

Modèle de corpus aligné pour une traduction automatique des langues Kwa à partir du traitement des contraintes de l'abron

Ambemou Oscar DIANÉ

ambemouoscar@gmail.com

Université Alassane Ouattara, Bouaké, Côte d'Ivoire

Tano Kouakou Frédéric MORIFIÉ

tkouakoufrederic@yahoo.fr

Université Alassane Ouattara, Bouaké, Côte d'Ivoire

Résumé : Les langues ivoiriennes, en particulier les langues kwa, ne sont pas suffisamment présentes parmi les ressources linguistiques numériques nécessaires au développement des systèmes de traduction automatique. Cette insuffisance s'explique par la rareté, voire l'inexistence de corpus alignés et de modèles de langue adaptés, capables de traiter leurs spécificités et leurs contraintes linguistiques. L'exigence du TAL étant le Deep Learning/Apprentissage profond à partir des données massives (Big Data), il se trouve que de nombreuses langues africaines ne disposent pas des grands corpus annotés nécessaires au développement de systèmes robustes de traitement automatique du langage naturel par Apprentissage Profond/Deep Learning de la Machine. Moufoutaou ADJERAN & Justine BASSABI 2024, p10). A l'effet de contribuer à réduire ces insuffisances relevées, le présent article vise à proposer un modèle de corpus aligné servant de base au traitement automatique des langues kwa. Pour atteindre cet objectif, les contraintes de traduction vers le français sont analysées à partir de l'Abron une langue kwa, précisément avec des données qualitatives issues de la recherche documentaire, notamment des thèses consacrées à cette langue. Sur le plan théorique, l'étude s'appuie sur l'approche statistique de la traduction automatique, issue des travaux des chercheurs d'IBM dans les années 1980, notamment ceux de Peter Brown et al. (1993). Face aux limites du modèle standard qui repose sur un découpage mot-à-mot, la réflexion ressort la nécessité de renforcer le modèle existant par l'ajout d'une segmentation syntagmatique au modèle de langue, en complément de l'approche fondée sur la traduction mot à mot. Cette approche, mieux adaptée aux particularités linguistiques de ces langues, devrait permettre d'obtenir une traduction plus satisfaisante.

Mots clés : corpus aligné, traduction automatique, approche statistique, langues kwa, contraintes

Abstract: Ivorian languages, particularly Kwa languages, are underrepresented in the digital linguistic resources required to develop machine translation systems. This shortfall stems from the scarcity, or even total absence, of aligned corpora and tailored language models capable of handling their specific linguistic characteristics and constraints. Given that modern Natural Language Processing (NLP) relies heavily on deep learning using massive datasets (Big Data), many African languages lack the large annotated corpora needed to develop robust deep learning-based NLP systems (Moufoutaou Adjeran & Justine Bassabi, 2024, p. 10). To help address these shortcomings, this article proposes an aligned corpus model to serve as a foundation for the automated processing of Kwa languages. To achieve this, the study analyzes translation constraints into French using Abron, a Kwa language, drawing on qualitative data from documentary research, specifically theses focused on the language. Theoretically, the study draws on the statistical approach to machine translation pioneered by IBM researchers in the 1980s, notably Peter Brown et al. (1993). Recognizing the limitations of the standard word-for-word model, the analysis highlights the need to enhance the existing framework by incorporating syntagmatic segmentation into the language model, complementing the word-for-word translation approach. This approach, which is better suited to the linguistic specificities of these languages, should yield more satisfactory translation results.

Keywords: aligned corpus, machine translation, statistical approach, Kwa languages, constraints

Introduction

La traduction automatique connaît de plus en plus des avancées majeures. Elle est désormais capable de fournir des résultats satisfaisants pour certaines langues, et

plus ou moins fiables pour d'autres, ce qui n'était pas le cas il y a quelques décennies. Des plateformes telles que *DeepL* ou *Google Translate* proposent aujourd'hui des traductions de qualité appréciable. Cette évolution est due en grande partie aux approches statistiques et neuronales de la traduction automatique. En effet, l'apparition de la traduction automatique statistique puis neuronale a fortement contribué au développement de cette discipline grâce à l'utilisation de corpus alignés de grande taille. Comme le souligne Boitet Christian (2007, p 25) : « On utilise maintenant des corpus gigantesques (jusqu'à 200 millions de mots) en TA empirique directe, surtout en TA statistique (...) ». Ces corpus sont exploités de différentes manières : « Alors que les systèmes statistiques utilisent les corpus comme ingrédients de leurs traductions, les moteurs neuronaux s'en servent pour apprendre à traduire par eux-mêmes », James Hadley (2020, p 15). Malgré ces avancées, ces approches ne sont pas exemptes de limites. Si, tentative après tentative, l'automatisation complète de l'activité de traduire s'avère impossible pour toutes les langues naturelles, c'est parce que celles-ci se distinguent les unes des autres à plusieurs niveaux linguistiques (Djebaili Farida, 2011, p 11). Cette impossibilité ne relève pas d'un manque de compétence des spécialistes de la traduction automatique ni des informaticiens, mais de ressources linguistiques numériques et de la complexité intrinsèque de l'activité de traduction, due en grande partie aux spécificités des langues naturelles. Elle demeure une tâche difficile, loin d'être parfaitement maîtrisée, même par les humains. C'est pourquoi, la présente étude vise à proposer un modèle de corpus aligné pour contribuer à une traduction automatique fluide et précise des langues kwa. Pour atteindre cet objectif, nous avons analysé les imperfections du modèle de corpus aligné sous la forme standard et décrit les spécificités à l'origine de ces imperfections.

Il s'agit primo de présenter l'approche statistique de la traduction automatique sur laquelle repose cette étude. L'alimentation de celle-ci par un corpus aligné abron-français nous permettra ensuite de mettre en évidence les imperfections, de les analyser et de décrire les spécificités de ces langues. De celles-ci naîtra une nouvelle approche de traduction automatique des langues kwa, susceptible d'apporter des solutions aux difficultés rencontrées dans l'approche standard.

1. Approche statistique de la traduction automatique

Les approches de traduction automatique basées sur corpus sont aux nombres de deux : l'approche statistique et neuronale. Nous optons dans le cadre de cette étude l'approche statistique. Appelée en anglais statistical machine translation (SMT), cette approche s'est développée vers la fin des années à partir des travaux des chercheurs d'IBM. L'idée fondamentale est de considérer la traduction comme un problème probabiliste. Dans un ensemble de traductions possibles, le système cherche celle la plus probable statistiquement. C'est une base de données linguistiques composées d'un corpus aligné ou parallèle et d'un modèle de langue qui n'est rien d'autre que du mot à mot et de leur occurrence (les informations statistiques associées). Les machines apprennent donc à traduire à partir d'un grand nombre d'exemples de traductions humaines. Cette base de données est alimentée, dans cette étude, par un corpus aligné tiré des thèses sur l'abron : Kossonou (2015), Loua Cyrille (2018) et Morifié Tano (2024), comportant déjà des textes traduits.

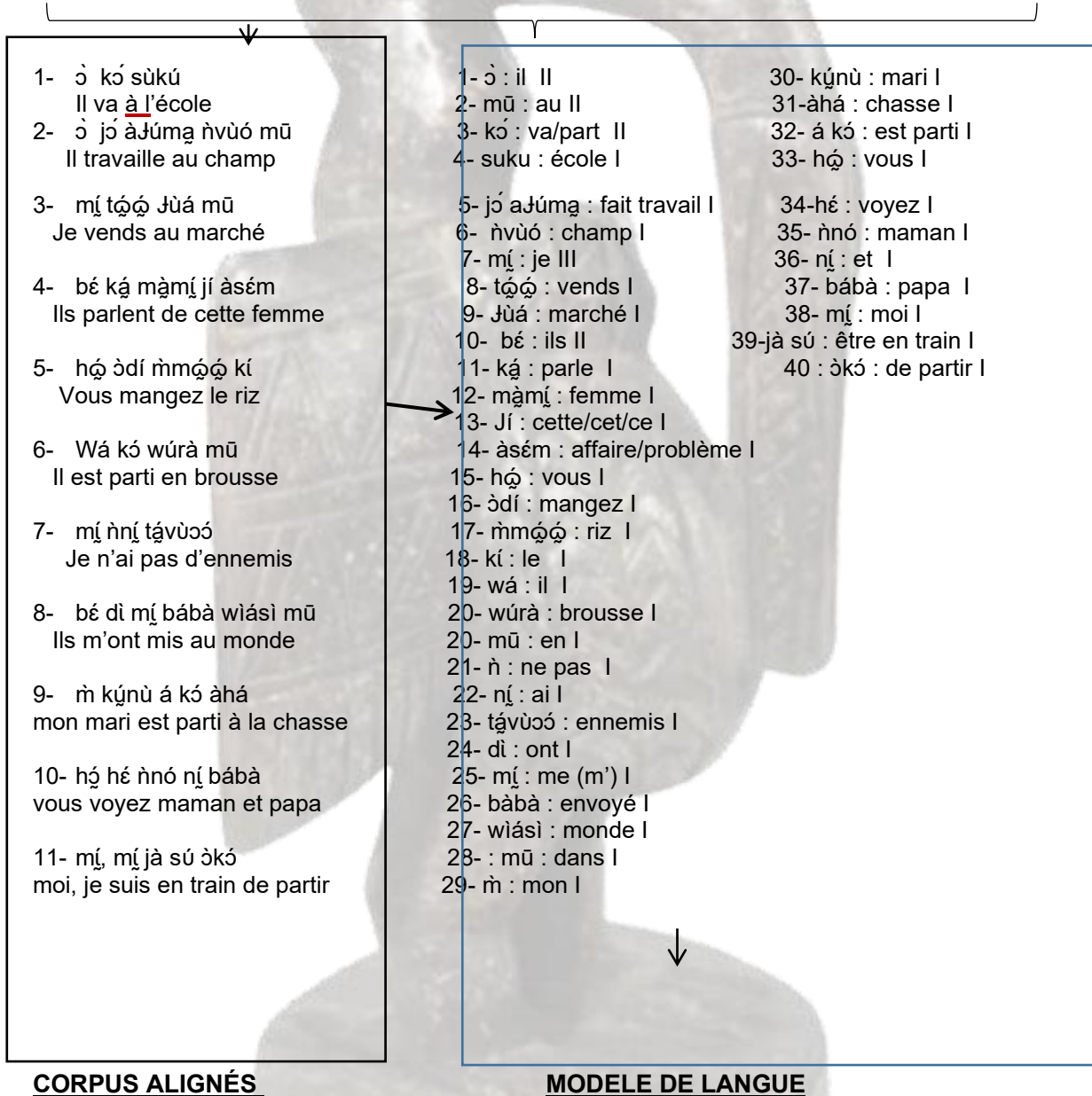
1.1. Présentation du système de l'approche statistique de traduction automatique

Le système de traduction que nous présentons repose sur quatre éléments fondamentaux. Deux de ces éléments sont les points focaux voire le cerveau du système, car ils sont à l'origine de toutes les opérations de traduction qui sont faites. Les deux autres

sont l'interface d'entrée grâce à qui les données sont saisies à l'entrée du système et l'interface de sortie qui affiche les résultats de la traduction. Ces deux interfaces sont celles visibles par l'internaute. Soit le schéma suivant :

(1)

Données d'entrées : mí kó Jùámū - bé jò àJúmā sùkú mū - h́́ ká Jùá mū àsém - ò kó àhá



Données de sorties : je va*au marché - vous parle marché affaire* - ils travaille* à l'école
Il va chasse*

Figure 1: Système de traduction automatique basé sur la linguistique de corpus
(MORIFIE Tano, 202, p 299)

Ce schéma-ci représente le système d'un corpus aligné alimenté par deux langues : l'abron, la langue source et le français, la langue cible. L'analyse porte sur le point focal du système, composé d'un « corpus aligné » à proprement dit réunissant des phrases de la langue source et leurs équivalences dans la langue cible et d'un modèle de langue, le découpage monématique des phrases de la langue source et leurs correspondants dans la langue cible. L'on y trouve également les occurrences des mots de la langue source. Ces occurrences sont indiquées par une numérotation en chiffres romains (I, II, III, etc.) placée devant chaque item dans le modèle de langue. Ces chiffres correspondent au nombre de fois qu'un item apparaît dans le corpus aligné. Ce décompte constitue ainsi une mesure statistique issue de l'analyse du corpus aligné.

1.2. Fonctionnement et analyse des données du système de traduction automatique

Le système de traduction automatique opère de deux manières. Premièrement, lorsqu'une demande de traduction est faite au niveau de l'interface d'entrée, le système procède par une mise en correspondance des données saisies dans le corpus aligné, c'est-à-dire le système recherche dans la base de données l'unité ou suite d'unités linguistiques saisies. Une fois trouvée, le système affiche cette correspondance. En second lieu, si celles-ci n'existent pas dans la base, le système procède par un découpage mot à mot et lance une requête dans le modèle de langue espérant trouver les mots et leurs correspondants dans la langue cible. C'est à ce niveau que le système rencontre des difficultés en termes d'erreurs de traduction. En effet, avec l'approche statistique, les mots ne s'affichent pas parce qu'ils sont les correspondants parfaits mais plutôt parce qu'ils ont une forte probabilité d'apparaître. La figure1, présentée ci-dessus, montre clairement que les exemples suivants tirés du modèle de langue en sont une illustration parfaite. L'item 1 : « ò » (il) a deux occurrences, l'item 2 : « mū » (au) a deux occurrences, l'item 22 « mū » (en) a une seule occurrence, l'item 28 : « mū » (dans) a également une seule occurrence. L'item 7 : « mǐ » (je) a trois occurrences, L'item 25 : « mǐ » (me, m') a une occurrence, l'item 38 : « mǐ » (moi) en a également une. Il ressort de cette analyse que des mêmes mots peuvent renvoyer à d'autres mots avec moins ou plus d'occurrence que les précédents ou subséquent. Soit la représentation suivante :

(2)



Figure 2: Correspondances des items mǐ et mū et occurrences

Au regard de la numérotation des items français en chiffres romains, le pronom « je », qui compte trois occurrences (III), présente une probabilité d'apparition plus élevée que les autres correspondances de « mǐ ». Il en est de même pour « au », qui compte deux occurrences (II), parmi les différentes correspondances de « mū ».

En d'autres termes, un même item de la langue source peut avoir plusieurs correspondants dans la langue cible. Cette multiplicité des correspondances constitue l'une des limites du système standard. Il s'agit là du premier niveau d'analyse des données dans le système, à savoir celui du corpus aligné et du modèle de langue.

Le deuxième niveau d'analyse est relatif aux données d'entrée et de sortie. Les quatre requêtes de traduction n'existant pas dans le corpus aligné. Le système fait recours au modèle de langue pour faire du mot à mot. Ces quatre propositions traduites ne sont pas exemptes d'erreurs : syntaxico-sémantique et orthographique. Soit les propositions suivantes traduites par le système:

(3)

mĩ kó* Jùámū « je va*au marché ».

bě jó àJúmā* sùkú mū « ils fait travail* à l'école ».

hó ká* Jùá mū àsém* « vous parle* marché affaire* ».

ò kó àhá « Il va chasse* ».

Les items et expressions, comportant des indices, sont les sources des différentes erreurs survenues lors de la traduction. Pour la proposition 1, le modèle de langue produit une traduction dont le rapport syntaxique entre le prédicat verbal et l'argument externe n'est pas respecté. Le verbe aller, conjugué à la première personne du singulier « je », prend la forme « vais » au lieu de « va » dans la langue cible. Pour la proposition 2, en plus de l'absence d'accord entre l'argument externe « bé » (ils) et le prédicat verbal « jó àJúmā » (faire travail), nous notons un problème de syntaxe du prédicat verbal de l'abron au français. Cela se traduit par l'obtention d'un résultat qui n'obéit pas à la structure syntaxique de la langue française. Cette expression se traduit comme suit : « travailler » plutôt que « faire travail ».

L'absence de la marque du pluriel dans la langue cible, quand-à- elle, est du fait que celle-ci n'est pas manifeste dans la langue source. Ceci est transposé dans la langue cible, car la langue source n'admet pas dans ce contexte ce type de rapport. La proposition 3 présente un autre phénomène, celui lié à la description d'une même réalité « parler d'un endroit en bien ou en mal » dans deux langues différentes. Dans ce genre de situation, le mot à mot de « Jùá mū àsém : affaire concernant le marché » ne produit pas de résultat satisfaisant. Une traduction contextuelle serait idéale. Le système pourrait traduire ce syntagme nominal comme suit : « du marché ». Ceci donnerait comme traduction finale « vous parlez du marché ». Pour la proposition 4, l'erreur se traduit en l'absence de préposition « à la » dans la langue source qui est transposée dans la langue cible qui n'admet pas cela. Nous pouvons conclure cette analyse sur les données de sortie que les erreurs constatées sont du ressort des spécificités ou des contraintes de la langue source. Cela s'explique par le fait de vouloir coute que coute appliquer les mêmes normes de traitement des langues indo européennes aux langues africaines. Ceci revient donc à ignorer ces spécificités qui se révèlent être la clé sans laquelle tout traitement serait approximatif. Moufoutaou ADJERAN & Justine BASSABI 2024, p12). Nous procédons, dans les lignes qui suivent, à l'interprétation de ces contraintes.

2. Les contraintes de la langue source et problèmes posés dans la langue d'arrivée

Nous appelons contraintes, les spécificités de la langue de départ (abron) qui parfois entraînent des imperfections dans la langue d'arrivée. Celles-ci sont loin d'être exhaustives dans cette étude. Nous ne présentons que les plus courantes et celles relevées dans le corpus aligné. Chaque contrainte posée est source de problème de

traduction dans la langue cible. Ces contraintes sont généralement d'ordre lexical, morphologique et syntaxique.

2.1. Les contraintes lexicales

Nous rangeons dans cette section les contraintes liées aux unités lexicales de la langue source. Nous relevons, à partir du corpus aligné, des contraintes liées aux locatifs, aux lexèmes polysémiques.

2.1.1. Les contraintes liées aux locatifs dans la langue source

Indiquant le lieu, l'endroit où se déroule l'action, les locatifs se manifestent différemment selon les langues. Dans les langues kwa, ils sont généralement constitués de particule postposée à valeur locative et d'un nom, formant ensemble un complément de lieu. L'analyse du corpus parallèle proposé, ainsi que plusieurs autres exemples, nous permet de constater que les locatifs apparaissent de deux manières en abron : d'une part, une construction où la « postposition à valeur locative », Kossonou (2015, p 339), est présente en surface, et d'autre part, une construction où elle est absente. Soit les énoncés suivants :

(4)

E₁. Ákȳá á kó Jȳá
Akua/agr+acc+être/aller /marché
« Akua est allée **au** marché »

E₂. Ádù jā fjé
Adou/acc+être/maison
« Adou est **à la** maison »

E₃. Kósùnú á kò àhá
Kossonou/agr+acc+être/aller/chasse
« Kossonou est parti **à la** chasse »

E₄. òmòfirá kínum bá kò súkù
Enfants/les/agr+acc+être/aller/ école
« Les enfants sont allés **à l'**école »

E₅. Ájwá tí fjé
Adjoua/ acc+asseoir /maison
« Adjoua est assise **à la** maison »

E₆. mí kó ájȳáribiré
1^{er} PS/ aller+prog/douche
« Je vais **dans la** douche »

E₇. mí kó ájȳáribiré **mū**
1^{er} PS/ aller+prog/douche/dans

« Je vais **dans la** douche »

E₈. nípà tí fjé **mū**
quelqu'un /vivre+acc /maison/**dans**
« quelqu'un vit **dans** la maison »

E₉. jé kó Jȳá **mū**
1pers.pl/aller+prog/marché/**dans**
« Nous allons **dans le** marché ».

E₁₀. ò tí ní tí **sū**
3^e pers.Sg/asseoir+acc/sa/ tête/**sur**
« Il est assis **sur** sa tête »

E₁₁. Ámà ò pírá Járímū
Ama/ balayer+prog/cuisine/**dans**
« Ama balaie la cuisine »
« Ama balaie **dans** la cuisine »

E₁₂. òmòfirá kínum bá kò súkù**mū**
Enfants/les/ acc+aller/ école/**dans**
« Les enfants sont allés **dans** école »
« Les enfants sont allés **à l'**école »

Il ressort de ces énoncés qu'en structure de surface, la postposition locative peut prendre la forme d'une particule grammaticale (*mū*) ou d'un lexème (*sū*). Le *mū* traduit tantôt « au » (à + le) et tantôt « dans » en français, la langue cible. Ainsi, une même postposition locative peut correspondre à deux prépositions locatives différentes. Cela nous amène à poser la question suivante : à quel moment la postposition *mū*

renvoie-t-elle à la préposition « à le » ou à la préposition « dans » ou encore « en »? Nous pouvons le constater dans les énoncés E_{9&11}. Alors, cela pose le problème d'une postposition locative en abron pour plusieurs prépositions locatives en français.

Contrairement à la langue source, qui admet l'absence de certaines postpositions locatives, la langue cible, impose la présence systématique de toutes les prépositions locatives. L'analyse des énoncés E₁, E₂, E₃, E₄ et E₅ nous permet de constater que, dans la langue cible, apparaissent systématiquement les prépositions « au », « à la » et « à l' », ce qui n'est pas le cas en abron. En d'autres termes, les nominaux locatifs ne sont pas toujours introduits par des postpositions locatives. En effet, ils possèdent intrinsèquement des sèmes locatifs. Les locuteurs dans ce cas procèdent à l'économie syntagmatique qui consiste à réduire le nombre de monèmes ou de syllabes d'un syntagme, Martinet (1996, p 176). La langue nous fait l'économie de certains mots quand cela est nécessaire. Dans les exemples E₁, E₂ et E₃, le locuteur est en droit de nous faire l'économie de deux monèmes (postposition locative et déterminant) des syntagmes postpositionnels locatifs respectifs : *Jyá*, *àhá* et *fjé*. Soit l'emploi de ses mêmes syntagmes dans d'autres propositions.

(5)

E₁₃ Akya ti fjé **kì mū** « Akua habite dans la maison »

E₁₄ Adū dè jā *àhá* **kì mū** « Adou est toujours dans la chasse »

E₁₅ Apiá dè kó jyá **mū kì** « Appia continue d'aller au marché »

Dans ces énoncés construits avec les mêmes lexèmes, le contexte impose la présence obligatoire des postpositions et des déterminants. Ici, ces lexèmes apportent des informations qui contraignent leurs présences. Dans ces énoncés par exemple, une situation antérieure devrait empêcher le sujet à demeurer dans la situation actuelle.

2.1.2. Les contraintes liées à la morphologie des prédicats verbaux

Les études sur l'abron (Kossonou (2015), Morifié Tano (2024)) démontrent qu'il s'y trouve deux catégories de verbes : simple et complexe. Dans la catégorie complexe, les verbes sont « construits à partir de plusieurs unités lexicales composées de : un verbe + un nom, un adjectif, une postposition, etc. Chacun de ces constituants est pris dans son entièreté pour être associé l'un à l'autre », Morifié Tano (2024, p 144). Soit les exemples suivants :

(6)

<i>bú ànīm</i>	<i>bú kótóJyé</i>	<i>dí ànīm</i>	<i>cùá kòtié</i>
Casser/visage	casser/genou	manger/visage	couper/pénis
« devancer »	« s'agenouiller »	« devancer »	« circoncire »
<i>bò nì tīmu</i>	<i>bú àní cǐé</i>	<i>əwí àní sw̄</i>	<i>bò ñzà mū</i>
nouer/tête/dans	casser/œil/derrière	taper/œil/ sur	taper/main/dedans
« imaginer »	« lorgner »	« gifler »	« applaudir »

Dans l'approche statistique où le modèle de langue n'est rien d'autre que du mot à mot et de leur occurrence, les résultats produits en terme de traduction, pourraient ne pas correspondre aux exigences syntaxiques et sémantiques de la langue cible (français).

Alors, pour *bú kótóJùé*, le système traduira « casser genou » plutôt que « s'agenouiller », et cela résulte des exigences de nos langues : très regardant sur les détails de l'action. Alors comment différencier les verbes simples des verbes complexes dans le modèle de langue pour une traduction réussite ?

2.2. Les contraintes liées au système de conjugaison de la langue source

Les langues kwa présentent un système de conjugaison très singulier. Dans la table de conjugaison, le verbe ne varie pas de la première personne du singulier à la dernière personne du pluriel. De plus, le pronom personnel-sujet peut subir des modifications selon le mode.

2.2.1. L'invariabilité du verbe

L'abron fait recours à sept modes dans son système de conjugaison. Le verbe, quel que soit le mode ou la forme, ne varie pas, mais subit une modification morphologique en passant de la forme positive à la forme négative en initiale. Autrement dit, Il ne s'accorde pas en nombre comme c'est le cas dans la langue cible (français). Soit les exemples suivants :

Pour mieux rendre compte des faits, nous allons conjuguer les verbes *kó* « aller » et *dí* « manger » (mode habituel) sans les pronoms sujets.

(7)

<i>kó</i> « aller »	<i>dí</i> « manger »
<i>kó</i> « vais »	<i>dí</i> «mange »
<i>kó</i> « vas »	<i>dí</i> «manges »
<i>kó</i> «va »	<i>dí</i> «mange »
<i>kó</i> «allons »	<i>dí</i> «mangeons »
<i>kó</i> «allez »	<i>dí</i> «mangez »
<i>kó</i> «vont »	<i>dí</i> «mangent »

Le modèle de corpus aligné standard pourrait-il traduire correctement chaque verbe sans le pronom sujet comme c'est le cas dans la langue cible ? Pour dire autrement, dans le modèle de langue basé uniquement sur le mot à mot, le système ne saurait traduire fidèlement de la langue source à la langue cible. Cela engendre donc un problème de traduction des verbes conjugués.

2.2.2. Le pronom personnel sujet et objet

Dans les langues kwa en général, les pronoms personnels sujets varient morphologiquement en nombre et en personne, Kossonou (2015, p 425), Yangra (2008, p 260). Les formes sont complètement distinctes du singulier au pluriel, contrairement au verbe. Les pronoms objets, quant à eux, sont morphologiquement identiques aux pronoms sujets. En effet, à la différence de leur position dans l'énoncé, il n'existe pas de distinction morphologique entre les pronoms sujets et objets, à l'exception de la troisième personne du pluriel qui présente deux formes distinctes (*bé* et *bé*) Kossonou (2015, p 429). La figure ci-dessous illustre ce propos.

(8)

	Pronoms sujets		Pronoms objets
SINGULIER	1 ^{er}	m̃ / m̃ « je »	m̃ / m̃ « moi/me »
	2 ^e	ò « tu »	ò « toi, te »
	3 ^e	ó / ó « il/elle »	ñ « lui, le »
PLURIEL	1 ^{er}	jé « nous »	jè « nous »
	2 ^e	Hó « vous »	hó « vous »
	3 ^e	Bé « ils/elles »	bé « les, eux »

Tableau 1. Source : Kossonou (2015, p 430)

Une seule question se pose à l'observation de ce tableau : comment le modèle mot à mot pourrait-il éviter la confusion entre les pronoms sujets et objets hors contexte?

2.3. Les contraintes syntaxiques

La liste des contraintes syntaxiques n'est pas exhaustive dans cet article. Nous en abordons deux à savoir : la position syntaxique du marqueur de la négation par rapport au verbe et celle des déterminants et compléments par rapport au nom.

2.3.1. La position syntaxique du marqueur de la négation par rapport au verbe

La négation, selon Creissels (2006 (b), p 136) désigne le

Mécanisme morphosyntaxique qui, appliqué à une phrase assertive, donne une phrase se distinguant de la première à la fois par une inversion de la valeur de vérité et par un ensemble de propriétés syntaxiques dont sont dépourvues les phrases dans lesquelles ne peut être décelée aucune espèce de négation.

La négation est ainsi un morphème qui dans une phrase change les conditions dans lesquelles elle est vraie. Ce qui lui permet de se démarquer de la première proposition. Dans les langues kwa, la négation se manifeste par un morphème nasal homorganique (N) pouvant prendre trois variantes : n, ŋ et m qui se préfixe généralement à la base verbale. Illustrons nos propos avec les exemples suivants :

(9)

- Conjugaison positive/négative de kó « aller »

m̃	kó	« je vais »	m̃	ŋgó	« je ne vais pas »
wú	ó kó	« tu vas »	wú	ó ŋgó	« tu ne vas pas »
ó	kó	« il va »	ó	ŋgó	« il ne va pas »
jé	kó	« nous allons »	jé	ŋgó	« nous n'allons pas »
hó	ó kó	« vous allez »	hó	ó ŋgó	« vous n'allez pas »
bé	kó	« ils vont »	bé	ŋgó	« ils ne vont pas »

- Conjugaison positive /négative de dí « manger »

m̃	dí	« je mange »	jé	dí	« nous mangeons »
wú	ó dí	« tu manges »	hó	ó dí	« vous mangez »
ó	dí	« il mange »	bé	dí	« ils mangent »

mí	nní	« je ne mange pas »	jé	nní	« nous ne mangeons pas »
wú	ó nní	« tu ne manges pas »	hú	ó nní	« vous ne mangez pas »
ó	nní	« il ne mange pas »	bé	nní	« ils ne pas mangent »

- Conjugaison positive/négative de brà « venir »

mí	bà	« je viens »	mí	nnmá	« je ne viens pas »
wú	ó bà	« tu viens »	wú	ó nmmá	« tu ne viens pas »
ó	bà	« il vient »	ó	nnmá	« il ne vient pas »
jé	bà	« nous venons »	jé	nnmá	« nous ne venons pas »
hú	ó bà	« vous venez »	hú	ó nmmá	« vous ne venez pas »
bé	bà	« ils viennent »	bé	nnmá	« ils ne viennent pas »

Le morphème de la négation ne se contente pas de se préfixer uniquement au verbe, mais nasalise l'initiale consonantique (b, k, d) du verbe ou encore sonorise celles sourdes. Un tel phénomène pourrait, dans le modèle mot à mot, entraîner un problème de reconnaissance de la négation, car ce phénomène de nasalisation ne s'applique pas uniquement qu'à la négation. Il se manifeste également dans la formation du pluriel des mots en abron. Soit les exemples suivants :

(10)

Singulier

Pluriel

E₁ àbòfirá kí á kò súkù

nnmòfirá kínum bá kò súkù

« L'enfant est parti à l'école »

« Les enfants sont allés à l'école »

E₂ biá kí á bè fá àkòkò kí

nnmiá kínum bá bè fá ñgòkòkò kínum

« La femme est venue prendre son poulet » « les femmes sont venues prendre leurs poulets »

E₃ Atándárié á tùtù dadié kí ási

ndáñdárié num bá tùtù ñnadié kínum ási

« Une perdrix a creusé sous le fer »

« Des perdrix ont creusé sous les fers »

Le morphème nasal homorganique (N), avec pour variantes : n, ñ et m, marqueur du pluriel, nasalise ou sonorise la première consonne des items en gras lors de la formation du pluriel.

Alors comment le modèle de langue pourrait-il distinguer la négation du pluriel ? De même, le mot à mot risque de produire des phrases incorrectes, car le système tel que conçu ne saura faire la distinction entre ces deux.

2.3.2. La position des déterminants et compléments du nom par rapport au nom

Les déterminants sont généralement postposés aux déterminés dans les langues kwa. Qu'il soit défini ou indéfini, le déterminant vient toujours après le nom qu'il détermine. Les études sur ces langues montrent que celui-ci n'est pas toujours marqué en surface. C'est le cas des indéfinis. En abouré et abron par exemple, il n'y a aucune marque formelle pour l'indéfini. Le nom est utilisé sans déterminant au niveau segmental, Yangra (2008, p 94), Kossonou (2015, p 293). Selon Abney (1987), Longobardi (1994), tout nominal est une projection d'une tête comme déterminant (Yangra (2008, p 98)); cela dit, le nominal (sans déterminant) est un DP dont la tête est un déterminant à morphème zéro, symbolisé $\emptyset$ par convention. C'est ce que Chomsky (1965, p 108)

appelle « the empty determiner » (déterminant nul ou vide). Illustrons nos propos avec les exemples ci-dessous.

(11)

Kwàkú cí kùsjé Ø òhè
 Kouakou/attraper+acc/rat/un/aujourd'hui
 « Kouakou a attrapé **un** rat aujourd'hui ».
 òmòfírá kínòm bè wí àkókò mùg Ø
 enfants/les/Agr/manger+acc /poulet/entier/un/
 « Les enfants ont mangé **un** poulet entier ».

La position qu'occupe le symbole, ensemble vide par endroit, est celle que devraient prendre ces indéfinis dans ces énoncés.

Avec le déterminant, nous pouvons rencontrer deux types de difficultés dans le modèle mot à mot. En premier lieu, le système risque de conserver l'ordre kwa (N+D), produisant des phrases non naturelles ou ambiguës en français. En second lieu, le système risque d'omettre les indéfinis dans la langue cible (le français) qui sont obligatoires. En un mot, ces phénomènes engendreront les problèmes d'omission des indéfinis et de conservation de la structure de la langue source. De plus, les déterminants ne varient pas en genre. Ils prennent la même forme quel que soit le genre du déterminé. Avec le modèle classique, il peut y avoir une confusion : usage de « un ou le » en lieu place de « une ou la » vis-versa dans la langue cible qui en fait une distinction claire et nette.

Les compléments du nom quant à eux ne dérobent pas à la structure grammaticale des langues kwa. Ils apparaissent également après le nom (postposition). En abron, le complément peut apparaître avant le nom. C'est le cas par exemple de certains adjectifs. Soit les exemples suivants :

(12)

Fjé késié kì (nom + CN (adjectif) + déterminant).
 Maison/grande/la
 « La grande maison » (déterminant+ adjectif + nom).

 biá késié kì (nom +CN (adjectif)+Dét)
 femme/grosse/la
 « la grosse femme »

 pùáfùò bāníní Ø (nom +CN (nom) +Dét)
 Malpropre/homme/Dét
 « Un homme malpropre » « un malpropre »

 kwàJófùò biá kì (nom +CN (nom) +Dét)
 Paresseux /femme/la
 « La femme paresseuse » « la paresseuse »

Nous constatons à partir de ces exemples que les compléments du nom (adjectifs, et noms) viennent après le nom auquel ils apportent des précisions. Au niveau de l'exemple pùáfùò bāníní Ø, le nom (bāníní) vient compléter le nom pùáfùò pour signifier le sexe masculin. Il vient donc préciser le genre du nom. Là où le français fait la

distinction de genre à partir de l'ajout de suffixe (-eux ou -euse). Ces types de formations sont légions dans la langue source (abron). Le mot à mot pourrait-il rendre les spécificités de chaque langue ? Le modèle rencontre ici les problèmes tels que la position du complément (nom et adjectif) par rapport au nom et celui d'adaptation du genre pour les noms. En effet, pour les expressions comme *pùáfùó bāníní* et *kwàJófùó biá*, le système devrait éviter le mot à mot pour ne pas obtenir des résultats comme « malpropre garçon » et « paresseux femme » qui sont rejetés en Français.

Les analyses sur les différentes contraintes ou spécificités de l'abron, nous ont permis de mettre en lumière les limites du modèle de langue basé sur le mot à mot dans un système de traduction automatique. L'abron, la langue source et le français, la langue d'arrivée sont divergentes à plusieurs niveaux. Alors, Ces différences entraînent plusieurs erreurs de traduction automatique, dont les principales sont celles liées aux contextes et à la syntaxe. Une meilleure gestion de ces divergences entre ces deux langues nous amène à la construction d'un modèle de langue basé sur le découpage par groupe de mot ou syntagmatique des phrases à traduire.

3.Modèle de corpus aligné adapté aux langues kwa

3.1. Composition et fonctionnement

Le modèle de corpus aligné que nous proposons dans cet article repose essentiellement sur les éléments constitutifs du modèle initial. Le corpus aligné, constitué d'exemples de phrases avec leur traduction ou de phrases bilingues, demeure le premier composant du système de traduction automatique basé sur corpus. Il s'agit de données linguistiques bilingues ou multilingues dans lesquelles le système puise lorsqu'une demande de traduction est formulée. Le moteur recherche la phrase saisie par l'utilisateur dans cette base et, une fois trouvée, en renvoie l'équivalent à l'écran. Pour obtenir des résultats de plus en plus satisfaisants, le système doit disposer d'un corpus aligné de grande taille. Cette composante reste intacte dans le nouveau modèle que nous proposons.

A ce premier niveau s'ajoute un deuxième appelé modèle de langue. Dans le modèle précédent, celui-ci se réduisait à une traduction mot à mot. Lorsque le moteur ne trouvait pas la phrase saisie dans le corpus aligné, il se tournait directement vers ce modèle pour effectuer une traduction lexicale, ce qui produisait souvent des résultats insatisfaisants à plusieurs niveaux de la langue. Dans la nouvelle approche que nous proposons,

Le modèle de langue de la nouvelle approche repose, en plus du mot à mot, sur une segmentation syntagmatique. Le mot à mot étant déjà connu, nous décrivons ici la segmentation syntagmatique. Étant la dernière étape du schéma de traduction automatique, cette segmentation consiste à sectionner la phrase en syntagmes (SN, SV, SN+V, V, SP, SN). Il inclut également la lexicalisation verbale (V+ V, V+ N, Adj., et V + Adv.), suivie de leur équivalence dans la langue cible.

Ce procédé permet de mieux gérer les divergences morphosyntaxiques entre les langues kwa et le français, en produisant des traductions plus cohérentes et plus proches de la structure attendue dans la langue cible.

(13)

Données d'entrées

– mĩ tǒǒ òmǒǒ kí - Wá kó Jyá mū – ǒ jó àǒúmǎ sùkú mū - ò kúnù jà

òvùó mū



1- ǒ kó sùkú : Il va à l'école	1- ǒ : il	1- ǒ kó : il va
2- ǒ jó àǒúmǎ òvùó mū : Il travaille au champ	2- kó	2- kó sùkú : va à l'école
3- mĩ tǒǒ Jyá mū : Je vends au marché	3- sùkú : école	3- ǒ jó àǒúmǎ : il travaille
4- bé kǎ mǎmĩ jí àsém Ils parlent de cette femme	4- jó àǒúmǎ : travaille	4- òvùó mū : au champ
5- hǒ ǒ dí òmǒǒ kí : Vous mangez le riz	5- òvùó : champ	5- jó àǒúmǎ òvùó mū : travaille au champ
6- Wá kó wúra mū : Il est parti en brousse	6- mū : au	6- 5-mĩ tǒǒ : je vends
7- mĩ ònĩ távùóó je n'ai pas d'ennemi	7- mĩ : je	7- tǒǒ Jyá mū : vends au marché
8- bé dì mĩ bábà wíási mū ils m'ont envoyé au monde	8- tǒǒ : vends	8- Jyá mū : au marché
9- ò kúnù á kó àhá mon mari est parti à la chasse	9- Jyá : marché	9- bé kǎ : ils parlent
10- hǒ ǒ hé ònó nĩ bābā vous regardez maman et papa	10- bé : ils	10- mǎmĩ jí àsém : de cette maman
11- mĩ, mĩ jà sú òkó moi, je suis en train de partir	11- kǎ : parlent	11- kǎ mǎmĩ jí àsém : parlent de cette maman
	12- mǎmĩ : maman	12- hǒ ǒ dí : vous mangez
	13- jí : cette	13- òmǒǒ kí : le riz
	14- àsém : affaire	14- ǒ dí òmǒǒ kí : mangez le riz
	15- hǒ : vous	15- Wá kó : il est parti
	16- ǒ dí : mangez	16- wúra mū : en brousse
	17- òmǒǒ : riz	17- kó wúra mū
	18- kí : le	18- mĩ ònĩ : je n'ai pas
	19- Wá : il	19- távùóó ø : d'ennemis
	20- kó : est parti	20- ònĩ távùóó ø : pas d'ennemis
	21- wúra mū	21- bé dì mĩ bábà : ils m'ont envoyé
	22- mĩ : moi	22- wíási mū : au monde
	23- ònĩ : n'ai pas	23- dì mĩ bábà wíási mū : m'ont envoyé au monde
	24- távùóó : ennemis	24- ò kúnù : mon mari
	25- dì mĩ bábà : m'ont envoyé	25- á kó àhá : est parti à la chasse
	26- wíási : monde	26- àhá : à la chasse
	27- mū : dans	27- hǒ ǒ hé : vous regardez
	28- ò : mon	28- ǒ hé ònó nĩ bābā : vous regardez maman et papa
	29- kúnù : mari	29- ònó nĩ bābā : maman et papa
	30- á kó : est parti	30- mĩ, mĩ jà sú : moi, je suis en train
	31- àhá : chasse	31- jà sú òkó : suis en train de partir
	32- hǒ : vous	
	33- ǒ hé regardez	
	34- ònó : maman	
	35- nĩ : et	
	36- bābā : papa	
	37- jà : être	
	38- jà sú : être en train	
	39- òkó : de partir	
	Mot à mot	Découpage syntagmatique
CORPUS ALIGNES	MODELE DE LANGUE	

Données de sorties

– je vends le riz — il est parti au marché – il travaille à l'école – mon mari est au champ.

Wá kó Jyá mū– ó jó àJúmā sùkú mū - m kúnù jà òvùó mū

Figure 3: Système de traduction automatique basé sur la linguistique de corpus

Avec ce système, nous avons essayé de traduire les phrases suivantes : *mí tṵṵ ṁmṵṵ kí ; ó jó àJúmā sùkú mū ; Wá kó Jyá mū; m kúnù jà òvùó mū*. N'existant pas dans le corpus aligné, le système a recours au modèle de langue précisément le découpage syntagmatique ou par groupe de mot. Comme résultats, nous avons : « je vends le riz — il est parti au marché – il travaille à l'école – mon mari est au champ ». Le constat est clair. Ces résultats sont plus ou moins satisfaisants que ceux produits par le mot à mot dans le modèle de langue classique.

Dans le cadre de la programmation, l'instruction suivante sera donnée à la machine. Pour toute demande de traduction, qu'elle soit de l'abron vers le français ou du français vers l'abron, le système s'appuiera, en première instance, sur un corpus parallèle (ou corpus aligné). Il procédera à une recherche de la phrase saisie dans la base de données phrastique afin d'en retrouver l'équivalent déjà enregistré. Lorsque la phrase recherchée est présente dans le corpus, le système affichera directement sa traduction correspondante.

En revanche, si aucune correspondance exacte n'est trouvée dans la base de données phrastique, le système déclenchera une segmentation lexicale de la phrase, c'est-à-dire une décomposition en unités lexicales ou en mots. Chaque unité sera alors recherchée dans le modèle de langue, notamment dans le lexique bilingue ou la base de données mot-à-mot, afin d'identifier son équivalent dans la langue cible. Les équivalents obtenus seront ensuite réassemblés pour produire une traduction de la phrase complète.

Toutefois, cette approche mot-à-mot peut entraîner des traductions approximatives, voire erronées, notamment lorsque le sens d'une expression dépend de la combinaison de plusieurs mots. En cas d'insatisfaction liée à la qualité de la traduction produite, le système aura alors recours à un troisième niveau de traitement fondé sur la segmentation syntagmatique. Il consistera à décomposer la phrase en groupes de mots ou en syntagmes porteurs de sens, puis à rechercher chacun de ces segments dans une base de données syntagmatique. Le système identifiera ainsi les équivalents correspondants dans la langue cible et procédera à leur réorganisation afin de générer une traduction plus cohérente et plus fidèle au sens de la phrase source. Ainsi, le processus de traduction reposera sur une approche hiérarchisée en trois niveaux : d'abord la recherche d'une correspondance phrastique dans le corpus aligné, ensuite la traduction lexicale mot-à-mot en cas d'absence de correspondance, et enfin la traduction par syntagmes lorsque la première traduction obtenue présente des insuffisances. Cette stratégie permettra au système de privilégier les traductions déjà attestées dans le corpus tout en disposant de mécanismes de repli capables de traiter les phrases nouvelles ou absentes de la base de données.

3.2. Avantages et inconvénients du modèle syntagmatique

Ce type de modèle vient renforcer le système de traduction automatique existant et permet de résoudre plusieurs problèmes liés à l'opération de traduction. En d'autres termes, il apporte des solutions aux difficultés d'ordre morphologique, syntaxique et sémantique.

Du point de vue morphologique, l'analyse des contraintes propres aux langues kwa, en l'occurrence l'abron, a montré que certains verbes sont en réalité constitués de l'association de plusieurs monèmes. Une traduction mot à mot entraînerait donc des problèmes de correspondance. Avec le nouveau modèle, les verbes sont traduits à partir des actions qu'ils expriment, en les associant, lorsque cela est nécessaire, au groupe sujet. Cette approche permet d'éviter les difficultés liées à l'invariance morphologique des verbes dans la table de conjugaison et facilite l'identification de leur équivalent français, même en l'absence d'un pronom sujet explicite. Désormais, grâce à la traduction syntagmatique, le verbe est d'abord associé à son pronom sujet avant d'être soumis à la traduction mot à mot en cas d'échec.

Du point de vue syntaxique, les pronoms, qu'ils soient sujets ou objets, conservent leur position dans le syntagme, ce qui évite les confusions de certains pronoms ayant la même forme en abron alors qu'ils correspondent à des formes distinctes en français. Cette approche réduit également les ambiguïtés entre la négation et le marqueur du pluriel de certains mots, ainsi que les perturbations liées à la position des déterminants et des compléments du nom. Chaque élément conserve ainsi sa place dans la structure syntaxique. Le modèle syntagmatique renforce ainsi de manière significative le système traditionnel. Il corrige plusieurs erreurs de traduction observées avec l'approche précédente. Son principal inconvénient réside dans le fait qu'il effectue parfois des segmentations peu conventionnelles. En effet, le découpage d'une phrase comportant un verbe transitif après le verbal (sujet verbe + objet) peut paraître inhabituel. Toutefois, malgré cette particularité, le modèle produit des résultats de traduction très satisfaisants.

Conclusion

Les résultats du système de traduction automatique dont le modèle de langue repose sur le mot à mot présentent généralement des erreurs à plusieurs niveaux de la langue cible. L'analyse de ces erreurs nous a permis de comprendre que les imperfections ne sont pas uniquement dues aux contraintes des langues en présence, mais également au modèle linguistique basé sur le mot à mot lui-même, qui n'est pas adapté aux langues kwa. Les différences linguistiques — morphologiques, syntaxiques et sémantiques — nous ont conduits à adapter le système de traduction aux exigences de la langue source et de la langue cible. Ainsi, avons-nous renforcé le système en y intégrant un modèle syntaxique.

L'opération de traduction peut désormais s'effectuer en trois étapes :

- Etape 1 : Recherche dans le corpus aligné : trouver l'équivalent de la phrase saisie au clavier dans une base bilingue préalablement constituée.
- Etape 2 : Traduction mot à mot : le système fait une segmentation mot à mot de la phrase saisie en cas de son absence dans le corpus aligné. Le système recherche ensuite l'équivalent de chaque segment de la phrase dans le modèle de langue mot à mot. Lorsque que le texte de sortie est rejeté, le système passe à l'étape 3.

- Etape 3 : Découpage syntagmatique: en cas d'insatisfaction, le système procède par un découpage par syntagme ou association de syntagme de la phrase. Il recherche par la suite ces syntagmes dans le modèle syntagmatique de la langue.

Avec un corpus de grande taille, le système réduit au maximum les erreurs de traduction et améliore la fluidité des textes de sortie produits. Ainsi, grâce à l'affluence des travaux, de développement de ressources en langues africaines, consacrés de plus en plus à la numérisation des langues africaines, à la conception de dictionnaires électroniques et d'autres ressources et outils du TAL dans et pour ces langues, (MBONING Elvis & Wandji Ornella, (2021, P125)), des corpus gigantesques pourront être constitués afin de pallier le manque de ressources linguistiques numériques pouvant servir à la traduction automatique et en général au traitement automatique des langues africaines.

Références bibliographiques

ABNEY, Steven (1987). *The English Noun Phrase in its sentential aspect*, Phd dissertation, MIT.

ADJERAN, Moufoutaou & BASSABI, Justine (2024). Traitement automatique des langues naturelles africaines : la linguistique, un allié indissociable. *Actes du colloque scientifique international sur les langues maternelles*, Editions du labo gur, Abomey-Calavi.

BOITET, Christian (2007). Corpus pour la TA : types, tailles et problèmes associés, selon leur usage et le type de système. *Revue française de linguistique appliquée*, XII-1(25-38)

CHOMSKY, Noam, (1957). *Syntactic Structures*, Mouton, la HAYE. (Trad. Fr : Structures Syntaxiques) Le Seuil, Paris 1969. Michel Baudeau.

CREISSELS, Denis (2006). *Syntaxe générale - Une introduction typologique* – Tome 2 – La phrase, Paris, Lavoisier.

DJEBAILI, Farida (2011). Corpus bilingues comparables et l'extraction automatique de terminologie bilingue français arabe dans les domaines de spécialité. RML7

HADLEY, James. (2020). *Traduction. Contrepoint*, (4), dossier spécial « Traduction automatique ».

KOSSONOU, Théodore. (2015a). *Description systématique d'un parler kwa : abron mérézon, Etude morpho-phonologique et syntaxique de l'abron de Côte d'Ivoire*, Editions universitaires européennes, Abidjan.

LOUA, Cyrille. (2018) *Les proverbes du bron : aspects morphosyntaxiques et sémantiques*. Linguistique descriptive. Thèse de doctorat unique à l'Université Félix Houphouët-Boigny.

LONGOBARDI Giuseppe. (1994). Reference and proper names. *Linguistic Inquiry*. Vol.25, N°4.

MARTINET, André. (1996). *Éléments de linguistique générale*. Armand colin/Masson, Paris, rue Laromiguière.

MBONING, Elvis & WANDJI, Ornella. (2021). Construire des ressources collaboratives pour les langues peu dotées: une modélisation orientée communauté. *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles*. Lille, France Volume 1 : conférence principale, p. 125-133.

MORIFIE, Tano Kouakou Frédéric (2024). *Etude logico-sémantique de l'abron et enjeux du traitement automatique d'une langue ivoirienne*, Thèse de doctorat unique, Département des Sciences du Langage et de la communication, Bouaké.

YANGRA, Aboa François (2008). *Syntaxe de l'abouré, langue kwa de Côte d'Ivoire*, Thèse de doctorat, Département des Sciences du Langage, Abidjan.

