), des études plus récentes semblent avoir divisé ce chiffre par deux, le fixant à 0,5 million de paires de segments ([Ibid.]). La limite de 0,1 million de paires de segments ayant été évoquée, par certains chercheurs, comme seuil d'inclusion dans la catégorie nouvelle des langues dites "très peu dotées" ([Ibid.]). [7]

Le reste de ce chapitre est consacré à la présentation de l'état de l'art de la traduction automatique d'une langue peu dotée — l'amharique — et introduit l'enjeu principal de l'expérience conduite dans la seconde partie de ce mémoire : le rôle possible de documents issus de la lexicographie bilingue en traduction automatique des langues peu dotées.

2.2 Amharique

L'amharique (ISO 639-1 "am", ISO 639-2 "amh") est une langue sémitique (ISO 639-2 "sem") d'Éthiopie, appartenant à ce titre à la macro-famille des langues afro-

6. On notera qu'une tendance récente de la recherche du domaine, et qui vise à contrôler l'impact formel du langage des traducteurs humains ("*translationese*") sur les modèles de traduction et leur évaluation, consiste à inclure dans l'organisation même des corpus parallèles des indications portant sur la direction de traduction et le caractère original ou traduit des segments compilés ; [Ni et al., 2022] présente un tel corpus.

7. On notera que si des données d'autres types existent sur lesquelles peut s'appuyer la TA comme tâche, elles ne semblent pas entrer en ligne de compte dans les définitions de la notion de langue peu dotée évoquées dans la littérature ; il s'agit principalement de corpus dits "comparables", formés par l'amalgamation de textes électroniques monolingues issus de domaines jugés similaires, des données monolingues étant elles-mêmes susceptibles – par ailleurs – d'être converties en données parallèles à la faveur d'une opération appelée "rétrotraduction" ("*backtranslation*") ([Sennrich et al., 2016a]) consistant à traduire automatiquement, et de manière potentiellement itérative, un corpus monolingue de langue cible vers la langue source de la TA comme tâche (l'opération inverse, moins évidemment motivée, consistant à traduire un corpus monolingue de langue source vers la langue cible pour générer un corpus parallèle est parfois appelée "forward translation" ([Zhuo et al., 2022])).

asiatiques (ISO 639-2 "afa").

2.2.1 Données démographiques et sociolinguistiques

Sur les 74 millions d'individus ayant participé au recensement de la population éthiopienne en 2007 [8], 29 millions ont fait le choix de caractériser leur idiolecte principal comme relevant de l'amharique [9] qui est en conséquence la langue sémitique "la plus parlée" au monde, après l'arabe (ISO 639-1 "ar", ISO 639-2 "ara").

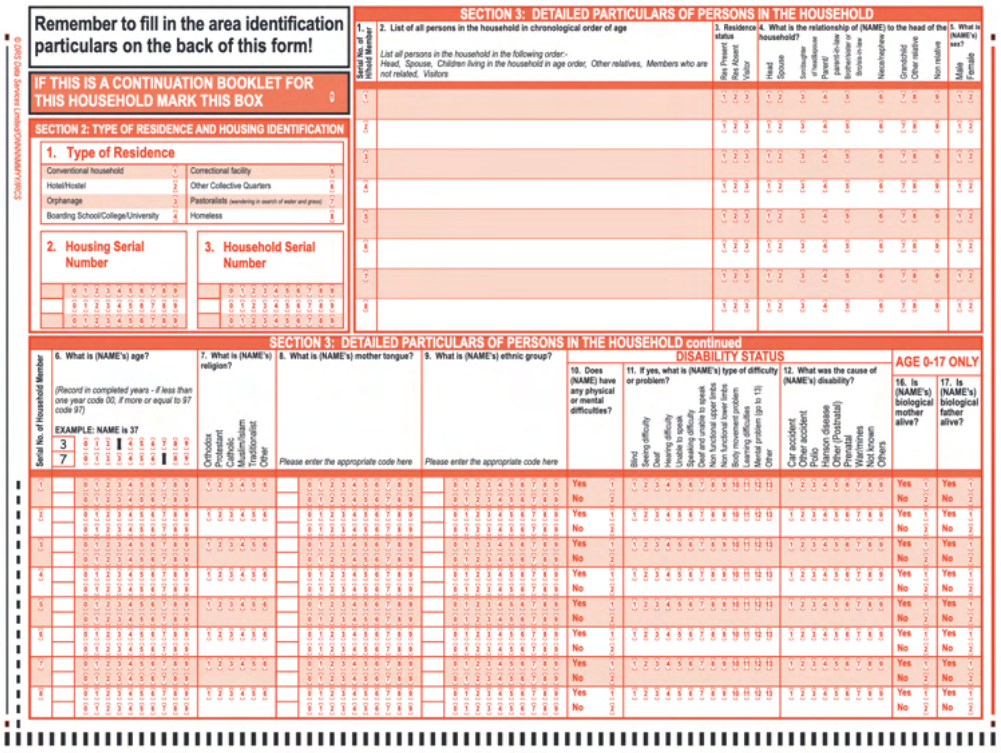


FIGURE 2.1 – Version anglaise d'un formulaire de recensement éthiopien ; le huitième champ concerne la langue des individus recensés

Si l'amharique ne jouit plus, depuis 1991, du statut de langue officielle de l'Éthiopie, son statut de langue de travail du gouvernement fédéral éthiopien est cependant garanti par la constitution de cet État [10], l'amharique demeurant par ailleurs la langue unique d'enseignement primaire et secondaire en Éthiopie (cf. D. Appleyard in [Brown and Ogilvie, 2008]), expliquant le nombre élevé de ses locuteurs de langue seconde déclarés (25 millions d'après Ethnologue/Wikipedia [11]). Ce rôle joué par l'amharique au sein du système éducatif éthiopien explique également, sans

8. https://www.statsethiopia.gov.et/census-2007-2/

9. En août 2022, la page Wikipedia de langue anglaise dédiée à l'amharique rapportait les données statistiques suivantes, avancées par Ethnologue pour l'année 2018 : 32 millions de locuteurs natifs de l'amharique et 25 millions de locuteurs ayant l'amharique pour langue seconde ; les données d'Ethnologue n'étant toutefois pas disponibles en libre accès, l'exactitude du report de ces chiffres par Wikipedia n'a pu être vérifiée.

10. Le statut méconnu, en aménagement linguistique, de langue de travail jouit dans les faits d'un grand prestige, supérieur, à certains égards, à celui des langues nationales et officielles (cf. [de Varennes, 2012]).

11. https://en.wikipedia.org/wiki/Amharic

doute, la remarquable uniformité diatopique que présente cette langue (Appleyard in [Brown and Ogilvie, 2008]). [12].

2.2.2 Contacts et traits pertinents en typologie linguistique

L'amharique — en tant qu'idiome en contact — a subi l'influence des langues couchitiques (ISO 639-2 "cus") dominantes de la Corne de l'Afrique, qu'il s'agisse de l'oromo (ISO 639-1 "om", ISO 639-2 "orm") à l'époque moderne, ou des langues agaw [13] dans l'Antiquité tardive et tout au long du Moyen Âge ([Hetzron, 1976]).

Ces influences couchitiques, substratales et adstratales, vis à vis de l'amharique, ainsi que ses racines sémitiques — jeu de mots volontaire — expliquent la situation originale de l'amharique au plan de la typologie linguistique, situation dont l'édition électronique du World Atlas of Linguistic Structures (WALS) offre une remarquable description sourcée synthétique, consultable en ligne [14].

Les traits typologiques les plus pertinents du point de vue de la TA comme tâche sont d'ordinaire considérés comme étant : l'ordre des mots [15], la complexité morphologique, la densité référentielle et la divergence lexicale ([Jurafsky and Martin, 2022]).

L'ordre canonique du sujet, de l'objet et du verbe amhariques est l'ordre SOV, l'ordre du génitif et du nom est l'ordre génitif-nom, et l'ordre de la proposition subordonnée relative par rapport à sa tête ou son "antécédent" est l'ordre relative-antécédent, faisant de l'amharique une langue foncièrement centripète (ou "head final"), à l'exemple du japonais (ISO 639-1 "ja", ISO 639-2 "jpn") [16].

La morphologie de l'amharique est ordinairement qualifiée de "complexe" ([Fabri et al., 2014], p. 8) — notamment dans le cas du verbe (cf. D. Appleyard in [Brown and Ogilvie, 2008], p. 34) — mettant en jeu des procédés sophistiqués de préfixation, de suffixation et d'infixation caractéristiques de la morphologie non-concaténative des langues sémitiques, et marquant :

— deux genres (masculin et féminin),
— deux nombres (singulier et pluriel, la pluralité indéfinie étant souvent exprimée par le singulier dans cette langue),
— deux cas (un cas non marqué et un cas objet réservé aux substantifs définis),
— le caractère défini ou indéfini des noms et des syntagmes nominaux,
— dix personnes pour les verbes et les pronoms personnels,
— quatre aspects verbaux (perfectif, imperfectif, impératif/jussif et gérondif),
— et cinq voix verbales (active, passive, réciproque, causative et factitive).

La densité référentielle de l'amharique n'a, semble-t-il, quant à elle, jamais fait l'objet d'études approfondies. En raison, néanmoins, de la complexité morphologique de cette langue, cette densité est susceptible d'être élevée. Toutes les formes verbales de l'amharique présentent, en effet, un indice de personne — distinguant, par surcroît, le masculin du féminin à la deuxième et à la troisième personne du singulier —

12. On notera cependant que des variations — négligeables du point de vue de la TA comme tâche — sont réputées exister dans la phonologie de cette langue, l'accent d'Addis-Abeba, la capitale, servant de norme à l'acrolecte (cf. D. Appleyard in [Brown and Ogilvie, 2008]).

13. L'awngi (ISO 639-3 "awn") et et le bilen (ISO 639-2 "byn") comptent parmi les derniers représentants contemporains de cette famille de langues menacées.

14. https://wals.info/languoid/lect/wals$_{c}ode_amh$

15. Il semble douteux, néanmoins, que l'ordre des mots soit toujours — à l'ère des modèles Transformer décrits précédemment — un problème d'envergure en TA neuronale...

16. Le caractère centripète de l'amharique est généralement attribué à l'influence de son substrat/adstrat couchitique.

et le complément d'objet du verbe, dès lors qu'il est défini, est fréquemment repris, de manière redondante, par un pronom anaphorique affixé au verbe, comme l'illustrent les deux exemples (a) et (b) présentés ci-dessous.

(a) lämma and t'rmus säbbär-ä
Lemma un(e) bouteille casser.PRF-3M
"Lemma a cassé une bouteille."

(b) lämma t'rmus-u-n säbbär-ä-w
Lemma bouteille-DEF-ACC casser.PRF-3M-PRON.ACC.3M
"Lemma a cassé la bouteille."

(d'après [Amberber, 2008])

Le degré de divergence lexicale présenté par l'amharique paraît également n'avoir jamais été quantifié vis à vis d'une quelconque autre langue. Relativement à l'anglais (ISO 639-1 "en") ou au français (ISO 639-1 "fr"), cette divergence est également susceptible d'être élevée, même si ces langues entretiennent avec l'amharique des liens indirects substantiels au plan du lexique. En effet, le guèze (ISO 639-2 "gez") — la langue classique de l'Éthiopie et qui demeure la langue liturgique de son Église orthodoxe — a subi l'influence marquée, au travers de son lexique, du grec hellénistique (ISO 639-2 "grc") par le truchement du Nouveau Testament, expliquant, le nombre substantiel de concepts et d'expressions empruntés ou calqués sur le texte biblique que l'amharique, l'anglais et le français ont en partage, telle l'expression "rendre à César ce qui est à César" des Évangiles synoptiques attestée dans ces trois langues ([Leslau, 1973]).

2.2.3 Système d'écriture

Aux enjeux découlant de la typologie linguistique de l'amharique vis-à-vis de la TA comme tâche, il convient d'ajouter ceux liés au système d'écriture de cette langue.

Ce système d'écriture, communément appelé "alphasyllabaire éthiopien", puise ses origines dans les scripts proto-alphabétiques attestés, dès le milieu du XIXe siècle avant notre ère, dans la péninsule du Sinaï, et qui devaient donner naissance, au cours de l'Âge du fer, à deux traditions scribales distinctes : celle, d'une part, de l'alphabet phénicien — lequel était voué à devenir l'un des artefacts majeurs de l'Histoire humaine [17] — et celle, d'autre part, de l'alphabet sudarabique, qui devait servir à l'écriture des langues sémitiques de la partie méridionale de la péninsule arabique et de la corne de l'Afrique dès le deuxième millénaire avant notre ère.

Le script de cette seconde tradition consistait en un abjad, ou alphabet consonantique, mais fut converti en un abugida (ou alphasyllabaire), au IVe siècle de notre ère, par des scribes du guèze ([Fischer, 2003]), donnant naissance au syllabaire éthiopien actuel, écrit de gauche à droite, dernier représentant encore en usage de la seconde

17. Adopté et adapté par diverses sociétés du monde antique, ce système d'écriture devait, en effet, servir de matrice aux alphabets grec, étrusque, romain, araméen et — par cet intermédiaire, en conséquence de l'adoption par l'empire achéménide de l'araméen comme "langue d'administration" — de base au développement des alphabets syriaque, arabe, hébreu, ainsi qu'à l'ensemble des scripts de l'Inde ([Fischer, 2003]).

FIGURE 2.2 – Stèle votive sabéenne présentant une inscription en écriture sudarabique, Musée du Louvre

branche de ces systèmes d'écriture sinaïtiques, du fait de l'extension contemporaine de l'alphabet arabe à l'ensemble de la péninsule arabique.

Représentant près de 26 consonnes et 7 voyelles [18] combinées en syllabes de structure CV (Consonne-Voyelle), l'alphasyllabaire éthiopien est communément présenté sous la forme d'un tableau (cf. figure 2.3) dont les rangs marquent les consonnes et les colonnes les voyelles, et dont les cellules contiennent ainsi les glyphes représentant la graphie des syllabes formées par la consonne du rang et la voyelle de la colonne.

Nom			ä	u	i	a	é	e	o	wa	yä
ሆይ	Hoy	h	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ		
ላዊ	Lawi	l	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ	ሏ	
ሐውት	Hawt	ḥ	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ	ሗ	
ማይ	May	m	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ	ሟ	ሿ
ሠውት	Säwt	ś	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ	ሧ	
ርእስ	Re'es	r	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ	ሯ	ፘ
ሳት	Sat	s	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ	ሷ	
ቃፍ	Qaf	q	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ	ቋ	
ቤት	Bét	b	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ	ቧ	
ታው	Taw	t	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ	ቷ	
ኀርም	Harm	ḫ	ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ	ኋ	
ነሐስ	Nähas	n	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ	ኗ	
አልፍ	Alef	ʾ	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ	ኧ	
ካፍ	Kaf	k	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ	ኳ	
ወዌ	Wäwé	w	ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ		
ዐይን	Ayn	ʿ	ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ		
ዘይ	Zäy	z	ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ	ዟ	
የመን	Yämän	y	የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ		
ድንት	Dent	d	ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ	ዷ	
ገምል	Gämel	g	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ	ጓ	
ጠይት	Täyt	ṭ	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ	ጧ	
ጰይት	Päyt	p	ጰ	ጱ	ጲ	ጳ	ጴ	ጵ	ጶ	ጷ	
ጸደይ	Sädäy	ṣ	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ	ጿ	
ፀጳ	Säpa	ṗ	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ		
አፍ	Af	f	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ	ፏ	ፗ
ፕሳ	Psa	p	ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ	ፗ	

FIGURE 2.3 – Alphasyllabaire amharique, Wikipedia

L'écriture éthiopienne possède également un système élaboré de ponctuations qui tend aujourd'hui à se simplifier mais comprenait, traditionnellement, quelques dix glyphes distinctes (un séparateur de mots "፡", une virgule "፣", un point-virgule "፤", un point final " ። ", un séparateur de paragraphe " ፧ ", des marqueurs de section, un point d'interrogation "፧", et plusieurs variantes des deux-points servant des fonctions distinctes sans équivalents en français : " ፥ " et " ፦ "), ainsi qu'un système de chiffres alphabétiques adapté du grec (cf. figure 2.4).

Les points de code attribués au script éthiopien par le consortium Unicode en 1999 ont des valeurs situées entre U+1200 et U+137F (en hexadécimal ; 4608 et 4991 en

18. Des caractères additionnels existent, en effet, pour noter les phonèmes absents du guèze dans les langues modernes de la Corne de l'Afrique.

FIGURE 2.4 – Systèmes de chiffres alphabétiques grec, copte et éthiopien ; Wikipedia

décimal) au sein du plan multilingue de base. [19]

On notera qu'un aspect des choix d'implémentation de ces encodages par le consortium Unicode est lourd de conséquences du point de vue du TAL, en général, et de la TA comme tâche, en particulier. Les glyphes de l'écriture éthiopienne sont, en effet, considérées comme non-compositionnelles par cet encodage, empêchant du même coup la décomposition des syllabes représentées par ces caractères selon la norme d'équivalence NFKD (cf. figure 2.5), laquelle permettrait d'en détacher la consonne de la rime, un atout majeur dans la perspective des approches BPE de la tokenisation précédemment évoquées.

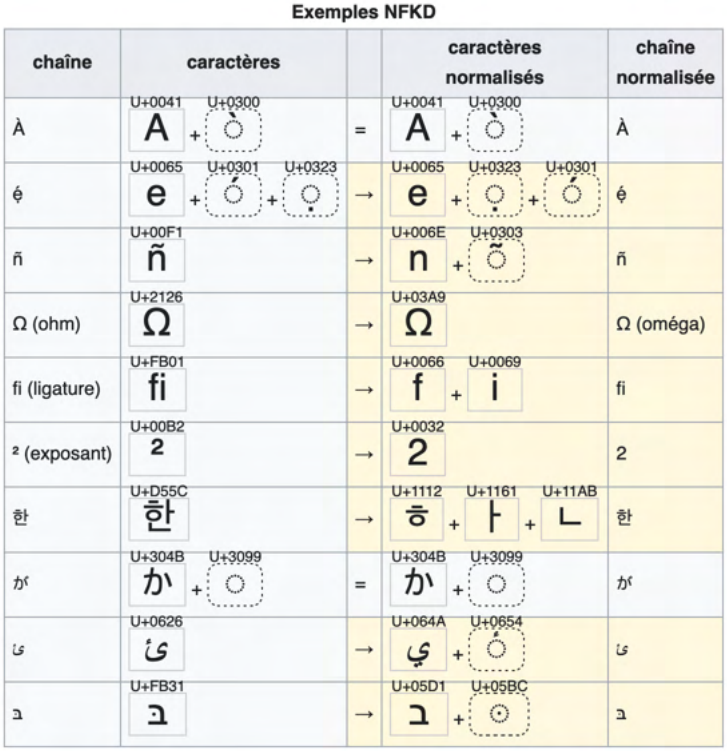


FIGURE 2.5 – Exemples de décompositions selon la norme NFKD des caractères composés du standard Unicode, Wikipedia

Le mot graphique *tawqaläčäw* "elle le sait" aurait ainsi pour segmentation "idéale", du point de vue de la morphologie, t-awq-al-äč-äw (glose : PRON.NOM.3F-

19. Cf. https://codepoints.net/basic$_{m}ultilingual_plane$, https://codepoints.net/ethiopic. On notera que depuis la version 4.1 du standard Unicode, le script éthiopien s'est par ailleurs vu attribuer plusieurs nouveaux blocs de points de code pour l'encodage de glyphes additionnelles, servant à l'écriture d'autres langues éthiopiennes, cf. https://codepoints.net/ethiopic$_{e}xtended$.

savoir-IPFV-PRON.NOM.3F-PRON.ACC.3M), une segmentation que la nature strictement non-compositionnelle de l'encodage Unicode des caractères éthiopiens rend inaccessible, dans la mesure où elle nécessiterait de couper au cœur des glyphes (cf. table 2.1). Une solution largement adoptée (notamment par Google Translate) pour résoudre cette difficulté consiste à procéder à la translittération systématique, en écriture latine, des textes amhariques à traiter.

ታ	ው	ቃ	ለ	ቸ	ው
ta	w'	qa	lä	čä	w'
t - aw'q- al - äč - äw'					

TABLE 2.1 – Décomposition du mot graphique *tawqaläcäw* en caractères et en morphèmes

Une translittération qui tend à résoudre – d'une pierre deux coups – un autre problème découlant pour le TAL de l'écriture de l'amharique : nombre d'oppositions phonologiques existant en guèze se trouvent neutralisées, en amharique, et les caractères des rangs ሀ, ሐ et ኀ représentent ainsi, à titre d'exemple, dans l'écriture de cette langue, un unique phonème /h/ (au plan phonétique la fricative glottale sourde [h]) suivi d'une voyelle. La figure 2.6 offre un tableau représentant l'ensemble de ces caractères homophones redondants de l'amharique contemporain. L'orthographe correcte de cette langue requiert une certaine érudition — la définition des bons usages se fondant généralement sur l'étymologie et faisant fréquemment l'objet de controverses — entraînant une instabilité sensible dans la graphie des mots de cette langue néfaste aux applications du TAL de manière général, et de la TA comme tâche en particulier. En effet — et à titre d'exemple — un modèle Transformer de TA ayant l'amharique pour langue source ou cible se verrait contraint d'apprendre les représentations distinctes d'un nombre artificiellement accru de synonymes résultant de ces variations orthographiques, et requerrait dès lors un surcroît de données d'apprentissage pour en saisir les contextes ; et — de façon similaire — dans le cas de l'évaluation par BLEU des performances de tels modèles de TA ayant l'amharique pour langue cible, ces instabilités orthographiques seraient susceptibles de pénaliser injustement les systèmes à évaluer en empêchant l'identification de n-grammes corrects mais dont l'orthographe différerait, de façon triviale, de celle de l'étalon-or.

La translittération systématiques des données textuelles amhariques au moyen de caractères adaptés de l'alphabet latin [20] résout néanmoins ces difficultés.

2.2.4 Données parallèles

La pénurie de données parallèles qui touche l'amharique constitue, toutefois, un obstacle plus grand encore au développement de systèmes de TA neuronale que ne le sont son script ou ses propriétés typologiques. Un inventaire exhaustif des ressources existantes, basé principalement sur la recension de [Gezmu et al., 2022], ne comprend ainsi, en tout et pour tout, que les cinq corpus parallèles suivants :

— un corpus bilingue amharique-anglais, compilé vers 2005 par la European Language Resource Association (ELRA) et d'accès payant, de 232 653 mots amhariques (aucune information n'étant offerte quant au nombre des segments)

20. Par exemple : https://github.com/andmek/AT4MT/blob/main/transliteration.py

	*	ə	u	i	a	e	o	ɨ	ʷi	ʷa	ʷe	ʷ
*			ኡ ዑ	ኢ ዒ	አ ኣ ዐ ዓ	ኤ ዔ	አ ዖ	እ ዕ				
g							ጎ ጐ					
h	ሀ ሐ ኀ ኸ		ሁ ሑ ኁ ኹ	ሂ ሒ ኂ ኺ	ሀ ሃ ሐ ሓ ኀ ኃ ኻ	ሄ ሔ ኄ ኼ	ሆ ሖ ኆ ኾ ኼ		ኍ ኹ	ሗ ኋ ኻ	ኌ ኼ	ኍ ኹ
k							ከ ኰ					
k'							ቀ ቈ					
q							ቆ ቈ					
s	ሥ ስ	ሠ ሰ	ሡ ሱ	ሢ ሲ	ሣ ሳ	ሤ ሴ	ሦ ሶ			ሧ ሷ		
s'	ጸ ፀ	ጸ ፀ	ጹ ፁ	ጺ ፂ	ጻ ፃ	ጼ ፄ	ጸ ፆ					

FIGURE 2.6 – Table des caractères redondants de l'amharique contemporain, in [Teshome and Besacier, 2012]

et 291 701 mots anglais, principalement issus des journaux d'actualité et des textes de droit [21] ;

— le "Lorelei Amharic representative language pack", compilé vers 2018 par le Linguistic Data Consortium, d'accès payant, et qui comprendrait un sous-corpus parallèle amharique-anglais long de 600 000 mots amhariques traduits vers l'anglais et de 80 000 mots anglais traduits vers l'amharique, issus de domaines textuels variés (réseaux sociaux, actualités) [22] ;

— un corpus compilé par [Abate et al., 2018] pour cinq langues éthiopiennes (amharique, tigrigna [23], oromo, wolaytta [24] et guèze) et dont le sous-ensemble amharique-anglais comporterait 628 474 mots amhariques, pour 969 345 mots anglais, issus de textes législatifs et religieux ;

— un corpus compilé par [Biadgline and Smaïli, 2021], mais n'ayant pas encore été mis à la disposition du public, qui comprendrait près de 225 000 segments anglais-amharique issus de textes religieux et de journaux d'information ;

21. http://catalog.elra.info/en-us/repository/browse/ELRA-W0074/
22. https://catalog.ldc.upenn.edu/LDC2018T04
23. ISO 639-1 "ti", ISO 639-2 "tir".
24. ISO 639-2 "wal".

— et un corpus compilé par [Gezmu et al., 2022] et mis gratuitement à la disposition des chercheurs ; ce corpus bâti, semble-t-il, sur les fondations de [Abate et al., 2018] comprend 145 000 paires de segments parallèles, représentant 1 704 972 mots amhariques pour 2 275 561 mots anglais, principalement issus de textes religieux (Bible et journaux des Témoins de Jéhovah) [25].

Des ressources complémentaires existent toutefois :

— deux petits corpus d'évaluation, l'un compilé par [Hadgu et al., 2020] consistant en 2865 segments issus des réseaux sociaux et des médias en ligne [26], et l'autre par Facebook/Meta, en 2021, dans le cadre du projet FLORES-101 (Facebook Low Resource MT Benchmark) dédié à l'évaluation de la TA des langues peu dotées et qui consiste en 3001 segments issus de la version anglaise de Wikipedia accompagnés de traductions amhariques professionnelles commanditées par Facebook [27] ;
— un patchwork de ressources de mauvaise qualité, et souvent mal alignées, distribuées par le projet OPUS et collectées automatiquement à partir du Web [28] ;
— et des corpus parallèles de très faibles dimensions pour l'amharique et des langues autres que l'anglais, telle le guèze [29].

Sur la base du rapport mot/segment observable au sein du corpus compilé par [Gezmu et al., 2022], il semble vraisemblable, par extrapolation, que l'intégralité des données parallèles disponibles pour la paire de langue amharique-anglais (pourtant de très loin le meilleur tandem de l'amharique) échoue à dépasser le seuil fatidique de 0,5 million de segments parallèles ([Ranathunga et al., 2021]) précédemment évoqué et en deçà duquel les langues sont considérées peu dotées, justifiant dès lors l'inclusion de l'amharique dans cette catégorie. Mais cette estimation nécessite sans doute encore d'être revue à la baisse : ces corpus, en effet, sont susceptibles de contenir un grand nombre de segments redondants issus des mêmes sources [30], et en leur sein, seul le corpus compilé par [Gezmu et al., 2022] se trouve accessible gratuitement. Le nombre de segments parallèles effectivement disponibles pour le développement libre de systèmes de TA amharique doit ainsi s'établir, plus vraisemblablement, aux alentours de 0,15 million, frôlant ainsi la limite précédemment évoquée pour la définition de langues dites extrêmement peu dotée ([Ranathunga et al., 2021]).

La rareté objective de ces données parallèles a manifestement entravé la recherche en TA statistique et neuronale en langue amharique, ainsi qu'en témoignent, indirectement, d'une part, la remarquable longévité des approches dites "par règles" de la TA comme tâche pour cette langue (cf. Gasser in [Pauw et al., 2012], p. 41, ou encore [Gasser, 2017]), et, plus directement d'autre part, le fait que la quasi-totalité des travaux majeurs dédiés aux paradigmes neuronaux et statistiques de la TA amharique ([Teshome and Besacier, 2012], [Biadgligne and Smaïli, 2022], [Gezmu et al., 2022]) est précisément consacrée à la compilation de nouveaux corpus.

25. https://www.findke.ovgu.de/findke/en/ Research/Data+Sets/Amharic$_{E}nglish+Parallel+Corpus-p-1144.html$
26. https://zenodo.org/record/3734260.Yw-DeyFByEs. Ce corpus était, à l'été 2022, malheureusement inutilisable car corrompu ; le choix des chercheurs l'ayant compilé de le distribuer sous la forme d'un fichier CSV plutôt qu'au moyen de fichiers texte fut malheureux.
27. https://ai.facebook.com/research/publications/the-flores-101-evaluation-benchmark-for-low-resource-and-multilingual-machine-translation
28. https://opus.nlpl.eu/
29. https://github.com/Amdework21/Geez-Amharic-DS
30. Le corpus compilé par [Gezmu et al., 2022] est ainsi le sur-ensemble du corpus amharique-anglais

ISO code	Language	Tokens (M)	Size (GiB)	ISO code	Language	Tokens (M)	Size (GiB)
af	Afrikaans	242	1.3	lo	Lao	17	0.6
am	Amharic	68	0.8	lt	Lithuanian	1835	13.7
ar	Arabic	2869	28.0	lv	Latvian	1198	8.8
as	Assamese	5	0.1	mg	Malagasy	25	0.2
az	Azerbaijani	783	6.5	mk	Macedonian	449	4.8
be	Belarusian	362	4.3	ml	Malayalam	313	7.6
bg	Bulgarian	5487	57.5	mn	Mongolian	248	3.0
bn	Bengali	525	8.4	mr	Marathi	175	2.8
-	Bengali Romanized	77	0.5	ms	Malay	1318	8.5
br	Breton	16	0.1	my	Burmese	15	0.4
bs	Bosnian	14	0.1	my	Burmese	56	1.6
ca	Catalan	1752	10.1	ne	Nepali	237	3.8
cs	Czech	2498	16.3	nl	Dutch	5025	29.3
cy	Welsh	141	0.8	no	Norwegian	8494	49.0
da	Danish	7823	45.6	om	Oromo	8	0.1
de	German	10297	66.6	or	Oriya	36	0.6
el	Greek	4285	46.9	pa	Punjabi	68	0.8
en	English	55608	300.8	pl	Polish	6490	44.6
eo	Esperanto	157	0.9	ps	Pashto	96	0.7
es	Spanish	9374	53.3	pt	Portuguese	8405	49.1
et	Estonian	843	6.1	ro	Romanian	10354	61.4
eu	Basque	270	2.0	ru	Russian	23408	278.0
fa	Persian	13259	111.6	sa	Sanskrit	17	0.3
fi	Finnish	6730	54.3	sd	Sindhi	50	0.4
fr	French	9780	56.8	si	Sinhala	243	3.6
fy	Western Frisian	29	0.2	sk	Slovak	3525	23.2
ga	Irish	86	0.5	sl	Slovenian	1669	10.3
gd	Scottish Gaelic	21	0.1	so	Somali	62	0.4
gl	Galician	495	2.9	sq	Albanian	918	5.4
gu	Gujarati	140	1.9	sr	Serbian	843	9.1
ha	Hausa	56	0.3	su	Sundanese	10	0.1
he	Hebrew	3399	31.6	sv	Swedish	77.8	12.1
hi	Hindi	1715	20.2	sw	Swahili	275	1.6
-	Hindi Romanized	88	0.5	ta	Tamil	595	12.2
hr	Croatian	3297	20.5	-	Tamil Romanized	36	0.3
hu	Hungarian	7807	58.4	te	Telugu	249	4.7
hy	Armenian	421	5.5	-	Telugu Romanized	39	0.3
id	Indonesian	22704	148.3	th	Thai	1834	71.7
is	Icelandic	505	3.2	tl	Filipino	556	3.1
it	Italian	4983	30.2	tr	Turkish	2736	20.9
ja	Japanese	530	69.3	ug	Uyghur	27	0.4
jv	Javanese	24	0.2	uk	Ukrainian	6.5	84.6
ka	Georgian	469	9.1	ur	Urdu	730	5.7
kk	Kazakh	476	6.4	-	Urdu Romanized	85	0.5
km	Khmer	36	1.5	uz	Uzbek	91	0.7
kn	Kannada	169	3.3	vi	Vietnamese	24757	137.3
ko	Korean	5644	54.2	xh	Xhosa	13	0.1
ku	Kurdish (Kurmanji)	66	0.4	yi	Yiddish	34	0.3
ky	Kyrgyz	94	1.2	zh	Chinese (Simplified)	259	46.9
la	Latin	390	2.5	zh	Chinese (Traditional)	176	16.6

FIGURE 2.7 – Métadonnées du corpus CC-100 ayant servi à l'entraînement du modèle XLM-R ([Conneau et al., 2020], p. 11)

Or ces travaux de compilation tendent à s'appuyer sur l'extraction automatique de contenus issus du Web ("*web scraping*") et se heurtent, en conséquence, à la faible représentation de l'amharique sur la Toile, faible représentation dont les métadonnées du modèle XLM-R, précédemment évoqué, offre une saisissante illustration.

XLM-R fut en effet entraîné sur des données monolingues extraites du Web pour 100 des 183 langues définies par la norme ISO 639-1. La figure 2.8 résente les dimensions respectives (en gigaoctets) des corpus compilés au départ du Web pour 88 de ces langues et l'échelle logarithmique de son axe des ordonnées — indispensable à sa lisibilité — n'entame en rien son caractère saisissant, tant est stupéfiante la disparité des degrés de représentations des langues dans le cyberespace. On retiendra notam-

de [Abate et al., 2018].

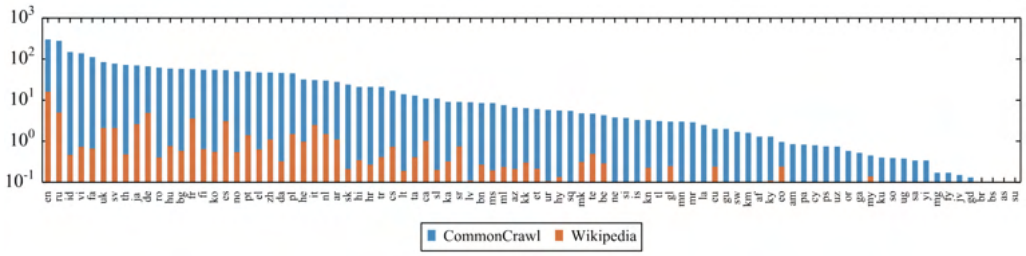


FIGURE 2.8 – Quantité de données (à l'échelle logarithmique) issues de Wikipedia (barres oranges) et de CommonCrawl (barres bleues) ayant servi à l'entraînement du modèle XLM-R ([Conneau et al., 2020])

ment qu'alors qu'une entreprise de l'envergure de Facebook/Meta est ainsi parvenue à extraire 300,8 Go de données textuelles en anglais (cf. figure 2.7), 56,8 Go de données textuelles en français, 3 Go de données textuelles en mongol (ISO 639-1 "mg") ou encore 1,6 Go en swahili (ISO 639 -1 "sw") — deux langues considérées comme peu dotées, voire très peu dotées ([Chimoto and Bassett, 2022], [Gao, 2022]) — elle n'a pu recueillir que 0,8 Go de données textuelles pour l'amharique ([Conneau et al., 2020], p. 11).

Dans l'espoir, donc, de pallier cette pénurie de données issues du Web, le présent mémoire se propose d'explorer la possibilité d'utiliser la lexicographie bilingue comme source de données parallèles en traduction automatique de l'amharique.

2.3 Lexicographie bilingue

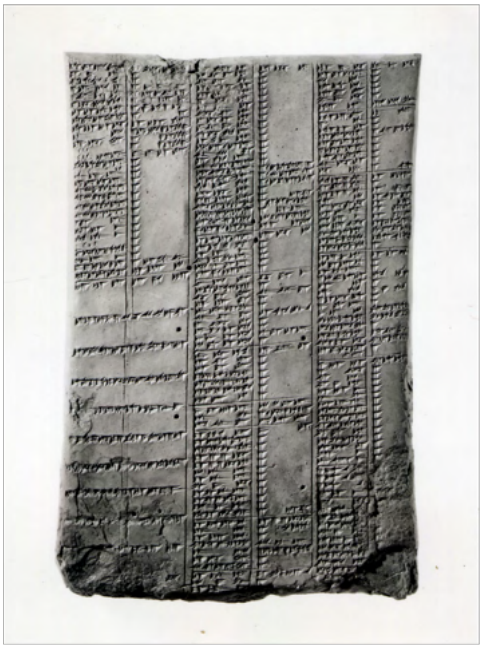


FIGURE 2.9 – Liste de mots sur tablette d'argile assyrienne, British Museum

Des listes de mots bilingues comptent parmi les plus anciens documents écrits

parvenus jusqu'à nous.

La tablette d'argile néo-assyrienne que représente la figure 2.9, numérotée K.4375 dans l'inventaire du British Museum et datant du VIIe siècle avant notre ère, fut exhumée à Ninive, aux abords de l'actuel Mossoul, à la fin des années 1840. Il s'agit de la troisième d'une série de neuf tablettes — dont existent de nombreuses copies antiques — appelée Malku = šarru (d'après la première ligne de la première des tablettes de cette série). Sur ces artefacts, figurent en colonnes des listes d'idéogrammes sumériens, leur transcription en phonogrammes akkadiens et leur traduction dans cette autre langue.

Les entrées de ces listes étaient généralement ordonnées par thème, et une autre série — plus longue — de ces tablettes-listes de mots, appelée HAR-ra = hubullu, se trouve ainsi organisée selon les rubriques suivantes : terminologie juridique et administrative (tablettes 1 et 2), arbres et objets en bois (3-7), roseaux et objets en roseau (8 et 9), poterie (10), peaux et objets de cuivre (11), autres métaux (12), animaux domestiques (13) et sauvages (14), parties du corps (15), pierres (16), plantes (17), oiseaux et poissons (18), textiles (19), termes géographiques (20-22), nourriture et boissons (23 et 24) ([Snell-Hornby, 1986]).

Ces séries de tablettes présentent de remarquables similarités, *mutatis mutandis*, d'avec les artefacts modernes que nous appelons "dictionnaires bilingues".

2.3.1 Nomenclature des dictionnaires bilingues

À la manière de ces tablettes d'argile, en effet, ces ouvrages consistent fondamentalement en des listes de mots d'une langue X systématiquement associés à des mots-formes équivalents, dans une langue Y, potentiellement accompagnés d'informations auxiliaires (indication de prononciation, partie du discours, genre, etc.). La liste des entrées d'un dictionnaire bilingue est communément appelée sa "macrostructure" [31], et si des dictionnaires bilingues dont les entrées sont ordonnées par thèmes [32] existent de nos jours, la majorité des dictionnaires bilingues modernes présentent des macrostructures organisées selon l'ordre alphabétique.

Ces ouvrages modernes peuvent, en outre, être monoscopiques ("*monoscopal*") ([Adamska-Sałaciak, 2022]), c'est-à-dire ne présenter d'entrées que dans une seule de leur deux langues, ou discopiques ("*biscopal*") ([Ibid.]) s'ils présentent une double macrostructure pour chacune de leurs deux langues [33].

L'organisation des entrées proprement dites de ces dictionnaires bilingues — appelée microstructure — peut se trouver réduite à un unique mot de langue Y, mais est bien plus susceptible — du fait de l'anisomorphie ("*anisomorphism*") ([Zgusta, 2010]) relative des langues humaines — d'offrir au lecteur des exemples de l'usage de ces mots-formes en contexte — exemples accompagnés de traductions permettant d'en

31. Le terme de "mégastructure" étant parfois appliqué à l'organisation d'un dictionnaire dans son ensemble, préface, annexes et quatrième de couverture comprises.

32. Une telle disposition est communément qualifiée d'onomasiologique ([Adamska-Sałaciak, 2022]).

33. La dictionnairique contemporaine classe également les dictionnaires bilingues dans deux catégories additionnelles relevant d'une distinction plus subtile entre des dictionnaires dits "bidirectionnels" ("*bidirectional*") et des dictionnaires dits "unidirectionnels" ("*monodirectional*") ([Adamska-Sałaciak, 2022]), selon que les auteurs de ces ouvrages aient fait ou non le choix de prendre en compte la langue maternelle de leurs lecteurs-utilisateurs, omettant, à titre d'exemple — dans le cas des dictionnaires unidirectionnels — des informations relative à la prononciation ou au genre grammatical des mots-formes appartenant à la langue maternelle des lecteurs visés, en vue d'une économie d'espace.

expliciter le sens — ainsi que des gloses et des paraphrases visant à l'explication des *realia*, soit des notions de langue X dénuées d'équivalents dans la langue Y.

2.3.2 La lexicographie bilingue et la TA comme tâche

Dans la mesure où les dictionnaires bilingues — notamment ceux produits pour des domaines techniques des sciences et de l'industrie — sont précisément susceptibles de servir d'outil de travail aux professionnels dont les traductions viennent alimenter les corpus parallèles sur lesquels s'appuie la TA statistique ou neuronale, il est possible d'affirmer que ces ouvrages jouent, à tout le moins, un rôle indirect au sein de l'état de l'art du domaine [34].

Quant au rôle direct joué, potentiellement, par les dictionnaires bilingues au sein de la recherche en TA, il se trouve malheureusement occulté par un fâcheux conflit de mots-clefs dans la littérature du domaine.

En effet le terme de "dictionnaire bilingue" ("*bilingual dictionary*") est largement employé, dans le contexte de la TA comme domaine, pour désigner des listes de paires de mots, automatiquement alignés, et généralement associés à la fréquence de leur occurrence conjointe au sein d'un corpus parallèle. Ces "dictionnaires bilingues" constituaient un élément essentiel des mécanismes techniques de la TA statistique, et de nombreux chercheurs semblent s'être employés à trouver à ces données une nouvelle utilité dans le contexte de la TA neuronale, les intégrant, à titre d'exemple, à des sous-modules visant à guider les étapes de génération du décodeur des modèles de traduction neuronaux ([Arthur et al., 2016], [Post and Vilar, 2018], [Takebayashi et al., 2018]), ou bien à des modules ayant partie liée avec la génération de plongements de mots ([Zhong and Chiang, 2020], ARTICLE RÉTRACTÉ), ou encore tout cela à la fois ([Zhang et al., 2021])! De tels "dictionnaires bilingues", automatiquement générés au départ de corpus parallèles, ont également été mis à profit, de manière remarquable, en TA des langues peu dotées, pour la génération de données de synthèse ("*data augmentation*") ([Fadaee et al., 2017, Fernando and Ranathunga, 2022]) ; associés à des modèles de langue du type d' XLM-R, mentionné précédemment, ces "dictionnaires bilingues" permettent, dans ces configurations, l'insertion de mots rares et de leur traductions dans de nouveaux contextes et améliorent ainsi la capacité impartie aux modèles de la TA neuronale d'en capturer les propriétés sémantiques.

```
safely sicher 0.0051237409068
safemode safemode 1
safeness antikollisionssystem 0.3333333
safer sicherer 0.09545972221228
```

FIGURE 2.10 – "Dictionnaire bilingue" anglais-allemand, au sens usuel de ce terme en TA comme domaine

34. Mais le rôle indirect joué, potentiellement, par les dictionnaires bilingues dans le domaine de la TA a d'autres facettes ; c'est ainsi, par exemple, aux travaux lexicographiques de Malinowski que nous devons sans doute, en dernière analyse, quoique par l'intermédiaire de J. R. Firth, les théories de la sémantique distributionnelle au cœur de l'état de l'art du TAL contemporain. (cf. [Robins, 2013], p. 246).

Ce sont néanmoins les dictionnaires bilingues du type des tablettes d'argile présentées ci-avant ou des artefacts modernes de l'édition lexicographique qui intéressent cette étude. Afin, donc, de se prémunir de cette confusion, le terme de lexicographie bilingue désignera, autant que faire se pourra, ces ouvrages pour le restant de cette étude.

La lexicographie bilingue ainsi définie semble, cependant, presque totalement absente de l'état de l'art contemporain de la TA comme domaine, quoique le conflit de mots-clefs précédemment évoqué puisse expliquer cette apparente absence par une occultation.

L'un des rares exemples d'un travail de recherche évoquant des enjeux similaires à ceux du présent mémoire est [Rijhwani et al., 2020]. Les auteurs de cette étude — prenant acte du manque de données à la disposition des chercheurs souhaitant développer des applications du TAL pour des langues dites "en danger" ("*endangered*") — s'emploient à évaluer la capacité de systèmes de reconnaissance optique des caractères ("*OCR*") à extraire ces données manquantes au départ de sources bibliographiques consistant en des livres imprimés et des images de textes scannés. Pourtant ces auteurs – bien qu'ils mentionnent des listes de mots ("*word lists*"), des gloses manuscrites de notes de chercheurs, des livres multilingues et des dictionnaires, et recourent à des traductions existantes des documents traités pour en assister la reconnaissance optique – n'évoquent jamais dans leur article l'extraction de segments parallèles au départ de ces textes, ni l'impact de données d'un tel type sur les performances "en aval" d'un système de TA pour une langue peu dotée ; et il en est de même d'une autre étude ([Bustamante et al., 2020]), laquelle s'applique également à extraire des données adéquates au traitement automatique de langues dites "vraiment peu dotées" ("*truly low-resource*") depuis des documents PDF multilingues, mais uniquement dans l'objectif d'obtenir par ce biais des corpus monolingues.

L'expérimentation présentée dans la seconde partie de ce mémoire s'attachera précisément à tenter d'extraire d'un document relevant de la lexicographie bilingue anglais-amharique un corpus parallèles de segments alignés et d'étudier l'impact de telles données sur les performances d'un système de TA comme tâche pour cette paire de langues.

Deuxième partie

Expérimentations / chaîne opératoire

CHAPITRE 3

DONNÉES FONDAMENTALES

Sommaire

3.1 Introduction **55**
3.2 Wolf Leslau et l'English-Amharic Context Dictionary **56**
3.3 Images scannées/photographiées **59**

3.1 Introduction

Au tournant des années 2020, la TA comme domaine, forte de remarquables succès d'estime, semble s'être mise en quête de nouveaux défis et avoir jeté son dévolu sur les applications de son objet d'étude à des langues dites "peu dotées", voire "très peu dotées" ou même "extrêmement peu dotées", soit à des langues pour lesquelles les données nécessaires au développement de systèmes de TA comme tâche — lesquelles données consistent en des corpus parallèles de traductions associées à leurs textes source — n'existent qu'en quantités infimes.

Au printemps 2022, dans un post de blog [1], des chercheurs employés par Google Translate allaient ainsi jusqu'à faire part de leur ambition de développer des systèmes de TA comme tâche dédiés à des langues dites "zéro-dotées" ou non dotées ("*zero-resource languages*") dont la liste comprenait nombre de langues africaines, tels le bambara [2], le lingala [3], le luganda [4], ou encore l'oromo [5] et le trigrigna [6], des langues éthiopiennes. Il est vrai que le degré zéro des approches visant à permettre l'obtention des données nécessaires au développement des systèmes contemporains de la TA comme tâche consiste à tenter d'extraire ces données depuis le Web où les langues du monde sont très inégalement représentées, ainsi qu'en attestent, à titre d'exemple, les métadonnées précédemment évoquées du modèle XLM-R (p. 47).

Il n'aura cependant pas échappé aux chercheurs en langues orientales que les idiomes "zéro-dotés" ainsi désignés ne sont pourtant pas dénués de ressources, compte tenu des nombreux travaux existants, fruits du labeur d'éminents linguistes, ayant notamment consisté en la compilation d'ouvrages relevant de la lexicographie bilingue destinés aux apprenants et aux chercheurs de ces langues.

1. Cf. https://ai.googleblog.com/2022/05/24-new-languages-google-translate.html
2. ISO 639-1 "bm", ISO 639-2 "bam".
3. ISO 639-1 "ln", ISO 639-2 "lin".
4. ISO 639-1 "lg", ISO 639-2 "lug".
5. ISO 639-1 "om", ISO 639-2 "orm".
6. ISO 639-1 "ti", ISO 639-2 "tir".

La seconde partie de ce mémoire s'attache à présenter une suite d'expériences spécifiquement dévolue à l'étude des moyens d'obtenir un corpus parallèle de données propices au développement d'un système de TA neuronale pour la paire de langue amharique-anglais, une suite d'expériences en forme de "chaîne opératoire" susceptible d'être, par là même, appliquée à d'autres de ces idiomes dits "peu dotés".

Le présent chapitre offre une présentation des données lexicographiques ayant servi de base à cette expérience.

3.2 Wolf Leslau et l'English-Amharic Context Dictionary

Le choix d'un ouvrage relevant de la lexicographie bilingue pour la paire de langues amharique-anglais s'est naturellement porté sur le remarquable English-Amharic Context Dictionary de Wolf Leslau ([Leslau, 1973]). Cet auteur, né à Krzepice dans l'actuelle Pologne, en 1906, étudiant, à Paris, du sémitiste Marcel Cohen qui l'initia à l'amharique, à l'École Nationale des Langues Orientales [7] dans les années 1930, fut interné au Camp des Milles et sauvé *in extremis* de la déportation vers Auschwitz, en 1942, par une organisation de secours international qui facilita sa fuite vers les États-Unis où il devait devenir l'un des plus éminents éthiopisants du XXème siècle et du début du XXIème siècle [8].

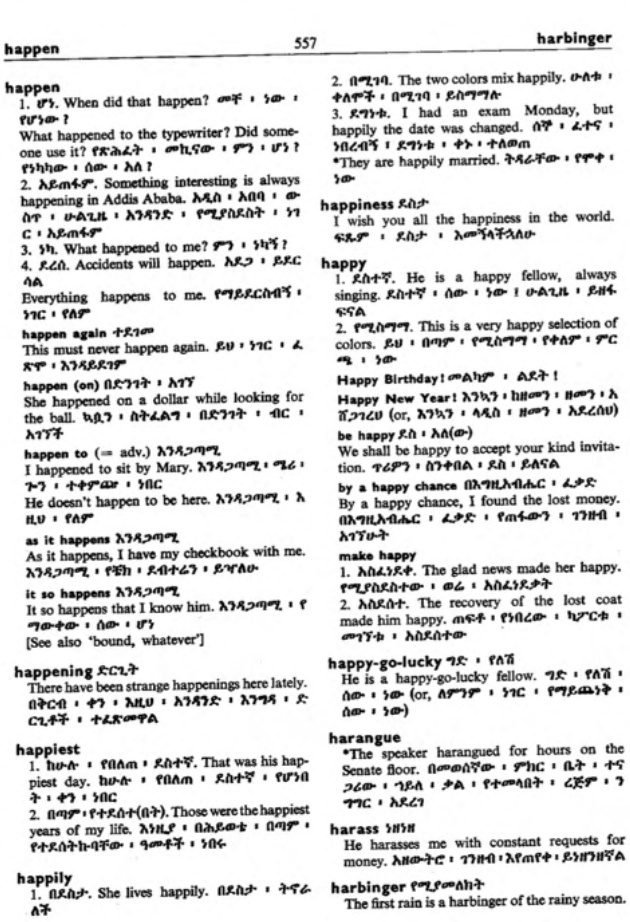
FIGURE 3.1 – Wolf Leslau en 2004, lors de la 32ème Conférence Nord Américaine de Linguistique Afro-Asiatique, Pete Unseth

La compilation du dictionnaire qui intéresse cette étude fut entreprise par cet auteur en 1966 et dura six années ; elle fut financée principalement par le Départe-

7. L'un des nombreux avatars passés de l'actuelle INALCO.
8. Wolf Leslau mourut centennaire à Fullerton, en Californie, en 2006.

ment de la Santé, de l'Éducation et des Services Sociaux des États-Unis, ainsi que par le Near Eastern Center, l'African Studies Center et le Committee on International and Comparative Studies de l'Université de Californie à Los Angeles (UCLA). De nombreux étudiants de Wolf Leslau colaborèrent également à la compilation de ce dictionnaire – notamment Tesfaye Shewaye et Girma Wolde Sellasie, Abebe Bekele et Thomas Leiper Kane, qui devait être l'auteur d'autres ouvrages en lexicographie de l'amharique – lequel fut publié, en 1973, par l'éditeur allemand Otto Harrassowitz (Wiesbaden) et dont les frais d'impression furent pris en charge par l'Agence des États-Unis pour le Développement International (USAID).

Le dictionnaire bilingue résultant de ces efforts est de nature monoscopique, n'étant consultable que dans la direction anglais → amharique, et peut également être qualifié d'unidirectionnel, semblant principalement dédié à des usagers ayant l'anglais pour langue maternelle, la prononciation des mots de la langue de Shakespeare n'y étant jamais indiquée aux lecteurs étrangers.



happen 557 harbinger

happen
1. ሆነ. When did that happen? መቼ ፣ ነው ፣ የሆነው ?
What happened to the typewriter? Did someone use it? የጽሕፈት ፣ መኪናው ፣ ምን ፣ ሆነ ? የነካካው ፣ ሰው ፣ አሰ ?
2. አይጠፋም. Something interesting is always happening in Addis Ababa. አዲስ ፣ አበባ ፣ ውስጥ ፣ ሁልጊዜ ፣ አንዳንድ ፣ የሚያስደስት ፣ ነገር ፣ አይጠፋም
3. ነክ. What happened to me? ምን ፣ ነካኝ ?
4. ደረሰ. Accidents will happen. አደጋ ፣ ይደር ሳል
Everything happens to me. የማይደርስብኝ ፣ ነገር ፣ የለም

happen again ተደገመ
This must never happen again. ይህ ፣ ነገር ፣ ፈጽሞ ፣ እንዳይደገም

happen (on) በድንገት ፣ አገኘ
She happened on a dollar while looking for the ball. ኳሷን ፣ ስትፈልግ ፣ በድንገት ፣ ብር ፣ አገኘች

happen to (= adv.) እንዳጋጣሚ
I happened to sit by Mary. እንዳጋጣሚ ፣ ሜሪ ፣ ጐን ፣ ተቀምጩ ፣ ነበር
He doesn't happen to be here. እንዳጋጣሚ ፣ እዚህ ፣ የለም

as it happens እንዳጋጣሚ
As it happens, I have my checkbook with me. እንዳጋጣሚ ፣ የቼክ ፣ ደብተሬን ፣ ይዣለሁ

it so happens እንዳጋጣሚ
It so happens that I know him. እንዳጋጣሚ ፣ የ ማውቀው ፣ ሰው ፣ ሆነ
[See also 'bound, whatever']

happening ድርጊት
There have been strange happenings here lately. በቅርብ ፣ ቀን ፣ እዚህ ፣ አንዳንድ ፣ እንግዳ ፣ ድርጊቶች ፣ ተፈጽመዋል

happiest
1. ከሁሉ ፣ የበለጠ ፣ ደስተኛ. That was his happiest day. ከሁሉ ፣ የበለጠ ፣ ደስተኛ ፣ የሆነበት ፣ ቀን ፣ ነበር
2. በጣም ፣ የተደሰተ(በት). Those were the happiest years of my life. እነዚያ ፣ በሕይወቴ ፣ በጣም ፣ የተደሰትኩባቸው ፣ ዓመቶች ፣ ነበሩ

happily
1. በደስታ. She lives happily. በደስታ ፣ ትኖራለች
2. በሚገባ. The two colors mix happily. ሁለቱ ፣ ቀለሞች ፣ በሚገባ ፣ ይስማማሉ
3. ደግነቱ. I had an exam Monday, but happily the date was changed. ሰኞ ፣ ፈተና ፣ ነበረብኝ ፣ ደግነቱ ፣ ቀኑ ፣ ተለወጠ
*They are happily married. ትዳራቸው ፣ የሞቀ ፣ ነው

happiness ደስታ
I wish you all the happiness in the world. ፍጹም ፣ ደስታ ፣ እመኝላችኋለሁ

happy
1. ደስተኛ. He is a happy fellow, always singing. ደስተኛ ፣ ሰው ፣ ነው ! ሁልጊዜ ፣ ይዘፋፍናል
2. የሚስማማ. This is a very happy selection of colors. ይህ ፣ በጣም ፣ የሚስማማ ፣ የቀለም ፣ ምርጫ ፣ ነው

Happy Birthday! መልካም ፣ ልደት !

Happy New Year! እንኳን ፣ ከዘመን ፣ ዘመን ፣ አሸጋገረህ (or, እንኳን ፣ ላዲስ ፣ ዘመን ፣ አደረሰህ)

be happy ደስ ፣ አለ(ው)
We shall be happy to accept your kind invitation. ጥሪዎን ፣ በንቀበል ፣ ደስ ፣ ይለናል

by a happy chance በእግዚአብሔር ፣ ፈቃድ
By a happy chance, I found the lost money. በእግዚአብሔር ፣ ፈቃድ ፣ የጠፋውን ፣ ገንዘብ ፣ አገኘሁት

make happy
1. አስፈነደቀ. The glad news made her happy. የሚያስደስተው ፣ ወሬ ፣ አስፈነደቃት
2. አስደሰተ. The recovery of the lost coat made him happy. ጠፍቶ ፣ የነበረው ፣ ካፖርቱ ፣ መገኘቱ ፣ አስደሰተው

happy-go-lucky ግድ ፣ የለሽ
He is a happy-go-lucky fellow. ግድ ፣ የለሽ ፣ ሰው ፣ ነው (or, ለምንም ፣ ነገር ፣ የማይጨነቅ ፣ ሰው ፣ ነው)

harangue
*The speaker harangued for hours on the Senate floor. በመወሰኛው ፣ ምክር ፣ ቤት ፣ ተናጋሪው ፣ ገይለ ፣ ቃል ፣ የተመላበት ፣ ረጅም ፣ ንግግር ፣ አደረገ

harass ነዘነዘ
He harasses me with constant requests for money. አዘውትሮ ፣ ገንዘብ ፣ እየጠየቀ ፣ ይነዘንዘኛል

harbinger የሚያመለክት
The first rain is a harbinger of the rainy season.

FIGURE 3.2 – Exemple d'une page de l'English-Amharic Context Dictionary, [Leslau, 1973]

La mégastructure de ce dictionnaire comprend – outre la macrostructure de ses entrées, longue de 1503 pages numérotées en chiffres arabes, organisée par ordre alphabétique et inspirée, ainsi que l'indique Wolf Leslau, par d'autres dictionnaires de langue anglaise – une préface de douze pages numérotées en chiffres romains ayant pour épigraphe une citation de Samuel Johnson. La microstructure des entrées com-

prend, elle, un lemme anglais imprimé en fonte grasse et dans une police sans sérif, accompagné, parfois, de la mention d'une partie du discours entre parenthèses visant à distinguer les lemmes anglais ambigus (tels "*abandon* (v.)" et "*abandon* (n.)"). Dans les cas d'isomorphie sémantique entre anglais et amharique, le lemme unique équivalent de l'amharique figure à la suite directe du lemme anglais ; dans les cas (fréquents) d'anisomorphie dans la structure relative des lexiques anglais et amharique, des sous-entrées numérotées en chiffres arabes introduisent chacun des lemmes amhariques distincts. La microstructure de la quasi-totalité des entrées comportent enfin des exemples d'usages des lemmes anglais, imprimés dans une police sérifée, suivis de leur traduction en amharique.

FIGURE 3.3 – "ale [une sorte de bière]", exemple de *realia* glosé entre crochets

La préface de ce dictionnaire apprend à son lecteur que les traductions amhariques des exemples anglais ont été produites et validées par de multiples traducteurs ayant reçu pour instruction de viser à l'idiomaticité de ces segments de texte traduits du point de vue de l'amharique et d'éviter, en conséquence, les traductions littérales, donnant parfois lieu à des adaptations ; ainsi, à titre d'exemples, les mots anglais "winter" et "summer" ont été traduits, respectivement, par ክረምት ("kèrèmt", la saison des pluies de la Corne de l'Afrique qui s'étend de juin à septembre) et par በጋ ("bèga", la saison sèche allant d'octobre à mai). De nombreux segments anglais ont, par ailleurs, été préfixés d'une astérisque ("*"), indiquant qu'aucun équivalent exact du segment en question n'existe en amharique et que la traduction donnée est dès lors une paraphrase, ou encore d'un signe plus surscrit ("+"), signalant que la diathèse du verbe anglais diffère de celle du verbe amharique en traduction, le sujet du premier devenant, par exemple, l'objet du second ou vice versa. Un symbole égal ("=") suivi du nom d'une partie du discours entre parenthèses signale, quant à lui, les cas dans lesquels le lemme anglais de l'entrée ou de la sous-entrée se voit traduit en amharique par un mot-forme relevant d'une autre partie du discours. Enfin les *realia* de la langue anglaise sont à l'occasion glosées entre crochets par une notice en amharique (cf. figure 3.3).

S'agissant des normes orthographiques, celle retenue par l'auteur pour l'anglais est américaine plutôt que britannique, et dans le cas de l'amharique, Wolf Leslau s'est employé, selon ses propres dires, à adopter, à un nombre très limité d'exceptions [9], l'orthographe étymologique des mots de cette langue dans la sélection des variantes d'écritures découlant de l'existence, en amharique contemporain, de caractères homophones issus de l'alphasyllabaire éthiopien.

9. Une telle exception réside, par exemple, dans le choix de l'auteur d'orthographier l'amharique *"s'ra"* "travail" en ሥራ plutôt qu'en ስራ — pourtant son orthographe étymologique — par respect pour la graphie dominante dans l'usage écrit de ce mot en amharique moderne, laquelle est susceptible de constituer un cas intéressant de synesthésie orthographique tenant du fameux et controversé effet "bouba/kikki" (cf. https://fr.wikipedia.org/wiki/Effet$_{b}ouba - kiki$), à la faveur duquel le trident du graphème ሥ pourrait être considéré comme l'emportant sur la rondeur du ስ dans la désignation de la notion de travail.

3.3 Images scannées/photographiées

FIGURE 3.4 – Système de reprographie pour la photographie d'ouvrages fragiles en possession de la BULAC, BULAC

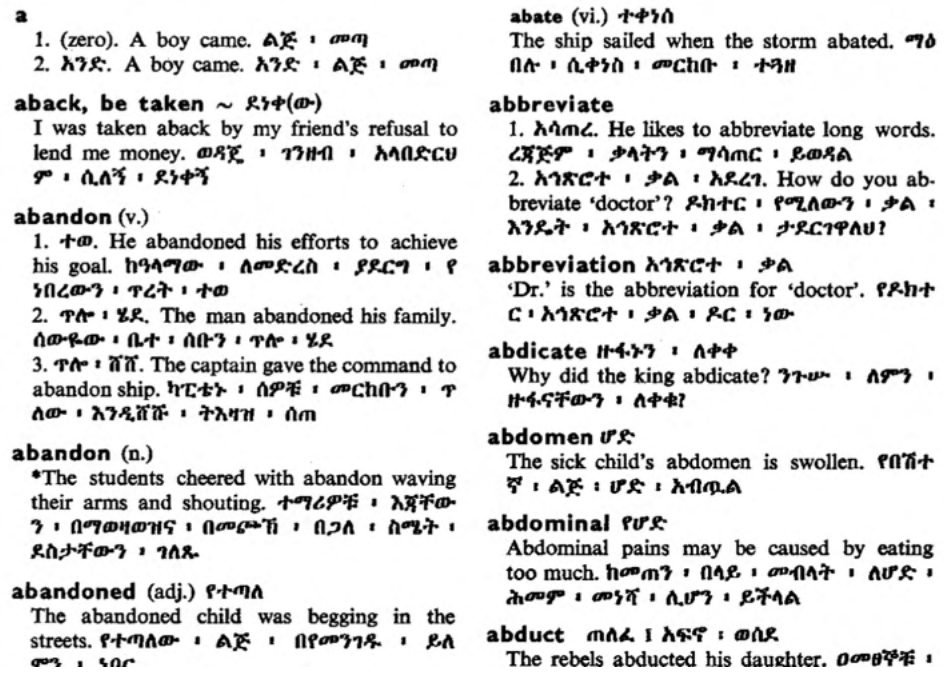
Le dictionnaire présenté ci-avant n'a pas été photographié/scanné dans le cadre de la présente étude, deux fichiers PDF de cet ouvrage — baptisés PDF-A et PDF-B pour le reste de ce chapitre — étant d'ores et déjà disponibles en ligne sur la Toile.

Le lecteur qui souhaiterait "scanner" un tel document par ses propres moyens devra se prémunir de quelques précautions, notamment dans l'éventualité où le document en question serait rare ou précieux. La photocopie classique, en effet, est susceptible d'endommager de tels ouvrages, l'électricité statique générée par les photocopieurs étant, par exemple, capable d'arracher une partie – infime – des pigments des encres de ces documents en les retenant contre la vitre de ces artefacts et risque ainsi d'estomper le texte imprimé ou manuscrit de ces ouvrages sur le long cours ; il est donc recommandé de photographier ces documents – plutôt que de les scanner au photocopieur – au moyen d'équipements d'un type dont la BULAC possédait, vers 2020, au moins deux exemplaires installés dans des cabines de reprographie en libre accès (cf. figure 3.6) [10].

Des deux exemplaires PDF de ce document disponibles en ligne, PDF-A est le plus complet, avec 1502 pages sur les 1503 que compte l'ouvrage, et semble avoir été photographié aux États-Unis, certaines de ses pages portant le tampon d'une fondation texane, et avoir été mis en ligne vers 2011 au plus tard. Les fichiers PNG extraits de ce document PDF sont de dimensions 1613×2212, PDF-A ayant par ailleurs fait l'objet d'une opération de reconnaissance optique des caractères hautement fautive pour l'amharique dont le script est interprété comme s'il s'agissait des caractères de l'alphabet latin. PDF-B est quant à lui très incomplet, ne présentant que 1175 pages sur les 1503, et est susceptible d'avoir été numérisé en Thaïlande ou en Malaisie et mis en ligne vers 2012, n'ayant par ailleurs fait l'objet d'aucune opération de reconnaissance optique des caractères et ne comporte donc aucune couche de texte.

10. Cf. tutoriel vidéo produit par la BULAC pour l'utilisation de ces outils : https://www.youtube.com/watch?v=wiTywNwHfd8 .

Si les dimensions des images de PDF-B sont supérieures à celles de PDF-A, valant 3333×4725, leur qualité est paradoxalement moindre et comme floutée et vague. Par chance, néanmoins, PDF-B possède la page manquante à PDF-A, et les données de départ de l'expérience décrite dans la suite de cette section résultent ainsi de la combinaison de ces deux documents, quoique dans des proportions extrêmement inégales (1502 pages pour 1).



a
1. (zero). A boy came. ልጅ ፡ መጣ
2. አንድ. A boy came. አንድ ፡ ልጅ ፡ መጣ

aback, be taken ~ ደነቀ(ው)
I was taken aback by my friend's refusal to lend me money. ወዳጄ ፡ ገንዘብ ፡ አላበድርህ ም ፡ ሲለኝ ፡ ደነቀኝ

abandon (v.)
1. ተወ. He abandoned his efforts to achieve his goal. ከዓላማው ፡ ለመድረስ ፡ ያደርግ ፡ የ ነበረውን ፡ ጥረት ፡ ተወ
2. ጥሎ ፡ ሄደ. The man abandoned his family. ሰውዬው ፡ ቤተ ፡ ሰቡን ፡ ጥሎ ፡ ሄደ
3. ጥሎ ፡ ሸሸ. The captain gave the command to abandon ship. ካፒቴኑ ፡ ሰዎቹ ፡ መርከቡን ፡ ጥ ለው ፡ እንዲሸሹ ፡ ትእዛዝ ፡ ሰጠ

abandon (n.)
*The students cheered with abandon waving their arms and shouting. ተማሪዎቹ ፡ እጃቸው ን ፡ በማወዛወዝና ፡ በመጮኽ ፡ በጋለ ፡ ስሜት ፡ ደስታቸውን ፡ ገለጹ

abandoned (adj.) የተጣለ
The abandoned child was begging in the streets. የተጣለው ፡ ልጅ ፡ በየመንገዱ ፡ ይለ

abate (vi.) ተቀነሰ
The ship sailed when the storm abated. ማዕ በሉ ፡ ሲቀንስ ፡ መርከቡ ፡ ተጓዘ

abbreviate
1. አሳጠረ. He likes to abbreviate long words. ረጃጅም ፡ ቃላትን ፡ ማሳጠር ፡ ይወዳል
2. አጎጽሮተ ፡ ቃል ፡ አደረገ. How do you abbreviate 'doctor'? ዶክተር ፡ የሚለውን ፡ ቃል ፡ እንዴት ፡ አጎጽሮተ ፡ ቃል ፡ ታደርገዋለህ?

abbreviation አጎጽሮተ ፡ ቃል
'Dr.' is the abbreviation for 'doctor'. የዶክተ ር ፡ አጎጽሮተ ፡ ቃል ፡ ዶር ፡ ነው

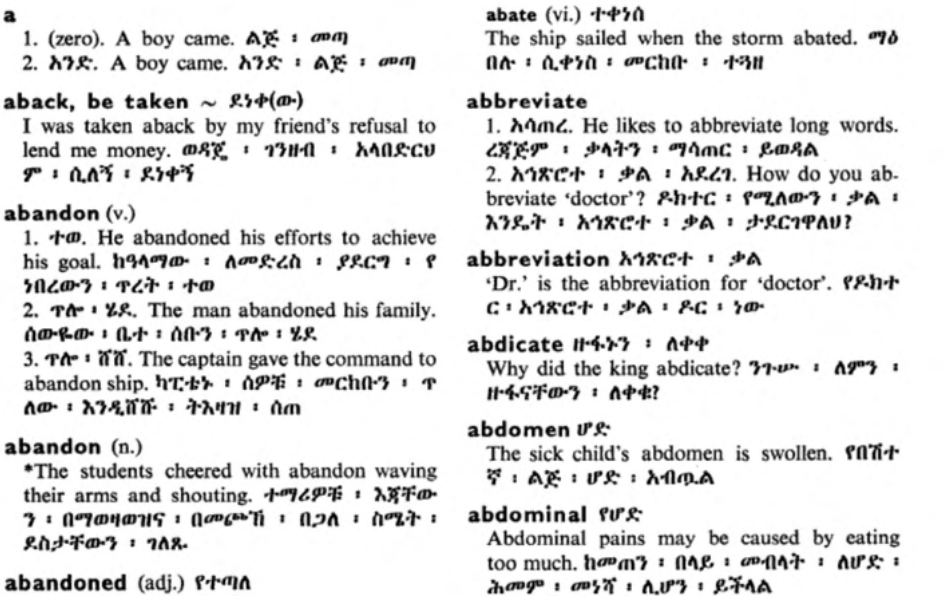
abdicate ዙፋኑን ፡ ለቀቀ
Why did the king abdicate? ንጉሡ ፡ ለምን ፡ ዙፋናቸውን ፡ ለቀቁ?

abdomen ሆድ
The sick child's abdomen is swollen. የበሽተ ኛ ፡ ልጅ ፡ ሆድ ፡ አብጧል

abdominal የሆድ
Abdominal pains may be caused by eating too much. ከመጠን ፡ በላይ ፡ መብላት ፡ ለሆድ ፡ ሕመም ፡ መነሻ ፡ ሊሆን ፡ ይችላል

abduct ጠለፈ I አፍኖ ፡ ወሰደ
The rebels abducted his daughter. ዐመፀኞቹ ፡

FIGURE 3.5 – "PDF-A", aperçu de la première page



a
1. (zero). A boy came. ልጅ ፡ መጣ
2. አንድ. A boy came. አንድ ፡ ልጅ ፡ መጣ

aback, be taken ~ ደነቀ(ው)
I was taken aback by my friend's refusal to lend me money. ወዳጄ ፡ ገንዘብ ፡ አላበድርህ ም ፡ ሲለኝ ፡ ደነቀኝ

abandon (v.)
1. ተወ. He abandoned his efforts to achieve his goal. ከዓላማው ፡ ለመድረስ ፡ ያደርግ ፡ የ ነበረውን ፡ ጥረት ፡ ተወ
2. ጥሎ ፡ ሄደ. The man abandoned his family. ሰውዬው ፡ ቤተ ፡ ሰቡን ፡ ጥሎ ፡ ሄደ
3. ጥሎ ፡ ሸሸ. The captain gave the command to abandon ship. ካፒቴኑ ፡ ሰዎቹ ፡ መርከቡን ፡ ጥ ለው ፡ እንዲሸሹ ፡ ትእዛዝ ፡ ሰጠ

abandon (n.)
*The students cheered with abandon waving their arms and shouting. ተማሪዎቹ ፡ እጃቸው ን ፡ በማወዛወዝና ፡ በመጮኽ ፡ በጋለ ፡ ስሜት ፡ ደስታቸውን ፡ ገለጹ

abandoned (adj.) የተጣለ

abate (vi.) ተቀነሰ
The ship sailed when the storm abated. ማዕ በሉ ፡ ሲቀንስ ፡ መርከቡ ፡ ተጓዘ

abbreviate
1. አሳጠረ. He likes to abbreviate long words. ረጃጅም ፡ ቃላትን ፡ ማሳጠር ፡ ይወዳል
2. አጎጽሮተ ፡ ቃል ፡ አደረገ. How do you abbreviate 'doctor'? ዶክተር ፡ የሚለውን ፡ ቃል ፡ እንዴት ፡ አጎጽሮተ ፡ ቃል ፡ ታደርገዋለህ?

abbreviation አጎጽሮተ ፡ ቃል
'Dr.' is the abbreviation for 'doctor'. የዶክተ ር ፡ አጎጽሮተ ፡ ቃል ፡ ዶር ፡ ነው

abdicate ዙፋኑን ፡ ለቀቀ
Why did the king abdicate? ንጉሡ ፡ ለምን ፡ ዙፋናቸውን ፡ ለቀቁ?

abdomen ሆድ
The sick child's abdomen is swollen. የበሽተ ኛ ፡ ልጅ ፡ ሆድ ፡ አብጧል

abdominal የሆድ
Abdominal pains may be caused by eating too much. ከመጠን ፡ በላይ ፡ መብላት ፡ ለሆድ ፡ ሕመም ፡ መነሻ ፡ ሊሆን ፡ ይችላል

FIGURE 3.6 – "PDF-B", aperçu de la première page

CHAPITRE 4

PROTOCOLE

Sommaire

4.1 Introduction **61**
4.2 Segmentation et reconnaissance optique des caractères **61**
4.2.1 Segmentation des images brutes 63
4.2.2 Reconnaissance optique des caractères des lignes 68
4.3 Extraction d'un corpus de segments alignés **71**
4.4 Entraînement de modèles et augmentation des données par rétro-traduction **77**

4.1 Introduction

Au départ des données fondamentales présentées ci-avant et consistant en des images scannées ou photographiées des pages de l'English-Amharic Context Dictionary, le protocole expérimental présenté tout au long du présent chapitre vise à l'extraction d'un corpus parallèle susceptible de permettre l'entraînement de systèmes de TA comme tâche pour l'amharique, langue peu dotée, et l'anglais. Ce protocole se découpe en trois phases principales, la première consistant à produire la transcription sous du contenu textuel des quelques 1500 pages photographiées de ce dictionnaire ; la seconde à extraire du texte informatique résultant de cette transcription un corpus de segments amharique-anglais rigoureusement alignés, comprenant les exemples anglais du dictionnaire et leur traduction amharique à l'exclusion du reste de la structure des entrées ; et la troisième phase à entraîner – au départ de ces données parallèles issues de la lexicographie bilingue mêlées à des segments tirés d'autres sources et de rétrotraductions – de multiples modèles Transformer afin d'évaluer l'impact de telles données sur la TA comme tâche. Ce protocole expérimental voudrait par ailleurs offrir au chercheur novice en TA des langues orientales un guide en forme de chaîne opératoire pour le développement des artefacts de la TA comme domaine, au départ de simples données brutes (les images des pages d'un dictionnaire), et s'accompagne en conséquence de – trop – nombreuses illustrations et d'exemples de code informatique, à la manière d'un tutoriel.

4.2 Segmentation et reconnaissance optique des caractères

La première étape de ce protocole expérimental/chaîne opératoire doit consister à produire, au départ des seules images scannées ou photographiées des pages du

dictionnaire présenté au fil du chapitre précédent, un texte informatique consistant en la transcription fidèle du contenu de cet ouvrage. Pour ce faire, le présent protocole prend appui sur un formidable outil appelé eScriptorium.

FIGURE 4.1 – Logo d'eScriptorium

Créé vers 2019, dans le cadre de l'Initiative de Recherches Interdisciplinaires et Stratégiques (IRIS) "Scripta" de l'université Paris Sciences Lettres (PSL), et résultant du travail des chercheurs et des techniciens Daniel Stökl Ben Ezra [1], Peter A. Stokes [2], Marc Bui [3], Benjamin Kiessling [4] et Robin Tissot [5], eScriptorium est une plateforme en libre accès [6] dédiée à la transcription automatique de documents paléographiques, y compris ceux rédigés dans des écritures non latines, et associe à la remarquable ergonomie d'une application Docker utilisable dans un navigateur Web le *nec plus ultra* de la reconnaissance optique des caractères, au travers du moteur OCR neuronal Kraken, basé sur pyTorch [7].

Le lecteur qui le souhaite pourra se reporter avec profit aux multiples présentations et tutoriels vidéos disponibles en ligne pour eScriptorium [8], de succincts tutoriels écrits et illustrés existant également, en français [9] et en anglais [10].

Afin de procéder à la transcription du contenu textuel des leurs images, le fichier PDF-A et sa page manquante issue du fichier PDF-B ont été subdivisés, par commodité, en sous-fichiers contenant, chacun, l'ensemble des pages d'une lettre donnée de la macrostructure alphabétique (cf. figure 4.2).

1. Directeur d'études, École Pratique des Hautes Études, Université PSL.
2. Directeur d'études, École Pratique des Hautes Études, Université PSL.
3. Directeur d'études, École Pratique des Hautes Études, Université PSL.
4. Ingénieur en vision industrielle, École Pratique des Hautes Études, Université PSL.
5. Chef de projet informatique, École Pratique des Hautes Études, Université PSL.
6. Code source : https://gitlab.inria.fr/scripta/escriptorium.
7. https://github.com/mittagessen/kraken
8. Dont, en français, https://vimeo.com/516105940, à partir de 1 :02 :51, ou bien encore en anglais, https://www.youtube.com/watch?v=N0hSNC3YvD4, https://vimeo.com/channels/1602497
9. https://lectaurep.hypotheses.org/documentation/prendre-en-main-escriptorium
10. https://lectaurep.hypotheses.org/documentation/escriptorium-tutorial-en

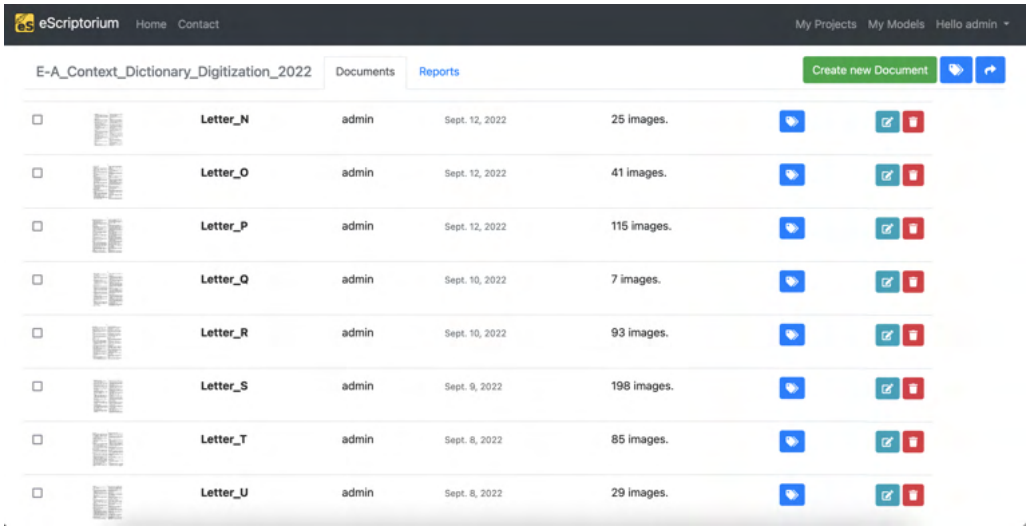


FIGURE 4.2 – Sous-fichiers PDF importés dans l'interface d'eScriptorium reprenant l'ordre alphabétique de la macrostructure du dictionnaire

Le module d'import des fichiers PDF d'eScriptorium étant capable sans intervention de débarrasser les images de ces fichiers PDF d'éventuelles couches de données textuelles et de les convertir au format PNG, l'utilisateur pourra faire l'économie de cette opération préalable (cf. figure 4.3).

FIGURE 4.3 – Méthodes d'import des données brutes dans l'interface d'eScriptorium

4.2.1 Segmentation des images brutes

La première étape du traitement de ces images brutes consiste à les subdiviser en un ensemble de segments de dimensions propres à être traités par l'OCR de cette application. La mise en page du texte de l'English-Amharic Context Dictionary en colonnes — une disposition fréquente pour les ouvrages relevant de la lexicographie bilingue, et ce depuis la plus haute antiquité ainsi qu'indiqué précédemment — fut ici la cause de quelques déboires, le segmenteur de page par défaut d'eScriptorium ne sachant pas gérer une telle disposition du texte des pages (cf. figure 4.4).

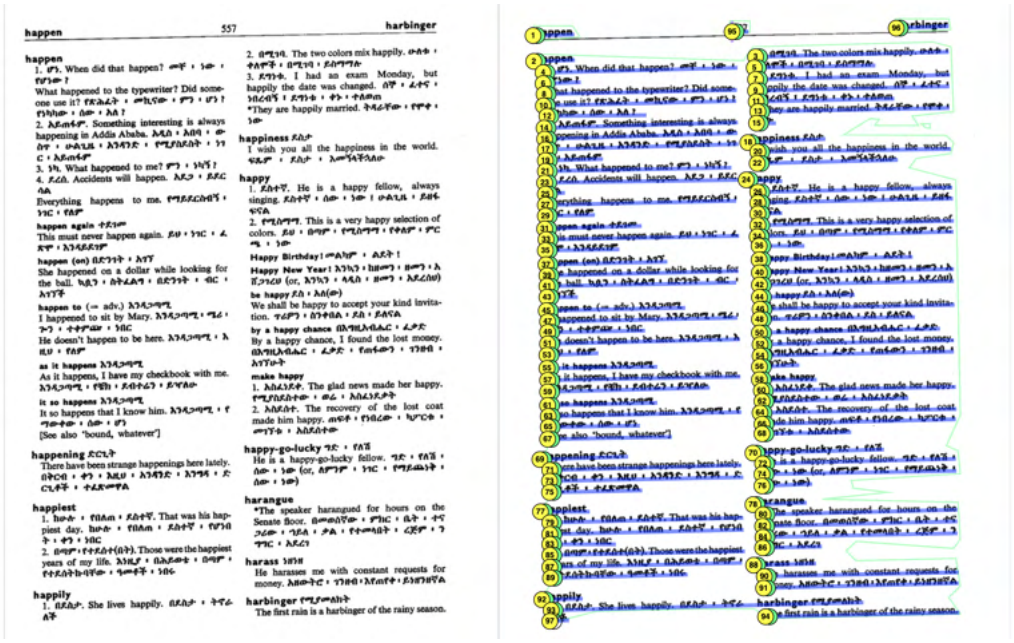


FIGURE 4.4 – Mauvaise segmentation des textes en colonnes du modèle par défaut d'eScriptorium

Les images brutes durent dès lors être manuellement segmentées en zones de textes afin d'entraîner un modèle de segmentation des pages adéquat.

Deux types de zones de textes (colonnes et en-tête) furent ainsi définies dans l'ontologie prévue à cet effet de l'interface d'eScriptorium (cf. figure 4.5).

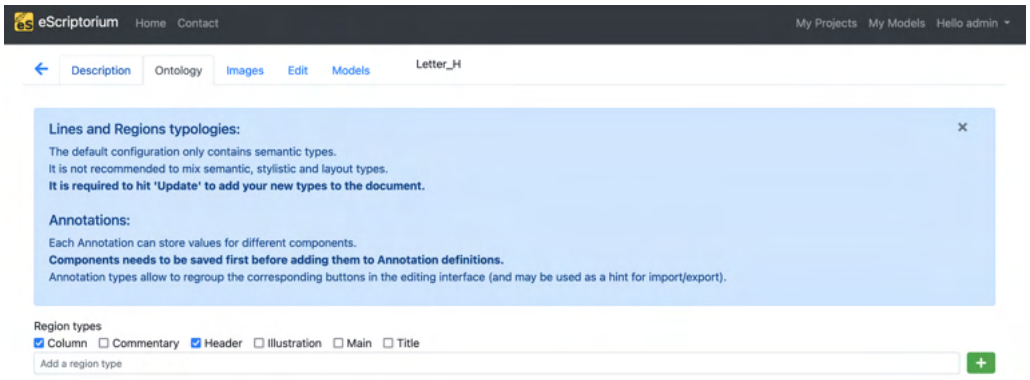


FIGURE 4.5 – Interface pour la définition d'une ontologie des zones de textes d'eScriptorium

Les zones de texte ainsi définies furent ensuite tracées manuellement à même les pages (cf. figure 4.6).

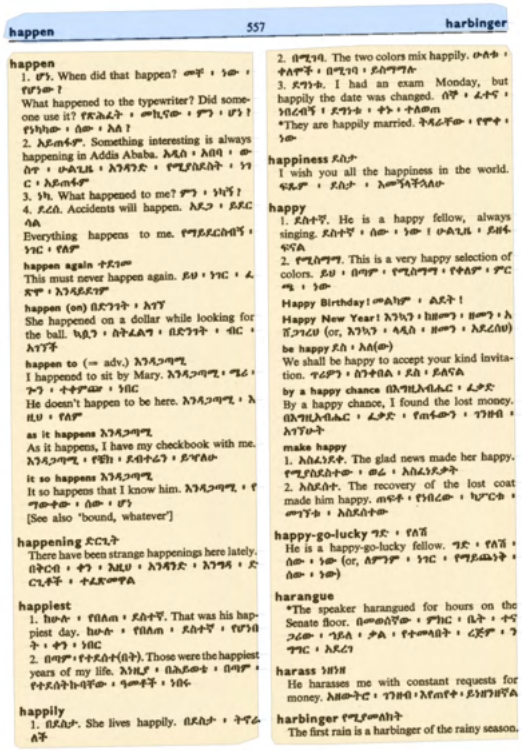


FIGURE 4.6 – Image PNG segmentée en zones de textes d'une page de dictionnaire (jaune = colonnes, bleu = en-tête)

Ces zones de textes furent par la suite segmentées en lignes, selon des modalités qui devaient faciliter l'extraction des segments d'exemples du dictionnaire. Ces segments d'exemples et leurs traductions, en effet, sont susceptibles – du fait de la mise en page de ce dictionnaire – de s'étendre sur de multiples lignes, de multiples colonnes, voire de multiples pages. Une ontologie des lignes fut donc créée visant à permettre la reconstitution, *a posteriori*, des entrées et des sous-entrées de ce dictionnaire, en dépit desdits sauts de lignes, de colonne ou de page. Outre les lignes par défaut, trois types de lignes furent définis dans cette optique : les lignes marquant les débuts d'entrée, consistant généralement en un lemme anglais en fonte grasse, les lignes marquant les débuts de sous-entrées consistant en un lemme anglais en fonte grasse mis en retrait par un alinéa, et les lignes marquant le début des sous-entrées numérotées de la microstructure du dictionnaire, permettant la gestion de l'anisomorphie sémantique de l'anglais et de l'amharique.

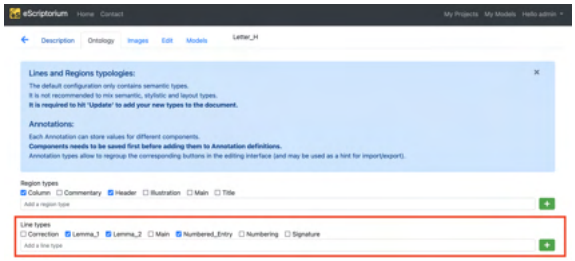


FIGURE 4.7 – Interface pour la définition d'une ontologie des lignes de textes d'eScriptorium (cadre rouge)

Cinquante et quelques pages de ce dictionnaire furent ainsi segmentées manuellement, en zones de textes et en lignes de divers types, pour fournir les données nécessaires à l'entraînement d'un modèle de segmentation automatique.

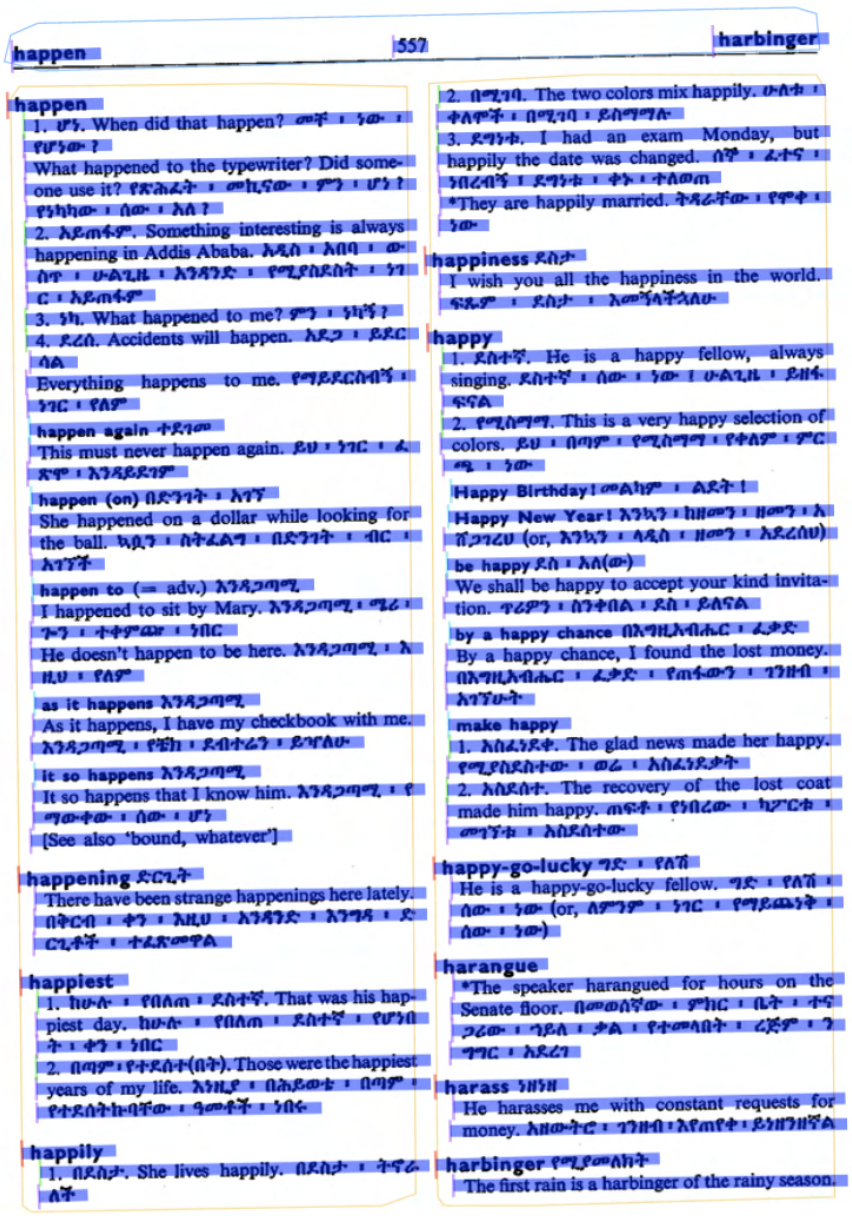


FIGURE 4.8 – Image PNG segmentée en zones de textes et en lignes d'une page de dictionnaire, les rectangles de couleur à la gauche des lignes marquent leur type

L'utilisateur disposant localement d'un processeur graphique ("*GPU*") se voit offrir par l'interface d'eScriptorium la possibilité d'entraîner un tel modèle de segmentation au sein même de cette application.

L'utilisateur dépourvu d'un processeur graphique adéquat — tel l'auteur de ces lignes — pourra toutefois mettre à profit la possibilité qu'offre eScriptorium d'importer et d'exporter des modèles de segmentation et de reconnaissance optique des caractères pour les entraîner sur un serveur distant au travers d'un notebook de

Google Colab. Une telle opération nécessite d'exporter les pages segmentées en zones de textes (fichiers PNG et documents XML Alto [11]) pour téléverser ces données sur un "drive" Google auquel il conviendra de se connecter dans l'interface des notebooks de Google Colab, puis d'installer Kraken – le module OCR sur lequel se base eScriptorium – au moyen d'une commande shell. On notera que l'environnement des notebooks Jupyter de Google Colab étant mis à jour de manière intempestive, il est fréquemment nécessaire de forcer l'installation de version anciennes de certaines bibliothèques Python 3.

```
# Connexion au drive personnel de l'utilisateur sur lequel ont été téléversées les données nécessaires à l'entraînement d'un modèle de segmentation dans un répertoire appelé "trainData"; un répertoire vide appelé "Segmenters" devra également avoir été créé :
from google.colab import drive
drive.mount('/content/drive')

# Installation du module requis :
!pip3 install kraken

# Installation forcée de modules dépréciés (nécessaire à l'été 2022, mais inutile à l'automne de cette même année), indiquée à titre d'exemple :
#!pip3 install torchtext==0.10.0
#!pip3 install torchvision==0.9.1
#!pip3 install protobuf==3.9.2

# Options de la commande segtrain du module Ketos, module d'entraînement de la librairie Kraken :
!ketos segtrain --help

# Entraînement d'un modèle de segmentation des pages ad hoc :

!ketos segtrain -d cuda:0 --augment --lag 100 -q early -o /content/drive/MyDrive/Segmenters/segModel_ -f alto /content/drive/MyDrive/trainData/*.xml
```

Code 4.1 – Code de cellule d'un notebook Google Colab pour l'entraînement d'un modèle Kraken de segmentation des pages

Les modèles de segmentation (dont la structure par défaut est donnée par la figure 4.10) entraînés successivement sur des ensembles de données progressivement étendus atteignirent des scores d'exactitude ("*accuracy*") [12] allant de 75% à 90%, des performances remarquables compte tenu de la complexité de la tâche requise (identification de colonnes très rapprochées et de multiples lignes particulières), et permirent de segmenter automatiquement, une fois importés dans eScriptorium, les pages du dictionnaire, nécessitant, toutefois, le contrôle minutieux des résultats de cette segmentation comme préalable à l'opération de reconnaissance optique des caractères.

11. Cf. http://www.loc.gov/standards/alto/.

12. L'exactitude de la segmentation des zones de textes et des lignes semble être mesurée par le module d'entraînement Ketos sur la base d'un jeu de données de validation représentant par défaut 10% des pages manuellement segmentées, et repose sur le ratio des pixels correctement classés par le segmenteur — dans des classes par ailleurs multiples (zone de texte, ligne ou interligne, ligne spéciale, sens de la ligne, etc.) — sur l'ensemble des pixels de l'image.

	Name	Type	Params
0	net	MultiParamSequential	1.3 M
1	net.C_0	ActConv2D	9.5 K
2	net.Gn_1	GroupNorm	128
3	net.C_2	ActConv2D	73.9 K
4	net.Gn_3	GroupNorm	256
5	net.C_4	ActConv2D	147 K
6	net.Gn_5	GroupNorm	256
7	net.C_6	ActConv2D	295 K
8	net.Gn_7	GroupNorm	512
9	net.C_8	ActConv2D	590 K
10	net.Gn_9	GroupNorm	512
11	net.L_10	TransposedSummarizingRNN	74.2 K
12	net.L_11	TransposedSummarizingRNN	25.1 K
13	net.C_12	ActConv2D	2.1 K
14	net.Gn_13	GroupNorm	64
15	net.L_14	TransposedSummarizingRNN	16.9 K
16	net.L_15	TransposedSummarizingRNN	25.1 K
17	net.l_16	ActConv2D	520

Trainable params: 1.3 M
Non-trainable params: 0
Total params: 1.3 M
Total estimated model params size (MB): 5

FIGURE 4.9 – Structure par défaut du modèle de segmentation Kraken.

4.2.2 Reconnaissance optique des caractères des lignes

Une fois ces pages segmentées en zones de textes et en lignes, il fut possible de transcrire, manuellement, ces lignes au moyen de l'interface d'eScriptorium, remarquablement ergonomique.

FIGURE 4.10 – Interface d'eScriptorium pour l'entrée manuelle des transcriptions

Au fur et à mesure de la transcription de ces données, des modèles de reconnaissance optique des caractères furent entraînés selon des modalités presque identiques à celles employées pour la segmentation des lignes et des pages. Dans le cas du modèle OCR, néanmoins le modèle par défaut de la bibliothèque Kraken étant apparu insuffisant, la structure du modèle de reconnaissance optique (cf. figure 4.11) fut redéfinie au moyen du langage de spécification des graphs VGSL [13] (Variable-size Graph Specification Language) au sein même de la commande Shell du sous-module Ketos de cette bibliothèque [14].

```
1 # Connexion au drive personnel de l'utilisateur sur lequel auront été téléversé
    es les données nécessaires à l'entraînement d'un modèle de segmentation
```

13. Cf. https://tesseract-ocr.github.io/tessdoc/tess4/VGSLSpecs.html
14. Cf. https://kraken.re/master/ketos.html

```
    dans répertoire appelé "trainData"; un répertoire vide appelé "OCRs" devra
    également avoir été créé :
from google.colab import drive
drive.mount('/content/drive')

# Installation du module requis :
!pip3 install kraken

# Installation forcée de modules dépréciés (nécessaire à l'été 2022, mais
    inutile à l'automne de cette même année), indiquée à titre d'exemple :
#!pip3 install torchtext==0.10.0
#!pip3 install torchvision==0.9.1
#!pip3 install protobuf==3.9.2

# Options de la commande train du module Ketos, module d'entraînement de la
    librairie Kraken :
!ketos train --help

# Entraînement d'un modèle de reconnaissance optique des caractères :
!ketos train -d cuda:0 -q early -w 0 -s '[1,120,0,1 Cr3,13,32 Do0.1,2 Mp2,2 Cr3
    ,13,32 Do0.1,2 Mp2,2 Cr3,9,64 Do0.1,2 Mp2,2 Cr3,9,64 Do0.1,2 S1(1x0)1,3
    Lbx200 Do0.1,2 Lbx200 Do.1,2 Lbx200 Do]' --augment --lag 5 --min-delta
    0.001 -N 250 -r 0.0001 -o /content/drive/MyDrive/OCRs/ocrModel_ -f alto /
    content/drive/MyDrive/trainData/*.xml
```

Code 4.2 – Code de cellule d'un notebook Google Colab pour l'entraînement d'un modèle de reconnaissance optique des caractères

	Name	Type	Params
0	net	MultiParamSequential	4.2 M
1	net.C_0	ActConv2D	1.3 K
2	net.Do_1	Dropout	0
3	net.Mp_2	MaxPool	0
4	net.C_3	ActConv2D	40.0 K
5	net.Do_4	Dropout	0
6	net.Mp_5	MaxPool	0
7	net.C_6	ActConv2D	55.4 K
8	net.Do_7	Dropout	0
9	net.Mp_8	MaxPool	0
10	net.C_9	ActConv2D	110 K
11	net.Do_10	Dropout	0
12	net.S_11	Reshape	0
13	net.L_12	TransposedSummarizingRNN	1.9 M
14	net.Do_13	Dropout	0
15	net.L_14	TransposedSummarizingRNN	963 K
16	net.Do_15	Dropout	0
17	net.L_16	TransposedSummarizingRNN	963 K
18	net.Do_17	Dropout	0
19	net.O_18	LinSoftmax	174 K

Trainable params: 4.2 M
Non-trainable params: 0
Total params: 4.2 M
Total estimated model params size (MB): 16

FIGURE 4.11 – Structure du modèle OCR *ad hoc*

On notera que certains des caractères de l'alphasyllabaire éthiopien étant d'occurrence rare en amharique, des données synthétiques furent créées, manuellement, au moyen de l'éditeur d'images Gimp, consistant en des pages du dictionnaire original dont le texte des entrées fut remplacé par des séquences aléatoires de caractères éthiopiens, représentés dans une police similaire à celle du document original et dont l'antialiasing avait été désactivé (cf. figure 4.12).

FIGURE 4.12 – Données synthétiques produites manuellement au moyen d'un éditeur graphique sur la base de séquences de caractères amhariques générées aléatoirement

En effet, si la correction d'erreurs pour la partie anglaise de ces données textuelles devait être fort aisée (le simple correcteur orthographique du navigateur Web dans lequel est déployée l'interface d'eScriptorium suffisant, par exemple, à identifier les erreurs les plus grossières), l'absence de correcteurs orthographiques fiables pour l'amharique rendait nécessaire cette opération de génération de données de synthèse, afin de garantir le traitement correct par l'OCR des caractères infréquents.

Les performances obtenues par les modèles ainsi entraînés sur la base du critère de l'exactitude ("*accuracy*")[15] furent remarquables, partant de 64.8% pour atteindre 98.1% après un mois de transcriptions/corrections manuelles, et ce en dépit de la mauvaise qualité des images des données de départ, du caractère multilingue, multiscript (et multifonte et multipolice dans le cas de l'alphabet latin) de ce document. Importé dans l'interface d'eScriptorium ces modèles permirent la transcription de l'intégralité des quelques 1500 pages de l'English-Amharic Context Dictionary en seulement trois mois d'un intense travail, mais sans connaissance préalable de cet outil technique.

15. L'exactitude de l'OCR est mesurée par le module d'entraînement Ketos sur la base d'un jeu de données de validation représentant par défaut 10% des lignes segmentées, et met en jeu, à la manière du calcul d'une distance de Levenshtein, le dénombrement des insertions, suppressions ("*deletions*") et substitutions séparant les chaînes de caractères inférées par l'OCR de la vérité terrain manuellement transcrite, le nombre de ces erreurs étant soustrait au nombre des caractères inférés puis divisé par cette même variable.

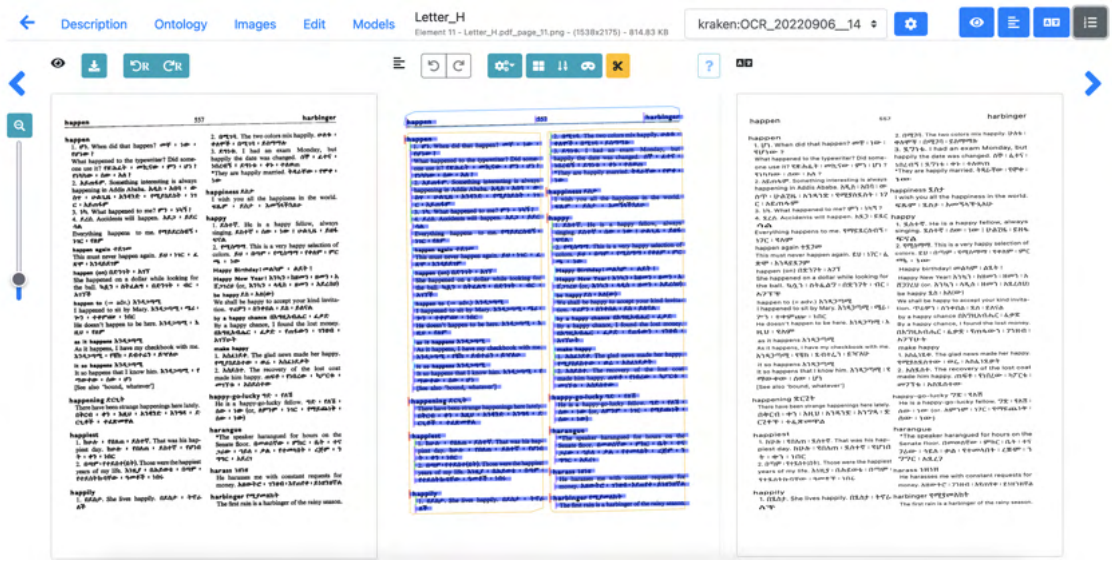


FIGURE 4.13 – Interface d'eScriptorium avec : l'image PNG brute d'une page de dictionnaire (à la gauche de l'écran), sa segmentation en zones de textes et en lignes (au centre) et sa transcription manuelle/automatique (à la droite de l'écran)

4.3 Extraction d'un corpus de segments alignés

Une fois achevée la segmentation et la reconnaissance optique des caractères des images des pages du dictionnaire servant de données fondamentales à cette expérience, les fichiers XML Alto furent exportées (pour le seul texte des colonnes de ce document, cf. figure 4.14) afin de faire l'objet d'un traitement visant à en extraire un corpus parallèle de segments alignés amharique-anglais.

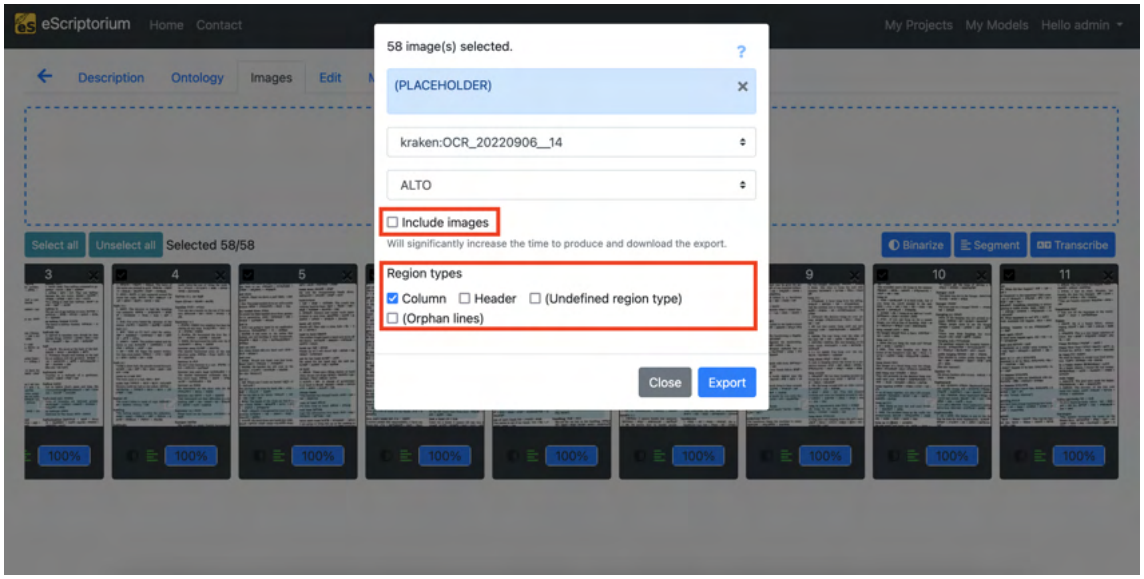


FIGURE 4.14 – Interface d'eScriptorium pour l'exportation des données de colonnes transcrites au format XML Alto

FIGURE 4.15 – Aperçu des données transcrites au format XML Alto

Pour une lettre donnée du dictionnaire, l'intégralité de ces fichiers XML Alto – un par page (la figure 4.15 en donne un aperçu) – fut concaténée dans l'ordre de ces pages et le contenu des lignes de textes assemblé – sur la base des balises résultant de l'action du segmenteur précédemment entraîné – en entrées et sous-entrées distinctes, tout en remplaçant les sauts de lignes, de colonnes ou de pages par le tiret bas ("_").

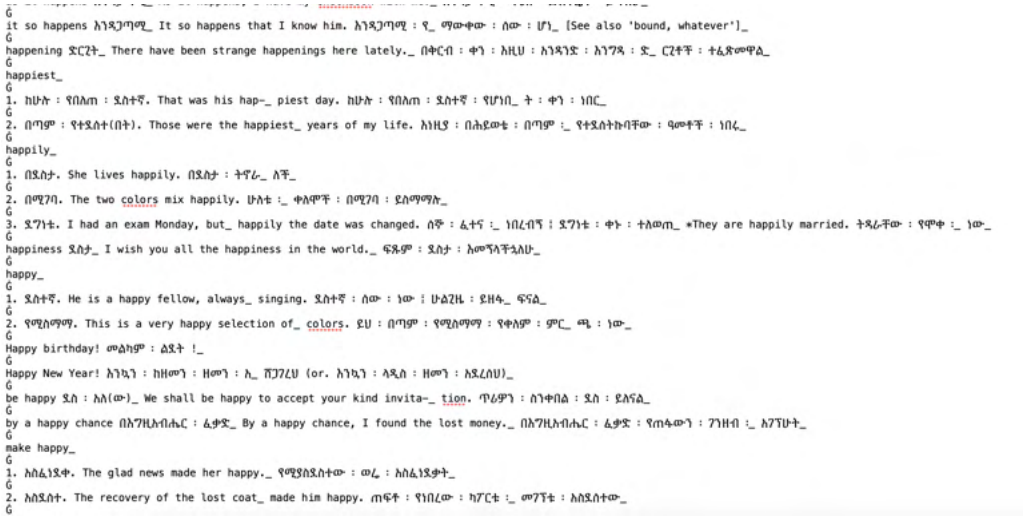


FIGURE 4.16 – Reconstitution des entrées et des sous-entrées du dictionnaire en dépit des sauts de ligne, de colonnes et de pages, au départ d'un document XML

Le document ainsi obtenu (cf. figure 4.16)fut ensuite importé dans une application source-ouverte pour l'annotation humaine des données appelée Doccano [16].

FIGURE 4.17 – Logo de l'application Doccano

16. https://github.com/doccano/doccano

Plusieurs milliers des entrées et sous-entrées reconstituées issues de la transcription du dictionnaire furent alors annotées manuellement afin d'y marquer les exemples anglais et leur traduction en amharique (cf. figures 4.18 et 4.19).

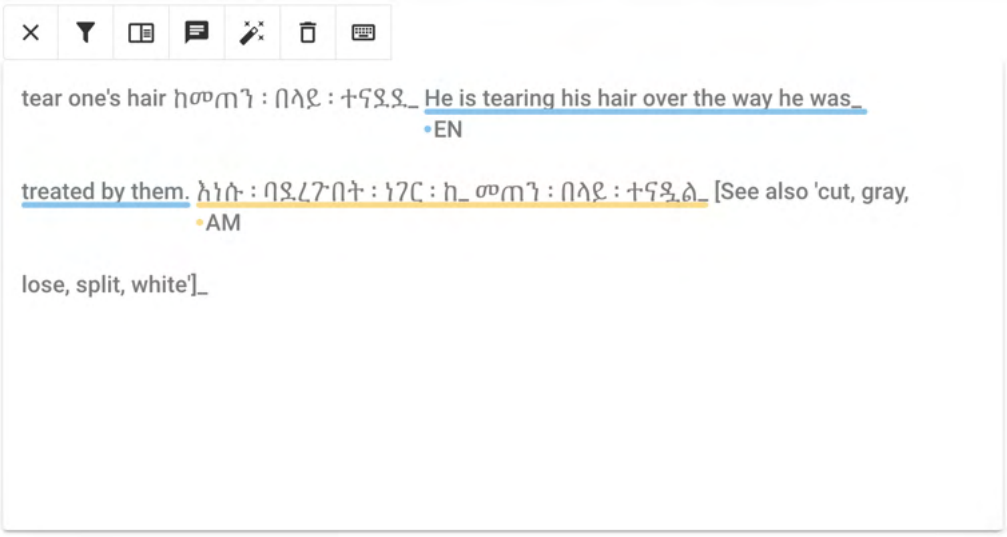


FIGURE 4.18 – Premier exemple d'annotation des entrées et sous-entrées au moyen de l'interface de l'application Doccano

FIGURE 4.19 – Second exemple d'annotation des entrées et sous-entrées au moyen de l'interface de l'application Doccano

Un fichier JSON (cf. figure 4.20) contenant le texte de ces entrées et les offsets de ces annotations fut ensuite exporté depuis cette application.

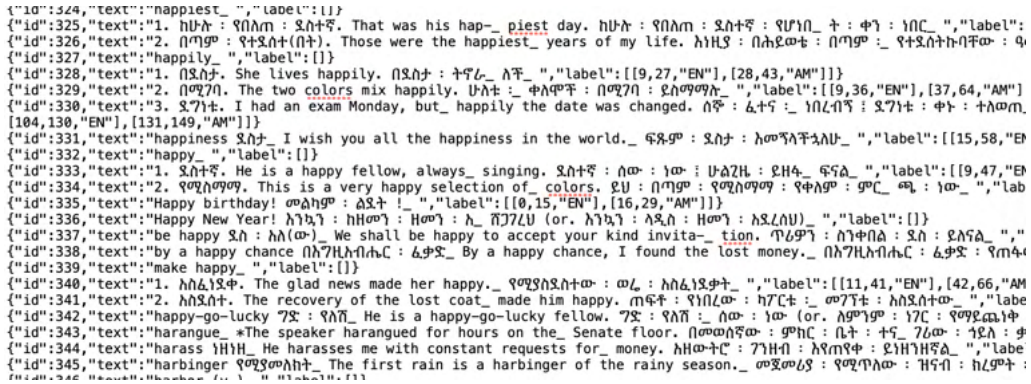


FIGURE 4.20 – Aperçu d'un fichier JSON d'entrées annotées exporté depuis Doccano

Une opération parfaitement triviale mais laborieuse de conversion de ces offsets au format BIO fut ensuite entreprise, dans le but d'entraîner un modèle Bi-LSTM pour l'étiquetage automatique du reste de ces entrées. Le document à traiter étant un dictionnaire, on a jugé que les mots graphiques contenus dans ces entrées feraient de mauvais tokens dans la perspective d'une pareille tâche. Les documents issus de la lexicographie bilingue présentant des régularités formelles remarquables, le choix fut fait de mettre ces propriétés à profit pour extraire d'un tel document l'ensemble des segments d'exemples anglais et leurs traduction amhariques, et un formalisme élémentaire fut mis au point consistant à réduire les mots graphiques des entrées aux seules informations portant sur la nature des caractères les composant. Cet algorithme devait consister, simplement, à substituer la lettre "e" à chacun des caractères de l'alphabet latin des entrées annotées (et la lettre "E" à chacune des majuscules de l'alphabet latin) et la lettre "a" à chacun des caractères du syllabaire éthiopien, les autres caractères restant inchangés, avant de réduire chacun des mots graphiques des segments de textes à traiter en un ensemble (au sens mathématique de terme, en y supprimant les caractères redondants) reconverti, ensuite, en une chaîne de caractères ordonnés dans un ordre fixe. La table 4.1 donne un exemple d'application de ce formalisme.

1.	በደስታ.	She	lives	happily.	በደስታ	:	ትኖራ_	ለች_
1.	aaaa.	Eee	eeeee	eeeeeee.	aaaa	:	aaa_	aa_
1.	a.	Ee	e	e.	a	:	a_	a_
O	O	B-E	I-E	I-E	B-A	I-A	I-A	I-A

TABLE 4.1 – Illustration du formalisme visant à l'abstraction du texte des entrées en vue de l'entraînement d'un modèle d'annotation automatique. De haut en bas : le texte de départ de la sous-entrée, la première phase de l'algorithme, la seconde phase (finale) de l'algorithme, les étiquettes au format BIO

Un fichier CSV (cf. figure 4.21) contenant ces données ainsi transformées devait permettre l'entraînement d'un modèle d'annotation automatique reposant sur la bibliothèque Python AllenNLP et inspiré du code offert par M. Pierre Magistry dans le cadre de son cours "Méthodes en apprentissage automatique" dispensé au second semestre de l'année 2021/2022 et des codes présents dans l'ouvrage [Hagiwara, 2021] [17].

17. Une version de ce code est consultable sur la page Github de cet auteur.

	ነበሩ_	a_	I-A
			B-
Sentence: 326	happily_	e_	O
			B-
Sentence: 327	1.	1.	O
	በደስታ.	a.	O
	She	Ee	B-E
	lives	e	I-E
	happily.	e.	I-E
	በደስታ	a	B-A
	:	:	I-A
	ትኖራ_	a_	I-A
	ለች_	a_	I-A
			B-
Sentence: 328	2.	2.	O
	በሚገባ.	a.	O
	The	Ee	B-E
	two	e	I-E
	colors	e	I-E
	mix	e	I-E
	happily.	e.	I-E
	ሁለቱ	a	B-A
	:_	:_	I-A
	ቀለሞች	a	I-A
	:	:	I-A
	በሚገባ	a	I-A
	:	:	I-A
	ይስማማሉ_	a_	I-A
			B-
Sentence: 329	3.	3.	O
	ያግኑቱ.	a.	O

FIGURE 4.21 – Aperçu d'un fichier CSV pour l'entraînement d'un modèle Bi-LSTM d'annotation automatique

Les performances du modèle d'annotation automatique obtenu furent excellentes (cf. figure 4.22) et permirent l'annotation de l'ensemble des entrées et sous-entrées du dictionnaire, l'alignement des segments découlant, de façon triviale, de ce bon étiquetage : les exemples anglais et leurs traductions amhariques dûment étiquetés alternant parfaitement.


```
{'best_epoch': 2,
 'peak_worker_0_memory_MB': 1203.13671875,
 'peak_gpu_0_memory_MB': 0,
 'training_duration': '0:36:29.717288',
 'epoch': 12,
 'training_accuracy': 0.9998782964059008,
 'training_prec': 0.9995281535073922,
 'training_rec': 0.9996391361474545,
 'training_f1': 0.9995836417468041,
 'training_loss': 0.0003580130825087093,
 'training_worker_0_memory_MB': 1203.13671875,
 'training_gpu_0_memory_MB': 0.0,
 'validation_accuracy': 0.9996588837264012,
 'validation_prec': 0.998725356900068,
 'validation_rec': 0.9993197857325058,
 'validation_f1': 0.9990224828934008,
 'validation_loss': 0.001978104452497252,
 'best_validation_accuracy': 0.9997412221372699,
 'best_validation_prec': 0.998980198861222,
 'best_validation_rec': 0.9994898392993793,
 'best_validation_f1': 0.9992349540971959,
 'best_validation_loss': 0.001096752067609246}
```


FIGURE 4.22 – "Stats" d'un modèle d'annotation automatique des exemples anglais et de leurs traductions amhariques au sein des entrées du dictionnaire transcrit

Au terme de la seconde étape de ce protocole expérimental, 45 047 paires de segments anglais-amharique – mis en forme au moyen d'expressions régulières pour

effacer les marques de sauts de lignes et les séparateurs de mots amhariques désuets – purent ainsi être extraites des données initiales. Les figures 4.23 et 4.24 donnent un aperçu de ce corpus.

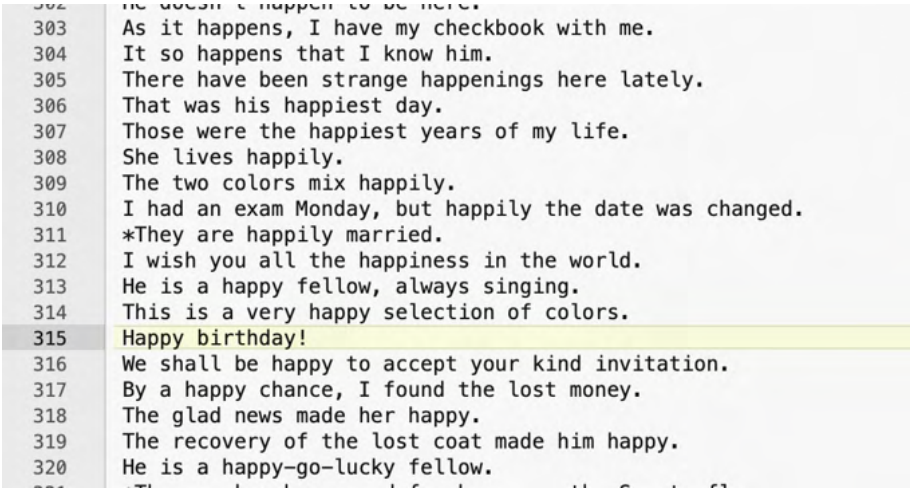


FIGURE 4.23 – Aperçu des segments anglais extraits d'un ouvrage relevant de la lexicographie bilingue

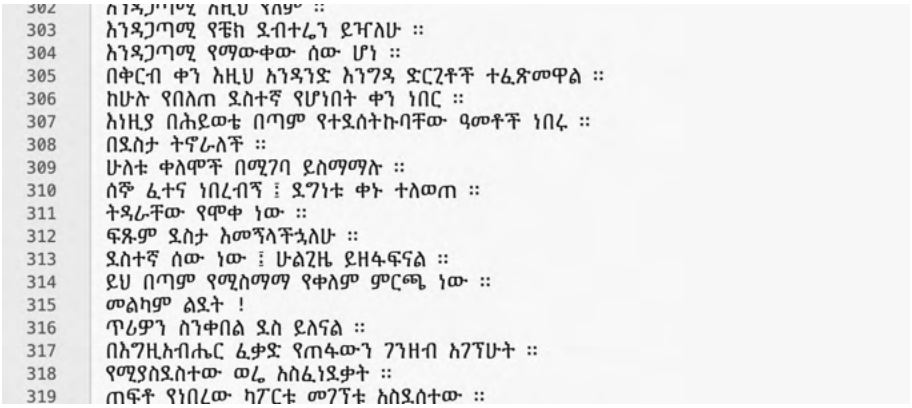


FIGURE 4.24 – Aperçu des segments amhariques extraits d'un ouvrage relevant de la lexicographie bilingue

4.4 Entraînement de modèles et augmentation des données par rétrotraduction

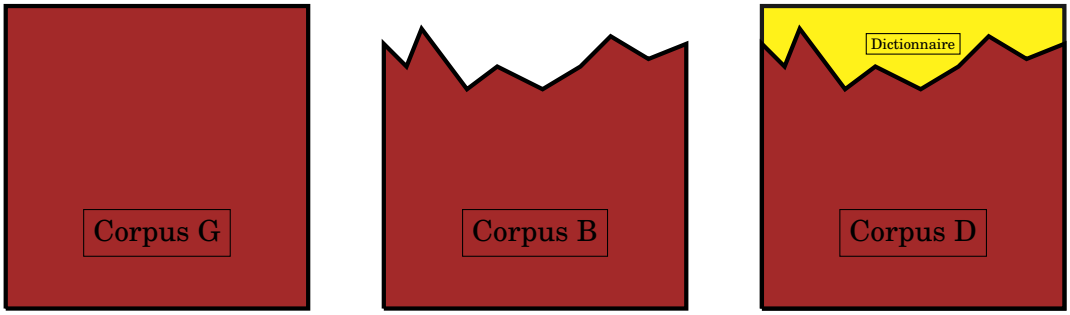


FIGURE 4.25 – Trois corpus d'entraînement

Afin d'évaluer les propriétés des données parallèles extraites au départ des seules images d'un document relevant de la lexicographie bilingue pour la paire de langue amharique-anglais, une expérience fut également mise au point qui mit en jeux trois corpus parallèles devant servir à l'entraînement de trois modèles de TA comme tâche distincts. Le premier de ces corpus — appelé "Corpus G" dans le restant de cette étude — est le corpus d'entraînement compilé par [Gezmu et al., 2022][18], long de 140 000 segments et comportant plus de 2 millions de mots anglais et amhariques. Le second corpus — appelé "Corpus B" dans le restant de cette étude et devant servir à l'entraînement d'une "baseline" — est une version du Corpus G aléatoirement amputé de segments représentant un nombre de tokens équivalent à celui des données parallèles précédemment extraites de l'English-Amharic Context Dictionary ; ce Corpus B comprend 120 368 segments, pour 2 204 699 mots graphiques anglais et 1 855 185 mots graphiques amhariques. Le troisième corpus — baptisé "Corpus D" — résulte de l'adjonction au Corpus B des segments précédemment issus du dictionnaire bilingue et comporte 165 415 segments [19] représentant 2 567 160 mots graphiques anglais et 2 205 325 mots graphiques amhariques [20].

Corpus	Segments	Mots graphiques anglais	Mots graphiques amhariques
Corpus G	140 000	2 566 325	2 158 899
Corpus B	120 368	2 204 699	1 855 185
Corpus D	165 415	2 567 160	2 202 325

TABLE 4.2 – Nombre des segments et des mots graphiques compris dans chacun des corpus constitués pour l'entraînement de modèles de TA distincts

Des corpus parallèles de validation et d'évaluation furent également constitués par la concaténation des jeux de données de validation et d'évaluation de

18. Corpus décrit p. 46

19. On notera que les segments issus du dictionnaire sont nettement plus brefs, en moyenne, que ceux du corpus de [Gezmu et al., 2022].

20. On notera que les nombres de mots ici rapportés sont calculés de deux manières distinctes pour l'anglais et l'amharique, le décompte des mots anglais ayant été opéré au moyen du tokeniseur de la bibliothèque NLTK, et celui des mots amhariques par l'application wc du shell Linux. La tokenisation sur laquelle s'est basé le découpage de ces corpus est celle des mots anglais.

[Gezmu et al., 2022] aux jeux de données de validation et d'évaluation pour les langues amharique et anglaise issus du corpus FLORES-101 [21] [22].

Corpus	Segments	Mots graphiques anglais	Mots graphiques amhariques
Corpus de validation issu de Gezmu et al.	2 864	45 827	44 666
Corpus de validation issu de FLORES-101	997	20 954	21 624
Total validation :	3 861	66 781	66 290
Corpus d'évaluation issu de Gezmu et al.	2 500	39 308	38 683
Corpus d'évaluation issu de FLORES-101	1 012	21 901	22 220
Total évaluation :	3 512	61 209	60 903

TABLE 4.3 – Nombre des segments et des mots graphiques compris dans chacun des corpus de validation et d'évaluation accompagné du détail de leurs sources.

Des modèles de TA comme tâche anglais-amharique et amharique-anglais furent donc entraînés au départ de chacun de ces corpus d'entraînement G, B et D, et validés sur la base du corpus de validation résultant de la concaténation des données *ad hoc* de [Gezmu et al., 2022] et de FLORES-101, au moyen de la suite d'outils Fairseq ([Ott et al., 2019]) déployée sur des notebooks Google Colab pour disposer d'un GPU.

FIGURE 4.26 – Logo de Fairseq

En manière de prétraitement, chacun de ces corpus fut tokenisé selon la méthode BPE-dropout précédemment évoqué, avec un vocabulaire de 8000 tokens ; les données amhariques ayant été préalablement transcrites en alphabet latin au moyen de fonctions développées à cet effet par Andargachew Mekonnen Gezmu [23].

```
# Installation de la librairie SentencePiece :
!pip3 install sentencepiece

import sentencepiece as spm
```

21. Cf. https://github.com/facebookresearch/flores/tree/main/previous_releases/flores101

22. On notera que l'ambition première de cette étude était de garder distincts les jeux de données provenant de ces deux sources, mais les moyens techniques limités dont nous disposions ainsi que des contraintes de temps n'ont pas permis de conduire ces multiples expériences, et la solution consistant à rassembler ces jeux de données en un seul nous est ainsi apparue comme un pis-aller acceptable, compte tenu de ces circonstances.

23. Cf. https://github.com/andmek/AT4MT

```
5
6  # Phase d'entraînement des tokeniseurs du Corpus B :
7  path_en = "/content/drive/Corpus_TA/Corpus_B.en"
8  path_am = "/content/drive/Corpus_TA/Corpus_B.am"
9
10 spm.SentencePieceTrainer.train(f'--input={path_en} --model_prefix=en_baseline
       --character_coverage=1.0 --vocab_size=8000 --model_type=bpe')
11 sp_baseline_en = spm.SentencePieceProcessor()
12 sp_baseline_en.load('en_baseline.model')
13
14
15 spm.SentencePieceTrainer.train(f'--input={path_am} --model_prefix=am_baseline
       --character_coverage=1.0 --vocab_size=8000 --model_type=bpe')
16 sp_baseline_am = spm.SentencePieceProcessor()
17 sp_baseline_am.load('am_baseline.model')
18
19 def tokenize_am(text):
20   return sp_baseline_am.encode_as_pieces(text)
21
22 def tokenize_en(text):
23   return sp_baseline_en.encode_as_pieces(text)
24
25 #Exemples de tokenisations :
26 print(tokenize_en("there are still others that are sown among the thorns."))
27 #['▁there', '▁are', '▁still', '▁others', '▁that', '▁are', '▁s', '
       own', '▁among', '▁the', '▁thorns', '.']
28 print(tokenize_am("b ziyaw gize akababi w lajoce w d bet ndmls
         yqu ."))
29 #['▁bziyaw', '▁gize', '▁akababi', '▁wlajoce', '▁wd', '▁bet
       ', '▁nd', '▁m l s', '▁y', '▁qu', '▁', '.']
```

Code 4.3 – Exemple de code pour la tokenisation des corpus anglais et amharique

Les segments parallèles des corpus d'entraînement, de validation et de test dûment tokenisés par BPE furent ensuite passés au module de "pré-processing" de Fairseq :

```
1  # Installation de la librairie Fairseq :
2  !pip3 install fairseq
3
4  # Création d'un répertoire pour l'archivage des données du modèle de TA
       baseline :
5  !mkdir -p /content/data/mt-bin_baseline
6
7  # "Pré-processing" :
8  !fairseq-preprocess \
9        --source-lang am \
10       --target-lang en \
11       --bpe sentencepiece \
12       --trainpref /content/baseline_train_bpe \
13       --validpref /content/baseline_dev_bpe \
14       --testpref /content/baseline_devtest_bpe \
15       --destdir /content/data/mt-bin_baseline \
16       --thresholdsrc 2 \
17       --thresholdtgt 2
```

Code 4.4 – Exemple de pré-processing Fairseq

Des modèles Transformer de TA comme tâche furent subséquemment entraînés sur la base d'hyperparamètres inspirés de [Àlex R. Atrio and Popescu-Belis, 2022] et de [Hagiwara, 2021] :

```
#Entraînement d'un modèle Transformer de "baseline" :
!fairseq-train \
    /content/data/mt-bin_baseline \
    --arch transformer \
    --share-decoder-input-output-embed \
    --max-epoch 500 --patience 50 \
    --optimizer adam --adam-betas '(0.9, 0.98)' \
    --clip-norm 0.0 --lr 5e-4 --lr-scheduler inverse_sqrt \
    --warmup-updates 32000 --dropout 0.3 \
    --criterion label_smoothed_cross_entropy \
    --label-smoothing 0.1 \
    --max-tokens 4096 \
    --no-epoch-checkpoints \
    --save-dir /content/data/mt-ckpt-transformer_baseline
```

Code 4.5 – Exemple d'entraînement d'un Transformer Fairseq

Chacun des modèles Transformer les plus performants en validation, entraînés sur la base des corpus G, B et D éponymes, fut employé pour générer une série d'hypothèses de traductions au départ du corpus d'évaluation dûment tokenisé [24] :

```
# Inférences sur les données d'évaluation :
!fairseq-interactive /content/data/mt-bin_baseline --input=/content/
    baseline_devtest_bpe.am --path /content/data/mt-ckpt-transformer_baseline/
    checkpoint_best.pt --beam 5 --source-lang am --target-lang en | grep -P "D
    -[0-9]+" | cut -f3 > /content/baseline_inference_bpe.txt
```

Code 4.6 – Exemple de génération d'inférences d'un modèle entraîné Fairseq

Ces traductions candidates furent subséquemment détokenisées et évaluées sur la base principalement du score BLEU et de l'algorithme COMET :

```
# Installation des modules requis :
!pip3 install sacrebleu==2.0.0
!pip3 install unbabel-comet==1.1.1

# Sacrebleu :
!sacrebleu /content/gold_standard.txt -i /content/dict_inference.txt -m bleu
    chrf ter -l am-en

# COMET :
!comet-compare -s /content/src_fid.am -t /content/baseline_inference.txt /
    content/G_inference.txt /content/D_inference.txt -r /content/drive/MyDrive/
    AmharicEnglish_Parallel_Corpus-version_1.0/test.am-en.base.en
```

Code 4.7 – Exemple d'évaluations des modèles entraînés sur la base de la TA document

FIGURE 4.27 – Logo de COMET

Des modèles de TA anglais → amharique G°, B° et D° furent entraînés qui servirent à la génération de données augmentées par rétrotraduction. Pour ce faire, 1 million de segments fut extrait de la partie anglaise du corpus giga-fren [25] et traduit

24. Chacun des modèles ayant un tokeniseur propre.
25. Cf. https://opus.nlpl.eu/giga-fren-v2.php

vers l'amharique au moyen de ces modèles Transformers. Fairseq s'étant avéré nettement inadéquat à l'accomplissement de cette tâche, le choix fut fait de recourir à la bibliothèque Python CTranslate2 [26] afin d'optimiser ces modèles.

```
# Installation des modules requis :
!pip3 install ctranslate2 OpenNMT-py sentencepiece fairseq

# Conversion d'un modèle Transformer Fairseq =
!ct2-fairseq-converter --model_path /content/drive/MyDrive/mt-ckpt-
    transformer_baseline/checkpoint_best.pt --data_dir /content/drive/MyDrive/
    mt-bin_baseline/ --output_dir /content/ct2_model_baseline

# Import des modules requis :
import ctranslate2
from tqdm import tqdm

# Instanciation du Transformer optimisé :
translator = ctranslate2.Translator("/content/ct2_model_baseline/", device="
    cuda", inter_threads=4)

# rétrotraduction :
result_batches = []
for batch in tqdm(batches):
  results = translator.translate_batch(batch, beam_size=5, allow_early_exit=
    False, normalize_scores=True)
  result_batches.append(results)
# Notes :
### batches = liste de listes
### batch = liste de 25 segments tokenisés à traduire
```

Code 4.8 – rétrotraduction par modèle Fairseq optimisé

Trois corpus d'entraînement additionnels – G+, B+ et D+ – furent ainsi compilés par l'adjonction de ces données retraduites [27] aux corpus de départ, leur conférant les dimensions reportées table 4.4.

Corpus +	Segments	Mots graphiques anglais	Mots graphiques amhariques
Corpus G+	1 133 145	25 145 874	20 213 439
Corpus B+	1 113 532	24 839 624	18 286 345
Corpus D+	1 159 203	25 245 826	19 223 517

TABLE 4.4 – Nombre des segments et des mots graphiques compris dans chacun des corpus augmentés G+, B+ et D+

Trois nouveaux modèles Transformer de la TA comme tâche furent enfin entraînés selon le même protocole, et les performances des six modèles de traduction amharique → anglais ainsi produits, ainsi que celles des trois modèles de traduction anglais → amharique, furent évaluées et comparées, pour tenter d'apprécier l'impact des données issues de la lexicographie bilingue de cette paire de langue sur la TA comme tâche.

26. https://github.com/OpenNMT/CTranslate2

27. Chacun des segments retraduits ayant été préalablement préfixé d'un token <$\dot{B}$>, cf. [Caswell et al., 2019].

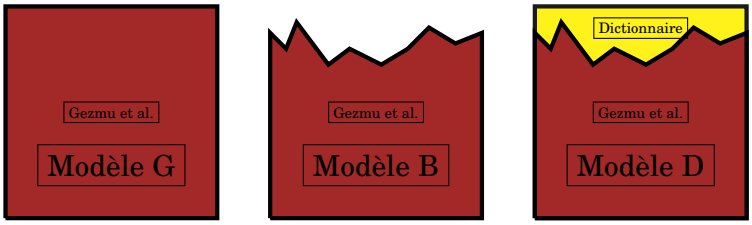


FIGURE 4.28 – Trois modèles de TA comme tâche amharique → anglais représentés et nommés en fonction de leurs données d'entraînement NON augmentées

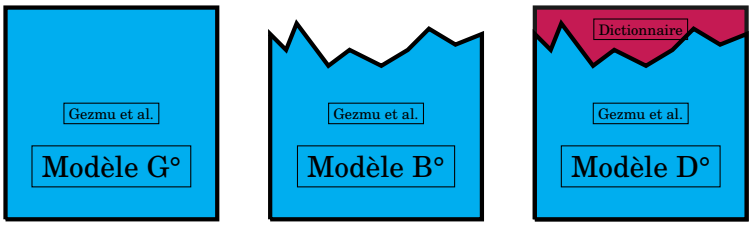


FIGURE 4.29 – Trois modèles de TA comme tâche anglais → amharique représentés et nommés en fonction de leurs données d'entraînement NON augmentées et visant à la production de données augmentées par rétrotraduction

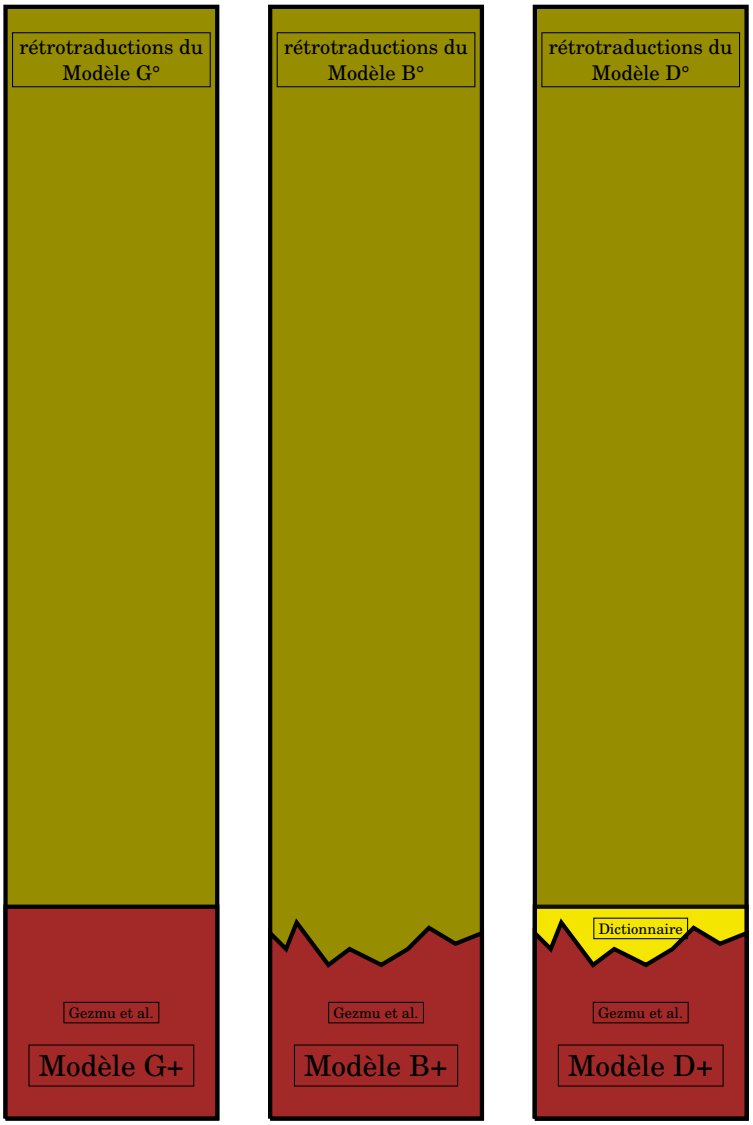


FIGURE 4.30 – Trois modèles de TA comme tâche amharique → anglais représentés et nommés en fonction de leurs données d'entraînement augmentées

CHAPITRE 5

RÉSULTATS ET DISCUSSION

Sommaire

5.1 Introduction **85**

5.2 Résultats **86**

5.2.1 Évaluation automatique des modèles amharique → anglais entraînés sur des données simples 86

5.2.2 Évaluation automatique des modèles anglais → amharique entraînés sur des données simples 87

5.2.3 Évaluation automatique des modèles amharique → anglais entraînés sur des données augmentées par rétrotraduction . 88

5.3 Discussion **89**

5.1 Introduction

Le protocole en forme de chaîne opératoire présenté au fil du précédent chapitre a permis l'extraction – au départ des seules images des pages d'un dictionnaire – d'un corpus de quelques 45 000 paires de segments parallèles pour l'anglais et l'amharique, langue peu dotée, et l'entraînement de neuf modèles Transformer de TA comme tâche visant à permettre d'évaluer l'impact de ces données issues de la lexicographie bilingue d'une langue orientale sur les performances de tels systèmes. Le présent chapitre présente et analyse les résultats de ces évaluations, basées principalement sur les métriques automatiques BLEU et COMET, les scores chrF ([Popović, 2015]) et le taux d'erreur TER (cf. [Koehn, 2020], p. 36) étant également rapportés à titre indicatif.

5.2 Résultats

5.2.1 Évaluation automatique des modèles amharique → anglais entraînés sur des données simples

Modèles AM → EN	BLEU (± 95% CI)
Modèle B	20.5 (20.5 ± 0.7) (baseline)
Modèle G	21.2 (21.2 ± 0.8) (p = 0.0020)*
Modèle D	22.9 (22.8 ± 0.8) (p = 0.0010)*

TABLE 5.1 – Scores BLEU résultant de l'évaluation automatique des modèles Transformer entraînés sur des données NON augmentées dans la direction amharique → anglais ; la présence d'une astérisque à la droite des valeurs-p, présentées entre parenthèses, indique que les performances du modèle considéré diffèrent significativement de celles de la baseline (Modèle B)

Les scores BLEU, reportés table 5.1, et leurs intervalles de confiance calculés sur le base de 1000 échantillons (par la technique dite du "*paired bootstrap resampling*") pour les Modèles G, B et D de traduction amharique → anglais semblent indiquer que les données parallèles issues d'un document relevant de la lexicographie bilingue, loin de nuire aux performances de systèmes de TA comme tâche (Modèle D > Modèle B, valeur-p = 0.0010), sont susceptibles d'offrir à ces derniers un surcroît significatif de performance, et ce relativement à des données de dimensions équivalentes provenant d'autres sources et notamment du Web (Modèle D > Modèle G, valeur-p = 0.0010 [1]).

Modèles AM → EN	chrF2 (± 95% CI)	TER (± 95% CI)
Modèle B	39.1 (39.1 ± 0.6) (baseline)	76.6 (76.6 ± 0.9) (baseline)
Modèle G	40.0 (40.0 ± 0.6) (p = 0.0010)*	73.5 (73.5 ± 0.8) (p = 0.0010)*
Modèle D	41.9 (41.9 ± 0.6) (p = 0.0010)*	71.9 (71.9 ± 0.8) (p = 0.0010)*

TABLE 5.2 – Résultats d'évaluation automatique (chrF et TER) des modèles entraînés sur des données NON augmentées dans la direction amharique → anglais ; (N.B. : Le score TER est un taux d'erreur et gagne, par conséquent, à être minimisé)

Des résultats que semblent confirmer les scores chrF et le taux d'erreur TER, reportés table 5.2, ainsi que les scores d'évaluation automatique produits par l'algorithme COMET, reportés table 5.3. Ces scores COMET ont été générés par le modèle wmt20-comet-da, entraîné – sur la base des données issues de campagnes d'évaluation directe humaine antérieures de la conférence WMT pour les années 2017 à 2019 – à simuler, par régression, la cote Z d'un score décerné par un évaluateur humain. Si ces scores COMET sont peu évocateurs, on retiendra que le classement des modèles opéré par cet algorithme est en tout point conforme à celui établi sur la base des scores BLEU (Modèle G > Modèle B, p = 0.0000 ; Modèle D > Modèle B, p = 0.0000 ; Modèle D > Modèle G, p = 0.0000, selon COMET).

1. Calculée indépendamment, avec le Modèle G pour baseline :

Modèles AM → EN	BLEU (± 95% CI)	chrF2 (± 95% CI)	TER (± 95% CI)
Baseline : Modèle G	21.2 (21.2 ± 0.8)	40.0 (40.0 ± 0.6)	73.5 (73.5 ± 0.8)
Modèle B	20.5 (20.5 ± 0.7) (p = 0.0020)*	39.1 (39.1 ± 0.6) (p = 0.0010)*	76.6 (76.6 ± 0.9) (p = 0.0010)*
Modèle D	22.9 (22.8 ± 0.8) (p = 0.0010)*	41.9 (41.9 ± 0.6) (p = 0.0010)*	71.9 (71.9 ± 0.8) (p = 0.0010)*

Modèles AM → EN	COMET (wmt20-comet-da)
Modèle B	-0.0064*
Modèle G	0.0367*
Modèle D	0.1264*

TABLE 5.3 – Résultats d'évaluation COMET des modèles Transformer amharique → anglais entraînés sur des données NON augmentées

5.2.2 Évaluation automatique des modèles anglais → amharique entraînés sur des données simples

Les modèles de TA comme tâche B°, G° et D°, entraînés sur des corpus identiques à ceux des modèles B, G et D, mais dans la direction opposée anglais → amharique, devaient permettre la génération de données retraduites. La hiérarchie de ces modèles sous le rapport de leurs scores BLEU (reportés table 5.4) est inchangée vis à vis de celle des modèles amharique → anglais précédemment évalués : Modèle G° > Modèle B° (p = 0.0020) ; Modèle D° > Modèle B° (p = 0.0010) ; Modèle D° > Modèle G° (p = 0.0020 [2]).

Modèles EN → AM	BLEU (± 95% CI)
Modèle B°	13.6 (13.6 ± 0.7) (baseline)
Modèle G°	14.1 (14.0 ± 0.7) (p = 0.0020)*
Modèle D°	14.6 (14.6 ± 0.7) (p = 0.0010)*

TABLE 5.4 – Scores BLEU résultant de l'évaluation automatique des modèles anglais → amharique entraînés dans l'objectif de générer des données retraduites ; (on notera que les scores BLEU sont incommensurables d'une paire de langues à l'autre et d'une direction de traduction à l'autre)

Si cette hiérarchie des modèles est confirmée par les scores chrF rapportés table 5.5, en revanche les taux d'erreur TER et les scores COMET obtenus par ces modèles anglais → amharique, et reportés table 5.6, ont en commun de ne pas discerner de différence significative entre les performances des Modèles B° et G° en traduction (Modèle G° > Modèle B° pour une valeur-p non significative de 0.1920, selon COMET, le Modèle G° l'emportant, pour cet algorithme, sur le Modèle B° dans 78% des échantillons produits par *bootstrap resampling*, le Modèle B° sur le Modèle G° dans 20.67% de ces échantillons, ces deux modèles étant impossibles à départager dans 1.33% des cas).

2. Calculée indépendamment, avec le Modèle G° pour baseline.

Modèles EN → AM	BLEU (± 95% CI)	chrF2 (± 95% CI)	TER (± 95% CI)
Baseline : Modèle G°	14.1 (14.0 ± 0.7)	29.4 (29.4 ± 0.6)	82.7 (82.7 ± 1.0)
Modèle B°	13.6 (13.6 ± 0.7) (p = 0.0020)*	28.7 (28.7 ± 0.6) (p = 0.0010)*	82.1 (82.1 ± 0.8) (p = 0.0509)
Modèle D°	14.6 (14.6 ± 0.7) (p = 0.0020)*	30.7 (30.7 ± 0.6) (p = 0.0010)*	80.1 (80.1 ± 0.8) (p = 0.0010)*

Modèles EN → AM	chrF2 (± 95% CI)	TER (± 95% CI)
Modèle B°	28.7 (28.7 ± 0.6) (baseline)	82.1 (82.1 ± 0.8) (baseline)
Modèle G°	29.4 (29.4 ± 0.6) (p = 0.0010)*	82.7 (82.7 ± 1.0) (p = 0.0509)
Modèle D°	30.7 (30.7 ± 0.6) (p = 0.0010)*	80.1 (80.1 ± 0.8) (p = 0.0010)*

TABLE 5.5 – Scores chrF et taux d'erreur TER résultant de l'évaluation automatique des modèles anglais → amharique entraînés dans l'objectif de générer des données retraduites

Modèles EN → AM	COMET (wmt20-comet-da)
Modèle B°	-0.3787
Modèle G°	-0.3606
Modèle D°	-0.2003*

TABLE 5.6 – Résultats d'évaluation COMET des modèles Transformer anglais → amharique entraînés sur des données NON augmentées dans l'objectif de générer des données retraduites

Toutes ces métriques s'accordent, cependant, à reconnaître la supériorité du Modèle D° sur les autres (Modèle D° > Modèle B° pour p = 0.0000, et Modèle D° > Modèle G° pour p = 0.0000, selon COMET).

5.2.3 Évaluation automatique des modèles amharique → anglais entraînés sur des données augmentées par rétrotraduction

Modèles AM → EN	BLEU (± 95% CI)
Modèle B+	19.8 (19.8 ± 0.6) (baseline)
Modèle G+	19.7 (19.6 ± 0.7) (p = 0.1518)
Modèle D+	19.8 (19.7 ± 0.6) (p = 0.3357)

TABLE 5.7 – Scores BLEU résultant de l'évaluation automatique des modèles Transformer amharique → anglais entraînés sur des données augmentées par rétrotraduction

Les données retraduites générées par les Modèles B°, G° et D° devaient permettre la constitution de corpus parallèles augmentés sur lesquels trois modèles de traduction amharique → anglais B+, G+ et D+ furent subséquemment entraînés.

Les performances d'aucun de ces trois derniers modèles ne semblent différer significativement de celles des autres sur la base des scores BLEU, reportés table 5.7, ou chrF2 (cf. table 5.8) [3].

3. Voir également les scores, calculés indépendamment, avec le Modèle G+ pour baseline.

Modèles AM → EN	BLEU (± 95% CI)	chrF2 (± 95% CI)	TER (± 95% CI)
Baseline : Modèle B+	19.7 (19.6 ± 0.7)	40.5 (40.5 ± 0.5)	74.5 (74.5 ± 0.9)
Modèle G+	19.8 (19.8 ± 0.6) (p = 0.1518)	40.7 (40.6 ± 0.5) (p = 0.1678)	73.6 (73.6 ± 0.8) (p = 0.0260)*
Modèle D+	19.8 (19.7 ± 0.6) (p = 0.1948)	40.7 (40.7 ± 0.5) (p = 0.1309)	71.9 (71.9 ± 0.9) (p = 0.0010)*

Modèles AM → EN	chrF2 (± 95% CI)	TER (± 95% CI)
Modèle B+	40.7 (40.6 ± 0.5) (baseline)	73.6 (73.6 ± 0.8) (baseline)
Modèle G+	40.5 (40.5 ± 0.5) (p = 0.1678)	74.5 (74.5 ± 0.9) (p = 0.0260)*
Modèle D+	40.7 (40.7 ± 0.5) (p = 0.3017)	71.9 (71.9 ± 0.9) (p = 0.0020)*

TABLE 5.8 – Scores chrF et taux d'erreur TER, résultant de l'évaluation automatique des modèles amharique → anglais entraînés sur des données augmentées

On notera cependant que les taux d'erreur TER (cf. table 5.8) et les scores COMET (reportés table 5.9) s'accordent à assigner au Modèle D+ des performances significativement supérieures à celles des autres modèles (Modèle D+ > Modèle B+ pour une valeur-p = 0.0000, et Modèle D+ > Modèle G+ pour une valeur-p = 0.0000, d'après COMET). Plus paradoxalement, ces modèles s'accordent également à assigner au Modèle B+ des performances supérieures à celles du Modèle G+ (Modèle B+ > Modèle G+, valeur-p = 0.0100 selon COMET).

Modèles AM → EN	COMET (wmt20-comet-da)
Modèle B+	0.0683*
Modèle G+	0.0472*
Modèle D+	0.0929*

TABLE 5.9 – Résultats d'évaluation COMET des modèles Transformer amharique → anglais entraînés sur des données augmentées

5.3 Discussion

Les résultats obtenus en évaluation des Modèles G, B, D, G°, B° et D° permettent d'affirmer que les données parallèles issues d'un document relevant de la lexicographie bilingue, loin de nuire aux performances de systèmes de TA comme tâche, sont susceptibles d'offrir à ces derniers un surcroît significatif de performance, et ce relativement à des données de dimensions équivalentes provenant d'autres sources et notamment du Web.

La motivation fondamentale de l'expérience conduite mettant en jeu les Modèles G+, B+ et D+, entraînés sur des données augmentées par rétrotraduction [4], était de s'assurer que ces gains de performances observés pour les modèles "simples" n'étaient pas de nature triviale, et que la lourde tâche consistant à numériser et transcrire des ouvrages relevant de la lexicographie bilingue, puis à en extraire des données parallèles, en valait bien la peine et était en mesure d'offrir de meilleurs résultats qu'une simple opération de rétrotraduction de données monolingues issues du Web, infiniment moins coûteuse en temps et en efforts.

Une interprétation intransigeante des résultats précédemment rapportés conduirait à affirmer que les données issues de la lexicographie bilingue offrent bel et bien un surcroît de performance aux systèmes de TA comme tâche vis à vis de données augmentées par rétrotraduction, lesquelles semblent donc entraîner une nette dégradation des performances des modèles. Deux hypothèses alternatives peuvent, cependant, être envisagées qui sont propres à nuancer cette interprétation littérale des

4. Données retraduites, pour chacun des Modèles G+, B+ et D+, par les modèles G°, B° et D°, respectivement.

résultats d'évaluation précédemment rapportés : l'hypothèse d'une "déspécialisation" des modèles "augmentés" et celle de leur sous-entraînement.

Hypothèse 1 : "Déspécialisation"

Corpus	Segments	Mots graphiques anglais	Mots graphiques amhariques
Corpus d'évaluation issu de Gezmu et al.	2 500 (71%)	39 308 (64%)	38 683 (64%)
Corpus d'évaluation issu de FLORES-101	1 012 (29%)	21 901 (36%)	22 220 (36%)
Total :	3 512	61 209	60 903

TABLE 5.10 – Sources des segments et des mots graphiques du jeu de données d'évaluation des modèles.

La dégradation des scores BLEU des modèles de traduction amharique → anglais entraînés sur des données augmentées pourrait, en effet, se voir expliquée par une perte de spécialisation de ces modèles vis à vis du jeu de données d'évaluation employé dans cette étude. La table 5.10 rappelle, en effet, que le corpus d'évaluation compilé, pour ces expériences, par concaténation de jeux de données issus de deux sources hétérogènes ([Gezmu et al., 2022] et FLORES-101), est déséquilibré (à hauteur de plus de 70%) en faveur des données provenant de [Gezmu et al., 2022], elles-mêmes constitutives d'une part majeure (voire totale) des corpus ayant permis l'entraînement des modèles simples G, B et D, ainsi que le détaille la table 5.11. Les données retraduites, consistant en un million de segments issus de la partie anglaise du corpus giga-fren, "submergent" quantitativement ces données et sont ainsi susceptibles d'éloigner les modèles, entraînés sur leur base, du domaine restreint des textes religieux (Bible et presse protestante en ligne) propres au corpus de [Gezmu et al., 2022], comme en témoigneraient, donc, les moindres performances des modèles G+, B+ et D+, mesurées en termes de score BLEU, sur le corpus d'évaluation (cf. tables 5.7 et 5.1). La dégradation du score COMET du modèle D+ vis à vis du modèle D (0.0929 contre 0.1264) pourrait-être une preuve supplémentaire de la validité de cette hypothèse de "déspécialisation", laquelle échoue, toutefois, à expliquer l'inversion des performances des Modèles G+ et B+ vis à vis des modèles "simples" G et B, une inversion que s'accordent pourtant à détecter les scores COMET et le taux d'erreur TER (cf. tables 5.9, 5.8, 5.2 et 5.3).

Hypothèse 2 : Sous-entraînement

Une hypothèse alternative consisterait à expliquer les moindres performances des modèles G+, B+ et D+, entraînés sur des données augmentées par rétrotraduction, comme les indices d'un sous-entraînement, conséquence du coût élevé – en terme de temps de calcul – de l'entraînement de modèles Transformer de la TA comme tâche sur des données parallèles de dimensions étendues, et compte tenu – également – des moyens techniques limités à notre disposition [5] pour la conduite de la dernière

5. Soit un unique GPU Google Colab de puissance variable selon les instances et d'usage intermittent.

Modèles	Part des données d'entraînement des modèles issues de Gezmu et al. (en nombre de segments)
Modèles G/G°	100%
Modèles B/B°	100%
Modèles D/D°	76%
Modèle G+	12%
Modèle B+	11%
Modèle D+	10%

TABLE 5.11 – Part des données d'entraînement des modèles issues de Gezmu et al. (en nombre de segments)

phase de cette suite d'expériences. Cette hypothèse expliquerait, ainsi, l'inversion significative des performances des Modèles G+ et B+ vis à vis de celles des Modèles G et B – que s'accordent donc à relever les scores COMET et les taux d'erreur TER (cf. tables 5.9, 5.8, 5.2 et 5.3) – comme une aberration ponctuelle vouée à disparaître à la faveur d'un plus long entraînement. On rappellera, en effet (cf. figures 4.28, 4.29 et 4.30), que le Corpus B ayant servi à l'entraînement du Modèle B est un sous-ensemble du Corpus G ayant servi à l'entraînement du Modèle G, et que le Modèle G° – entraîné sur le Corpus G et ayant servi à la rétrotraduction des données d'entraînement augmentées du Modèle G+ – affichait des performances supérieures à celles du Modèle B° [6] ; les moindres performances, par conséquent, du Modèle G+ vis à vis de celles du Modèle B+ ont donc bien l'apparence d'une anomalie.

Synthèse

Les deux hypothèses précédemment évoquées ne sont pas mutuellement exclusives, et la "déspécialisation" comme le sous-entraînement pourraient expliquer, conjointement, les résultats d'évaluation paradoxaux des Modèles G+ et B+.

On retiendra, néanmoins, que des indices tels que :

— le taux d'erreur de traduction (TER) du Modèle D+, significativement inférieur à ceux des autres modèles "augmentés",
— ou le score COMET de ce même modèle, significativement supérieur à ceux des autres modèles "augmentés",
— ou encore les scores du Modèle D "simple", supérieurs – toutes métriques confondues – à ceux de tous les autres modèles amharique → anglais en lice, "simples" comme "augmentés"

tendent à indiquer que l'extraction de données parallèles au départ de documents relevant de la lexicographie bilingue vaut d'être entreprise en TA comme tâche des langues peu dotées, en ce qu'une telle opération est *susceptible* d'offrir des performances supérieures à la rétrotraduction :

— de manière hautement probable, dans des contextes où les moyens matériels à la disposition des chercheurs et des ingénieurs pour l'entraînement de systèmes de TA comme tâche seraient limités ;

6. Du moins sous le rapport des scores BLEU ; de manière remarquable, les scores COMET et les taux d'erreur TER s'accordaient à considérer comme non significative la différence observée entre les performances de ces deux modèles B° et G°, signalant peut-être un sous-entraînement des modèles de rétrotraduction eux-mêmes.

— et de manière vraisemblable dans les cas où ces modèles pourraient être entraînés plus longuement sur les données retraduites et leurs hyperparamètres optimisés.

Inférence causale

Une étude qualitative poussée des hypothèses de traduction produites par les modèles en lice – et qui reposerait en partie sur une approche en "boîte de verre" de l'évaluation des modèles de la TA comme tâche précédemment évoquée – serait nécessaire à la mise au jour du mécanisme causal à la faveur duquel les données issues de la lexicographie bilingue offriraient à ces systèmes ce surcroît de performance. La suite d'outils Fairseq rendant, malheureusement, laborieuse – voire interdisant – toute analyse de ce type, il est recommandé au lecteur qu'intéresserait une telle recherche de remplacer, dans ses expériences, la suite d'outils Fairseq par celle d'OpenNMT [7], compatible, à titre d'exemple, avec la bibliothèque BertViz précédemment évoquée.

Il est certain, quoi qu'il en soit, que les données parallèles issues de la lexicographie bilingue semblent présenter, sans surprise, des propriétés textométriques singulières, nettement différentes de celles des données provenant d'autres sources, ainsi qu'en atteste, à titre d'exemple, le calcul comparé des ratios type/token standardisé (cf. [Volkart and Bouillon, 2022], p. 5) des corpus mis en jeu par ces expériences, lesquels ratios semblent témoigner de la plus grande richesse lexicale des documents issus de la lexicographie bilingue, une richesse lexicale elle-même susceptible de placer les tokens BPE dans des contextes d'occurrence plus divers permettant aux modèles Transformer de la TA comme tâche d'en mieux saisir les propriétés sémantiques en traduction.

Corpus	Ratio Type/Token
Corpus parallèle issu des données lexicographiques	0.555477813194052
Corpus compilé par Gezmu et al. au départ du Web	0.456362042210556

TABLE 5.12 – Ratios types/tokens comparés du corpus de segments parallèles issu d'un dictionnaire et d'un corpus compilé au départ du Web (moyenne calculée sur la base de 10 000 échantillons aléatoires de 100 segments de langue anglaise)

7. Cf. https://opennmt.net/

Conclusion générale

La TA comme domaine, au tournant des années 2020, semble s'être prise de passion pour les applications de son objet d'étude à des langues pour lesquelles les données nécessaires au développement de systèmes de TA comme tâche n'existent qu'en quantités rares ou infimes, des langues peu dotées, voire très peu dotées, voire extrêmement peu dotées, voire même "zéro-dotées" !

Les approches dominantes visant à pallier ce défaut de données adéquates consistent principalement à tenter de les extraire du Web. De telles approches se heurtent néanmoins à la faible représentation sur la Toile des langues peu-dotées en question.

Nombre de ces langues disposent pourtant de remarquables ressources d'un autre ordre, celles issues de la lexicographie bilingue, mais ces documents n'existent que rarement sous une forme permettant le traitement informatique de leur contenu textuel. Si quelques travaux ([Bustamante et al., 2020, Rijhwani et al., 2020]) s'étaient employés à extraire d'ouvrages imprimés relevant du champ plus général de la didactique des langues (dictionnaires, méthodes de langues, grammaires) des corpus monolingues destinés au traitement automatique des langues peu dotées, le présent mémoire constitue une tentative fructueuse d'extraction de données parallèles au départ de tels documents imprimés, photographiés ou scannés, mettant à profit la régularité de leurs propriétés formelles – trait caractéristique des documents relevant de la lexicographie bilingue depuis la plus haute Antiquité – et les avancées récentes de la vision par ordinateur. Cette étude, en dépit de moyens limités, a permis la compilation, en seulement cinq mois d'un travail assidu, d'un corpus parallèle de quelques 45 000 paires de segments anglais et amharique, et a été l'occasion d'une expérience tendant à démontrer la valeur tangible d'un tel corpus pour la TA comme tâche d'une langue peu dotée.

On s'est efforcé, au fil de la seconde partie du présent mémoire, de documenter en détail le protocole ayant permis d'obtenir ces résultats, à la faveur d'une "chaîne opératoire" dont on espère qu'elle permettra au lecteur qui le souhaiterait de développer pour d'autres de ces langues "zéro-dotées" précédemment mentionnées les artefacts du TAL et de la TA comme tâche qui leur font défaut.

Outre la possibilité de remplacer la suite d'outils Fairseq par celle d'OpenNMT précédemment évoquée, en discussion, le lecteur pourra tenter avec avantage d'appliquer au corpus parallèle, obtenu au départ de documents lexicographiques, les approches d'augmentation des données par synthèse d'exemples illustrées chez [Fadaee et al., 2017] ou encore, plus récemment, chez [Liu et al., 2021], et susceptibles d'être particulièrement appropriées à de tels jeux de données. Le recours comparé à des dictionnaires bilingues monoscopiques et discopiques – notamment dans le contexte d'expériences de rétrotraduction – est également susceptible d'être digne d'intérêt.

De tels travaux valent, sans doute possible, la peine d'être entrepris. Si l'intérêt des complexes militaro-industriels des grands empires contemporains pour la TA

comme domaine doit conduire à modérer toute naïveté, il convient de noter que les artefacts de la TA comme tâche des langues peu dotées sont susceptibles de faire plus de bien que de mal en ce bas monde, notamment en contribuant à la survie des langues en danger, ainsi que se plaît à l'imaginer Mari C. Jones, dans un ouvrage dédié au rôle des nouvelles technologies dans le sauvetage des langues menacées ; cette autrice y dépeint la vision optimiste de l'avenir heureux que ces artefacts pourraient offrir à l'humanité dans toute sa diversité, écrivant :

> "Communication among those who have different native languages can be achieved in at least five ways : (a) by the services of bilingual interpreters ; (b) by general adoption of a single lingua franca ; (c) by systematic translation of key texts ; (d) by language pedagogy ; and (e) by automatic language conversion.
>
> These five solutions have arisen historically at different times. [...]
>
> Currently, the world's international communications are dominated by the lingua franca model, with English as medium. [Our prediction] is that the world will lose its motivation to maintain English as a convenient lingua franca just as automatic language conversion becomes a realistic, and crucially an easier and cheaper, alternative to support interlingual communication. This use of automatic language conversion (automatic interpreting and machine translation) will be increasingly egalitarian between languages, since it will be based solely on the amount of electronically recorded data available for each language involved. Although the languages of great powers which have fostered its development will have an initial advantage, there is little or nothing intrinsic to the technologies used for conversion which will favour these languages over others : in the long run, the ability to generate data, and have it recorded and statistically analysed, is open to all. The net effect will be that the smaller 99 per cent of the world's languages, namely those with fewer than sixteen million speakers, will have a corresponding opportunity to become accessible ; the 'long tail' need no longer be disregarded. [...]
>
> *In this world of aspiration, all will speak as they like, and yet the world will understand them.*"
>
> ([Jones, 2014], pp. 1-13, nos italiques)

Il n'y a donc plus qu'à ! Haut les cœurs !

BIBLIOGRAPHIE

[Abate et al., 2018] Abate, S. T., Melese, M., Tachbelie, M. Y., Meshesha, M., Atinafu, S., Mulugeta, W., Assabie, Y., Abera, H., Ephrem, B., Abebe, T., Tsegaye, W., Lemma, A., Andargie, T., and Shifaw, S. (2018). Parallel Corpora for bi-lingual English-Ethiopian Languages Statistical Machine Translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3102–3111, Santa Fe, New Mexico, USA. Association for Computational Linguistics. – Cité pages 45, 46 et 47.

[Adamska-Sałaciak, 2022] Adamska-Sałaciak, A. (2022). Issues in compiling bilingual dictionaries. *The Bloomsbury Companion to Lexicography, 3rd Ed. p.193*. – Cité page 49.

[Amberber, 2008] Amberber, M. (2008). *4. Semantic primes in Amharic*. Studies in Language Companion Series. John Benjamins Publishing Company. Pages: 83-119 Publication Title: Cross-Linguistic Semantics. – Cité page 40.

[Anderson, 2006] Anderson, N. (2006). Defense Department funds massive speech recognition and translation program. – Cité page 31.

[Arthur et al., 2016] Arthur, P., Neubig, G., and Nakamura, S. (2016). Incorporating Discrete Translation Lexicons into Neural Machine Translation. arXiv:1606.02006 [cs]. – Cité page 50.

[Bahdanau et al., 2016] Bahdanau, D., Cho, K., and Bengio, Y. (2016). Neural Machine Translation by Jointly Learning to Align and Translate. arXiv:1409.0473 [cs, stat]. – Cité pages 22 et 25.

[Baker and Saldanha, 2019] Baker, M. and Saldanha, G. (2019). *Routledge Encyclopedia of Translation Studies*. Routledge. Google-Books-ID: c9CwDwAAQBAJ. – Cité page 21.

[Baruah et al., 2020] Baruah, R., Mundotiya, R. K., Kumar, A., and Singh, A. k. (2020). NLPRL System for Very Low Resource Supervised Machine Translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1075–1078, Online. Association for Computational Linguistics. – Cité page 14.

[Biadgligne and Smaïli, 2022] Biadgligne, Y. and Smaïli, K. (2022). Offline Corpus Augmentation for English-Amharic Machine Translation. – Cité page 46.

[Biadgline and Smaïli, 2021] Biadgline, Y. and Smaïli, K. (2021). Parallel Corpora Preparation for English-Amharic Machine Translation. – Cité page 45.

[Brown and Ogilvie, 2008] Brown, K. and Ogilvie, S., editors (2008). *Concise Encyclopedia of Languages of the World*. Elsevier Science, Amsterdam Oxford, 1st edition edition. – Cité pages 38 et 39.

[Bustamante et al., 2020] Bustamante, G., Oncevay, A., and Zariquiey, R. (2020). No Data to Crawl? Monolingual Corpus Creation from PDF Files of Truly low-Resource Languages in Peru. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2914–2923, Marseille, France. European Language Resources Association. – Cité pages 14, 51 et 93.

[Caswell and Bapna, 2022] Caswell, I. and Bapna, A. (2022). Unlocking Zero-Resource Machine Translation to Support New Languages in Google Translate. – Cité page 14.

[Caswell et al., 2019] Caswell, I., Chelba, C., and Grangier, D. (2019). Tagged Back-Translation. arXiv:1906.06442 [cs]. – Cité page 81.

[Chambers and Trudgill, 1998] Chambers, J. K. and Trudgill, P. (1998). *Dialectology*. Cambridge University Press. Google-Books-ID: 9bYV43UhKssC. – Cité page 35.

[Chimoto and Bassett, 2022] Chimoto, E. and Bassett, B. (2022). Very Low Resource Sentence Alignment: Luhya and Swahili. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 1–8, Gyeongju, Republic of Korea. Association for Computational Linguistics. – Cité page 48.

[Cho et al., 2014] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. arXiv:1406.1078 [cs, stat]. – Cité page 25.

[Chomsky, 1986] Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin, and Use*. Greenwood Publishing Group. Google-Books-ID: b0VZPtZDL8kC. – Cité page 35.

[Church and Kordoni, 2022] Church, K. W. and Kordoni, V. (2022). Emerging Trends: SOTA-Chasing. *Natural Language Engineering*, 28(2):249–269. Publisher: Cambridge University Press. – Cité page 34.

[Conneau et al., 2020] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. arXiv:1911.02116 [cs]. – Cité pages 8, 33, 47 et 48.

[Cori, 2020] Cori, M. (2020). *Le traitement automatique des langues en question: des machines qui comprennent le français?* Cassini. Google-Books-ID: x0zjvgEACAAJ. – Cité page 13.

[Coupaye, 2022] Coupaye, L. (2022). Making 'Technology' Visible: Technical Activities and the Chaîne Opératoire. In Bruun, M. H., Wahlberg, A., Douglas-Jones, R., Hasse, C., Hoeyer, K., Kristensen, D. B., and Winthereik, B. R., editors, *The Palgrave Handbook of the Anthropology of Technology*, pages 37–60. Springer, Singapore. – Cité page 15.

[Davidson, 1986] Davidson, D. (1986). A Nice Derangement of Epitaphs. In Lepore, E., editor, *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*, pages 433–446. Blackwell. – Cité page 35.

[de Varennes, 2012] de Varennes, F. (2012). Language policy at the supranational level. In Spolsky, B., editor, *The Cambridge Handbook of Language Policy*, Cambridge Handbooks in Language and Linguistics, pages 149–173. Cambridge University Press, Cambridge. – Cité page 38.

[Ethayarajh and Jurafsky, 2021] Ethayarajh, K. and Jurafsky, D. (2021). Utility is in the Eye of the User: A Critique of NLP Leaderboards. arXiv:2009.13888 [cs]. – Cité pages 30 et 34.

[Fabri et al., 2014] Fabri, R., Gasser, M., Habash, N., Kiraz, G., and Wintner, S. (2014). Linguistic Introduction: The Orthography, Morphology and Syntax of Semitic Languages. pages 3–41. – Cité page 39.

[Fadaee et al., 2017] Fadaee, M., Bisazza, A., and Monz, C. (2017). Data Augmentation for Low-Resource Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics. – Cité pages 50 et 93.

[Fernando and Ranathunga, 2022] Fernando, A. and Ranathunga, S. (2022). Data Augmentation to Address Out-of-Vocabulary Problem in Low-Resource Sinhala-English Neural Machine Translation. arXiv:2205.08722 [cs]. – Cité page 50.

[Fischer, 2003] Fischer, S. R. (2003). *History of Writing*. Reaktion Books. Google-Books-ID: Ywo0M9OpbXoC. – Cité page 40.

[Gage, 1994] Gage, P. (1994). A New Algorithm for Data Compression. page 14. – Cité page 23.

[Gao, 2022] Gao, S. (2022). Système de traduction automatique neuronale français-mongol (Historique, mise en place et évaluations) (French-Mongolian Neural Machine Translation System (History, Implementation, and evaluations) Machine Translation (hereafter abbreviated MT) is currently undergoing rapid development, during which less-resourced languages nevertheless seem to be less developed). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 2 : 24e Rencontres Etudiants Chercheurs en Informatique pour le TAL (RECITAL)*, pages 97–110, Avignon, France. ATALA. – Cité page 48.

[Gasser, 2017] Gasser, M. (2017). Minimal Dependency Translation: a Framework for Computer-Assisted Translation for Under-Resourced Languages. arXiv:1710.00923 [cs]. – Cité page 46.

[Gehring et al., 2017] Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017). Convolutional Sequence to Sequence Learning. arXiv:1705.03122 [cs]. – Cité page 25.

[Geva et al., 2021] Geva, M., Schuster, R., Berant, J., and Levy, O. (2021). Transformer Feed-Forward Layers Are Key-Value Memories. arXiv:2012.14913 [cs]. – Cité page 29.

[Gezmu et al., 2022] Gezmu, A. M., Nürnberger, A., and Bati, T. B. (2022). Extended Parallel Corpus for Amharic-English Machine Translation. arXiv:2104.03543 [cs]. – Cité pages 44, 46, 77, 78 et 90.

[Haddow et al., 2021] Haddow, B., Bawden, R., Barone, A. V. M., Helcl, J., and Birch, A. (2021). Survey of Low-Resource Machine Translation. arXiv:2109.00486 [cs]. – Cité page 14.

[Hadgu et al., 2020] Hadgu, A. T., Beaudoin, A., and Aregawi, A. (2020). Evaluating Amharic Machine Translation. arXiv:2003.14386 [cs]. – Cité page 46.

[Hagiwara, 2021] Hagiwara, M. (2021). *Real-World Natural Language Processing: Practical Applications with Deep Learning*. Simon and Schuster. Google-Books-ID: Ok5NEAAAQBAJ. – Cité pages 74 et 79.

[Hassan et al., 2018] Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T.-Y., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Wu, S., Xia, Y., Zhang, D., Zhang, Z., and Zhou, M. (2018). Achieving Human Parity on Automatic Chinese to English News Translation. arXiv:1803.05567 [cs]. – Cité page 13.

[Hetzron, 1976] Hetzron, R. (1976). *The Agaw Languages*. Undena Publ. Google-Books-ID: g7xZGQAACAAJ. – Cité page 39.

[Hinton, 2017] Hinton, G. (2017). Geoffrey Hinton: What is wrong with convolutional neural nets? – Cité page 25.

[Hirst, 2013] Hirst, G. (2013). Computational Linguistics. In Allan, K., editor, *The Oxford Handbook of the History of Linguistics*, page 0. Oxford University Press. – Cité page 21.

[Huang et al., 2022] Huang, F., Tao, T., Zhou, H., Li, L., and Huang, M. (2022). On the Learning of Non-Autoregressive Transformers. arXiv:2206.05975 [cs]. – Cité page 30.

[Jones, 2014] Jones, M. C. (2014). *Endangered Languages and New Technologies*. Cambridge University Press. Google-Books-ID: HpBEBQAAQBAJ. – Cité page 94.

[Jurafsky and Martin, 2022] Jurafsky, D. and Martin, J. H. (2022). *Speech and Language Processing (3rd ed. draft)*. Prentice Hall. Google-Books-ID: fZmj5UNK8AQC. – Cité page 39.

[King, 1996] King, M. (1996). *On the notion of validity and the evaluation of machine translation systems*. Benjamins Translation Library. John Benjamins Publishing Company. Pages: 189 Publication Title: Terminology, LSP and Translation: Studies in language engineering in honour of Juan C. Sager. – Cité page 32.

[Knowles, 2021] Knowles, R. (2021). On the Stability of System Rankings at WMT. In *Proceedings of the Sixth Conference on Machine Translation*, pages 464–477, Online. Association for Computational Linguistics. – Cité page 34.

[Koehn, 2010] Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press. Google-Books-ID: 4v_Cx1wIMLkC. – Cité page 19.

[Koehn, 2020] Koehn, P. (2020). *Neural Machine Translation*. Cambridge University Press. Google-Books-ID: iRzhDwAAQBAJ. – Cité pages 19, 21, 33 et 85.

[Kreyenbroek and Sperl, 1992] Kreyenbroek, P. G. and Sperl, S. (1992). *The Kurds: A Contemporary Overview*. Routledge. Google-Books-ID: DkI1u4ta5w4C. – Cité page 35.

[Krubiński et al., 2021] Krubiński, M., Ghadery, E., Moens, M.-F., and Pecina, P. (2021). Just Ask! Evaluating Machine Translation by Asking and Answering Questions. In *Proceedings of the Sixth Conference on Machine Translation*, pages 495–506, Online. Association for Computational Linguistics. – Cité page 34.

[Kudo and Richardson, 2018] Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. arXiv:1808.06226 [cs]. – Cité page 24.

[Leslau, 1973] Leslau, W. (1973). *English-Amharic Context Dictionary*. Otto Harrassowitz Verlag. Google-Books-ID: x4JdPU7wfpQC. – Cité pages 8, 40, 56 et 57.

[Liu et al., 2021] Liu, Q., Kusner, M., and Blunsom, P. (2021). Counterfactual Data Augmentation for Neural Machine Translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 187–197, Online. Association for Computational Linguistics. – Cité page 93.

[Mathur et al., 2020] Mathur, N., Baldwin, T., and Cohn, T. (2020). Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics. arXiv:2006.06264 [cs]. – Cité page 34.

[Mielke et al., 2021] Mielke, S. J., Alyafeai, Z., Salesky, E., Raffel, C., Dey, M., Gallé, M., Raja, A., Si, C., Lee, W. Y., Sagot, B., and Tan, S. (2021). Between words and characters: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP. arXiv:2112.10508 [cs]. – Cité page 22.

[Ni et al., 2022] Ni, J., Jin, Z., Freitag, M., Sachan, M., and Schölkopf, B. (2022). Original or Translated? A Causal Analysis of the Impact of Translationese on Machine Translation Performance. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5303–5320, Seattle, United States. Association for Computational Linguistics. – Cité page 37.

[Ott et al., 2019] Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics. – Cité page 78.

[Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. – Cité page 32.

[Pauw et al., 2012] Pauw, G., Schryver, G.-M. d., Forcada, M., Sarasola, K., Tyers, F., and Wagacha, P. (2012). *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages (SaLTMiL 8 - AfLaT 2012)*. – Cité page 46.

[Poibeau, 2017] Poibeau, T. (2017). *Machine Translation*. MIT Press. Google-Books-ID: LYc3DwAAQBAJ. – Cité page 19.

[Poibeau, 2019] Poibeau, T. (2019). *Babel 2.0: où va la traduction automatique?* Odile Jacob. Google-Books-ID: SrJmxQEACAAJ. – Cité page 19.

[Popović, 2015] Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics. – Cité page 85.

[Post and Vilar, 2018] Post, M. and Vilar, D. (2018). Fast Lexically Constrained Decoding with Dynamic Beam Allocation for Neural Machine Translation. arXiv:1804.06609 [cs]. – Cité page 50.

[Provilkov et al., 2020] Provilkov, I., Emelianenko, D., and Voita, E. (2020). BPE-Dropout: Simple and Effective Subword Regularization. arXiv:1910.13267 [cs]. – Cité pages 8 et 24.

[Ranathunga et al., 2021] Ranathunga, S., Lee, E.-S. A., Skenduli, M. P., Shekhar, R., Alam, M., and Kaur, R. (2021). Neural Machine Translation for Low-Resource Languages: A Survey. arXiv:2106.15115 [cs]. – Cité pages 14, 37 et 46.

[Rei et al., 2022] Rei, R., Farinha, A. C., de Souza, J. G., Ramos, P. G., Martins, A. F., Coheur, L., and Lavie, A. (2022). Searching for COMETINHO: The Little Metric That Could. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 61–70, Ghent, Belgium. European Association for Machine Translation. – Cité page 34.

[Rei et al., 2020] Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics. – Cité page 33.

[Rijhwani et al., 2020] Rijhwani, S., Anastasopoulos, A., and Neubig, G. (2020). OCR Post Correction for Endangered Language Texts. arXiv:2011.05402 [cs]. – Cité pages 51 et 93.

[Robins, 2013] Robins, R. H. (2013). *A Short History of Linguistics*. Routledge. Google-Books-ID: CAIvAgAAQBAJ. – Cité page 50.

[Sennrich et al., 2016a] Sennrich, R., Haddow, B., and Birch, A. (2016a). Improving Neural Machine Translation Models with Monolingual Data. arXiv:1511.06709 [cs]. – Cité page 37.

[Sennrich et al., 2016b] Sennrich, R., Haddow, B., and Birch, A. (2016b). Neural Machine Translation of Rare Words with Subword Units. arXiv:1508.07909 [cs]. – Cité page 23.

[Snell-Hornby, 1986] Snell-Hornby, M. (1986). *The Bilingual Dictionary - Victim of its own tradition?* Studies in the History of the Language Sciences. John Benjamins Publishing Company. Pages: 207 Publication Title: The History of Lexicography. – Cité page 49.

[Sutskever et al., 2014] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. arXiv:1409.3215 [cs]. – Cité page 25.

[Swales, 2004] Swales, J. M. (2004). *Research Genres: Explorations and Applications*. Cambridge University Press. Google-Books-ID: 6dcBYY96No8C. – Cité page 30.

[Takebayashi et al., 2018] Takebayashi, Y., Chenhui, C., Arase\dag, Y., and Nagata, M. (2018). Word Rewarding for Adequate Neural Machine Translation. In *Proceedings of the 15th International Conference on Spoken Language Translation*, pages 14–22, Brussels. International Conference on Spoken Language Translation. – Cité page 50.

[Tars et al., 2021] Tars, M., Tättar, A., and Fišel, M. (2021). Extremely low-resource machine translation for closely related languages. arXiv:2105.13065 [cs]. – Cité page 14.

[Teshome and Besacier, 2012] Teshome, M. and Besacier, L. (2012). Preliminary experiments on English-Amharic statistical machine translation. *Proc. 3rd Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2012)*, pages 36–41. – Cité pages 8, 45 et 46.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need. arXiv:1706.03762 [cs]. – Cité pages 14, 24 et 25.

[Volkart and Bouillon, 2022] Volkart, L. and Bouillon, P. (2022). Studying Post-Editese in a Professional Context: A Pilot Study. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 71–79, Ghent, Belgium. European Association for Machine Translation. – Cité page 92.

[Wang et al., 2019] Wang, C., Cho, K., and Gu, J. (2019). Neural Machine Translation with Byte-Level Subwords. arXiv:1909.03341 [cs]. – Cité page 23.

[Wang et al., 2021] Wang, R., Tan, X., Luo, R., Qin, T., and Liu, T.-Y. (2021). A Survey on Low-Resource Neural Machine Translation. arXiv:2107.04239 [cs]. – Cité page 14.

[Wang and Chen, 2020] Wang, Y.-A. and Chen, Y.-N. (2020). What Do Position Embeddings Learn? An Empirical Study of Pre-Trained Language Model Positional Encoding. arXiv:2010.04903 [cs]. – Cité page 27.

[Way, 2010] Way, A. (2010). A critique of Statistical Machine Translation. *Linguistica Antverpiensia*, 8. – Cité page 21.

[Wu et al., 2016] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, , Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv:1609.08144 [cs]. – Cité page 13.

[Zgusta, 2010] Zgusta, L. (2010). *Manual of lexicography*. Walter de Gruyter. Google-Books-ID: aZyBzC212BoC. – Cité page 49.

[Zhang et al., 2021] Zhang, T., Zhang, L., Ye, W., Li, B., Sun, J., Zhu, X., Zhao, W., and Zhang, S. (2021). Point, Disambiguate and Copy: Incorporating Bilingual Dictionaries for Neural Machine Translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3970–3979, Online. Association for Computational Linguistics. – Cité page 50.

[Zhong and Chiang, 2020] Zhong, X. J. and Chiang, D. (2020). Look It Up: Bilingual and Monolingual Dictionaries Improve Neural Machine Translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 538–549, Online. Association for Computational Linguistics. – Cité page 50.

[Zhuo et al., 2022] Zhuo, T. Y., Xu, Q., He, X., and Cohn, T. (2022). Rethinking Round-trip Translation for Automatic Machine Translation Evaluation. arXiv:2209.07351 [cs]. – Cité page 37.

[Àlex R. Atrio and Popescu-Belis, 2022] Àlex R. Atrio and Popescu-Belis, A. (2022). On the Interaction of Regularization Factors in Low-resource Neural Machine Translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 111–120, Ghent, Belgium. European Association for Machine Translation. – Cité page 79.