ROYAUME DU MAROC
Université Mohammed V
- RABAT -
جامعة محمد الخامس
– الرباط –
Faculté des sciences
كلية العلوم
CENTRE D'ETUDES DOCTORALES - SCIENCES ET TECHNOLOGIES

N° d'ordre : 3087

THESE

En vue de l'obtention du : ***DOCTORAT***

Structure de Recherche : Laboratoire de Recherche en Informatique et Télécommunications
Discipline : Sciences de l'ingénieur
Spécialité : Informatique et Télécommunications

Présentée et soutenue le : *19/04/2018* ***par :***

Imane TAGHBALOUT

Traduction automatique de la langue amazighe

JURY

Moulay Driss RAHMANI	*PES, Faculté des Sciences de Rabat*	*Président*
Faouzia BENABBOU	*PES, Faculté des Sciences Ben M'Sik, Casablanca*	*Rapportrice*
Abderrahim EL QADI	*PES, Ecole Supérieure de Technologie de Salé*	*Rapporteur*
Mohamed EL MARRAKI	*PES, Faculté des Sciences de Rabat*	***Directeur de Thèse***
Salma MOULINE	*PES, Faculté des Sciences de Rabat*	*Examinatrice*
Fadoua ATAA ALLAH	*CH, Institut Royal de la Culture Amazighe, Rabat*	*Encadrante*
Siham BOULAKNADEL	*CH, Institut Royal de la Culture Amazighe, Rabat*	*Examinatrice*

Année Universitaire : 2017-2018

Remerciements

Ce travail de thèse a été effectué au sein du Laboratoire de Recherche en Informatique et Télécommunications (LRIT) à la Faculté des Sciences de Rabat (FSR), sous la direction du Monsieur **Mohamed EL MARRAKI,** Professeur à la Faculté des Sciences de Rabat, et l'encadrement de Madame **Fadoua ATAA ALLAH**, Chercheure Habilitée à l'Institut Royal de la Culture Amazighe (IRCAM).

Je tiens, d'abord, à remercier mon directeur de thèse, Pr. Mohamed EL MARRAKI, d'avoir accepté de diriger mes travaux, et de m'avoir intégrée au sein du laboratoire LRIT. Je le remercie également pour son aide, sa confiance et ses précieux conseils. Je profite de cette occasion pour lui exprimer tous mes sentiments de respect et d'estime.

Je tiens particulièrement à exprimer mes profonds remerciements à mon encadrante Madame Fadoua ATAA ALLAH pour avoir toujours été disponible pour m'apporter son aide, me donner ses précieuses directives pour mener à bien ce travail de thèse. Je la remercie aussi pour sa confiance, son soutien, et ses encouragements. Ces quelques mots ne sauraient exprimer mon sentiment de reconnaissance à son égard.

Je tiens également à exprimer mes remerciements les plus sincères aux membres du jury pour l'honneur qu'ils me font d'évaluer mon travail de thèse. Je sais combien ils sont sollicités et je sais mesurer l'honneur qu'ils me font en acceptant de me consacrer leur temps.

Que Monsieur Mr. **Moulay Driss RAHMANI**, Professeur d'enseignement supérieur à la Faculté des Sciences de Rabat, trouve ici l'expression de mes remerciements les plus sincères d'avoir accepté de présider cette thèse.

Mes remerciements vont également à Mme **Faouzia BENABBOU**, Professeur d'Enseignement Supérieur à la Faculté Des Sciences Ben M'Sik de Casablanca d'avoir accepté de rapporter ma thèse et de faire le déplacement depuis Casablanca pour assister à ma soutenance en tant que membre de jury.

Que Mr. **Abderrahim EL QADI**, Professeur d'Enseignement Supérieur à l'Ecole Supérieure de Technologie de Salé, trouve ici mes remerciements pour l'intérêt qu'il a porté à cette thèse en acceptant de la rapporter et de faire partie de mon jury**.**

Je remercie aussi Mme **Salma MOULINE**, Professeur d'Enseignement Supérieur à la Faculté des Sciences de Rabat pour avoir accepté d'examiner mon travail.

Mes sincères remerciements vont également à Mme **Siham BOULAKNADEL**, Chercheure Habilitée à l'Institut Royal de la Culture Amazighe (IRCAM), de m'avoir fait l'honneur d'examiner ma thèse.

Je remercie également tous ceux qui, de près ou de loin, ont contribué à la réalisation de ce travail. Un merci très spécial à Mme Salima EL KOULALI d'avoir accépté de valider la traduction d'un corpus vers l'amazighe.

Par ailleurs, je voudrais témoigner toute ma reconnaissance et ma gratitude aux personnes que je ne saurais jamais remercier assez et à qui je dédie cette thèse. A mes chers parents qui m'ont toujours soutenu tout au long de mes études. Merci à vous, pour vos sacrifices, et soutiens indéfectibles et inconditionnels. A mon cher époux pour ses encouragements, sa compréhension et son appui quotidien. A ma petite Inas pour la joie et l'amour qu'elle nous procurre chaque jour. Par la même occasion, je tiens à remercier mes sœurs et mon frère : Layla, Lamya, et Ilyas, mes beaux parents, ainsi que toute ma famille pour leurs encouragements. Aussi je remercie toutes mes amies, en particulier Fatima Zahra Nejme et Khadija El Gajoui.

Résumé

La mondialisation a influencé considérablement l'essor de l'industrie des langues, spécialement en traduction automatique où la demande ne cesse de croître. Ainsi, les besoins en matière de systèmes de traduction automatique fiables augmentent de plus en plus. L'objectif principal de cette thèse est de réaliser un système de traduction automatique au profit de la langue amazighe dans un contexte national visant à sa promotion et son développement. En fait, l'amazighe est une langue peu dotée qui manque de ressources linguistiques électroniques nécessaires pour toute application du Traitement Automatique des Langues Naturelles (TALN) en général et pour la traduction automatique en particulier. Face à cette limitation en ressources, le choix de l'approche de traduction automatique à adopter n'était pas évident. Certes l'approche à base des statistiques est celle la plus utilisée de nos jours vu ses avantages en termes de rapidité de développement et de facilité de maintenance mais elle reste tributaire de l'existence d'un corpus parallèle de taille suffisamment importante afin de bien entrainer des modèles probabilistes capables de donner de bons résultats de traduction. Partant de ce constat, nous avons proposé dans un premier temps un système de traduction automatique unidirectionnel de l'anglais vers l'amazighe à base de l'interlangue UNL (Universal Networking Language). L'utilisation de cette approche a produit plusieurs ressources linguistiques de valeur très importante à savoir un dictionnaire électronique bidirectionnel multilingue et un corpus parallèle anglais-amazighe. Ensuite, nous avons exploité ce corpus pour constituer, après avoir été enrichi par de nouveaux couples de phrases, le corpus d'apprentissage des modèles probabilistes nécessaires pour assurer la traduction automatique statistique bidirectionnelle de l'anglais vers l'amazighe et vice-versa.

Mots clés : *La langue amazighe, Traitement automatique des langues naturelles, Traduction automatique, Traduction automatique par l'interlangue, Traduction automatique à base des statistiques, Traduction hybride, UNL, Dictionnaire électronique multilingue, Analyse morphologique et syntaxique de l'amazighe.*

Abstract

Globalization has significantly influenced the language industry development, especially in machine translation domain where demands continue to grow. As a result, the need for reliable machine translation systems is growing more and more. The main objective of this thesis is to design an Amazigh multilingual machine translation system for the benefit of Amazigh language. In fact, the Amazigh is an under-resourced language that lacks electronic linguistic resources needed for any Natural Language Processing (NLP) tool development in general and for machine translation in particular. Given this limitation on resources, the adequate choice of machine translation approach is not obvious. Certainly, the statistical based approach is the most used nowadays given its advantages in terms of speed development and ease maintenance. However, it remains dependent on the existence of an important parallel corpus in order to well train probabilistic models able to give good translation results. Thus, we proceed at first to use a rule-based machine translation founded on the UNL interlanguage (Universal Networking Language). The proposed system is unidirectional, it allows translation toward Amazighe language. In this approach, many important linguistic resources have been produced namely a multilingual bi-directional electronic dictionary and an English-Amazigh parallel corpus. In a second time, we have exploited this corpus to constitute, after been enriched with new sentences' pairs, the training corpus for producing language and translation models necessary for English-Amazighe bi-directional statistical machine translation.

Keywords: *Amazigh language, Natural Language Processing, Machine translation, Interlingual-based machine translation, Statistical-based machine translation, Hybrid-based machine translation, UNL, Multilingual electronic dictionary, Morphological and syntactical analysis of Amazighe.*

Liste des acronymes

Adjt	Modificateur (Adjunct)
AP	Syntagme adverbial (Adverbial Phrase)
Comp	Complément
CTS	Etat d'annexion (Construct State)
FEM	Féminin
IRCAM	Institut Royal de la Culture Amazighe
JP	Syntagme adjectival (Adjectival Phrase)
LEX	Catégorie lexicale
LN	Langue Naturelle
LN_C	Langue Naturelle cible
LN_S	Langue Naturelle source
MCL	Masculin
NOM	Etat Libre (Nominative state)
NP	Syntagme nominal (Nominal Phrase)
PAR	Paradigme flexionnel
POS	Catégorie morphosyntaxique

RSC	Représentation Syntaxique Cible
Spec	Spécificateur
SVO	Ordre des mots Sujet- Verbe- Objet
TA	Traduction automatique
TALN	Traitement automatique des langues Naturelles
TAR	Traduction Automatique à base des règles linguistiques
TAS	Traduction Automatique à base des statistiques
UNL	Universal Networking Language
UW	Mot universel (Universal Word)
VP	Syntagme verbal (Verbal Phrase)
VSO	Ordre des mots Verbe-Sujet-Objet

Table des matières

Liste des figures

Liste des tableaux

Introduction générale

Sommaire

1. Contexte général et motivation

La Traduction Automatique (TA), sujet autour duquel s'articulent les travaux de cette thèse, s'inscrit dans le cadre de la discipline du Traitement Automatique des Langues Naturelles (TALN). C'est un domaine de recherche pluridisciplinaire, au carrefour de plusieurs domaines, à savoir : la linguistique, l'informatique, les statistiques et l'Intelligence Artificielle (IA). Il vise la réalisation des systèmes informatiques permettant de manipuler le langage humain dans tous ses aspects. Ils transforment des données linguistiques, qui peuvent être des textes écrits ou des oraux, en des textes traduits, des textes corrigés, des résumés, ou bien en des informations extraites, etc. Ce passage entre le texte en entrée d'un système de TALN et le texte reproduit fait intervenir un ensemble de niveaux de traitement : le traitement phonétique, dans le cas où l'entrée du système est vocale, le traitement morphologique, le traitement syntaxique, le traitement sémantique et pragmatique.

La TA, en tant qu'une application du TALN, connait un essor majeur ces dernières années, surtout pour certains couples de langues comme anglais – français, anglais – chinois, anglais – espagnol. Cependant, il reste encore des efforts à fournir pour les autres langues. En effet, la TA est considérée comme l'une des applications les plus délicates en TALN surtout lorsqu'il s'agit d'une langue peu dotée telle que la langue amazighe, qui constitue la langue objet d'étude de cette thèse. En fait, le projet de cette thèse s'inscrit dans le cadre d'un large mouvement national visant la conservation, le développement et la promotion de la langue amazighe. Plusieurs travaux du traitement automatique de l'amazighe ont été entamés ces dernières années, en l'absence à notre connaissance de vrais travaux sur la TA au niveau national, la chose qui nous a suscité de choisir de travailler sur ce sujet.

2. Contributions proposées

Dans ce mémoire, nous présentons nos contributions majeures visant à doter la langue amazighe d'un système de traduction automatique.

Notre première contribution s'est focalisée sur l'analyse morphologique de l'amazighe, la collecte et la formalisation de ses paradigmes flexionnels. Nous avons implémenté, au total 175 paradigmes flexionnels. La chose qui reflète la complexité de la morphologie de cette langue.

La deuxième contribution porte sur le développement d'un dictionnaire bilingue UNL-amazighe qui constituera le noyau de la mise en place d'un dictionnaire électronique multilingue amazighe-anglais-arabe-espagnol-français de taille de 8827 lemmes.

La troisième contribution se focalise sur la préparation des règles grammaticales de transformation UNL-amazighe qui permettront le transfert sémantique et syntaxique entre la représentation sémantique UNL et le texte amazighe.

A la base de ces deux ressources linguistiques : le dictionnaire bilingue UNL-amazighe et les règles de transformation UNL-amazighe, nous pouvons aussi bien générer via l'UNL des corpus parallèles et traduire depuis n'importe quelle langue, disposant de son enconvertisseur UNL, vers l'amazighe.

Notre quatrième contribution consiste à proposer une méthodologie de TA hybride qui exploite le corpus parallèle anglais-amazighe, généré à partir de l'UNL pour servir, après l'avoir enrichi par de nouveaux couples de phrases, comme corpus d'apprentissage des modèles probabilistes nécessaires pour assurer la TA statistique.

3. Plan de la thèse

Ce mémoire de thèse se décline, en plus de cette introduction et d'une conclusion générale, en trois parties principales :

La première partie présente le contexte général dans lequel s'inscrit l'axe de recherche de cette thèse. Elle comporte deux chapitres : le premier présente un état de l'art des différentes approches de TA. Le deuxième définit la langue amazighe, objet de notre étude, en abordant sa morphologie, sa syntaxe, ainsi qu'un aperçu sur les travaux TALN réalisés pour son compte.

La deuxième partie traite l'approche de TA par l'interlangue UNL. Elle s'articule autour de trois chapitres. En premier lieu, elle introduit l'interlangue UNL, son historique et sa syntaxe dans le troisième chapitre de cette thèse. En second lieu, elle présente la préparation des ressources requises par l'UNL pour assurer la TA. Il s'agit du développement du dictionnaire UNL-amazighe, qui constitue l'objet du quatrième chapitre, et de la mise en place des règles grammaticales transformationnelles UNL-amazighe, expliquée dans le cinquième chapitre.

La troisième partie traite une approche de TA hybride qui consiste à exploiter les corpus parallèles LN_S-LN_C construits et enrichis à base de l'approche linguistique par l'interlangue

UNL, pour apprendre des modèles statistiques capables de générer des traductions relativement de bonne qualité dans les deux sens : depuis LNs vers LNc et vice-versa. A titre d'étude de cas, nous avons traité la traduction pour le couple de langues anglais-amazighe.

4. Publications

❖ **Revues scientifiques Internationales**

- 2017: "Pivot-based multilingual dictionary model for under- resourced languages". International Journal of Applied Engineering Research, Volume 12, Number 20 pp. 10342-10350. (**Scopus**)

- 2016: "Towards UNL based machine translation for Amazigh language". International Journal of Computational Science and Engineering. **DOI: 10.1504/IJCSE.2016.10009693.** In press (**Scopus***)*

- 2015: "Amazigh Noun Inflection in the Universal Networking Language". International Journal of Education and Information Technologies, Volume 9, pp. 122-128. (**Copernicus**)

- **Article soumis**: "Natural language generation from semantic graph structures: A case study of a minority language". Malysian Journal of Computer Science. (**Scopus**)

❖ **Chapitre**

- 2014: "Towards an Amazigh UNL dictionary". In book: Recent Advances in Electrical Engineering and Educational Technologies, International Conference on Systems, Control and Informatics (SCI 2014), 28-30 Novembre, Athens, Greece, pp. 129-132.

❖ **Articles de conférences internationales**

- 2017: "Sentence-aligned parallel corpus Amazigh-English". International Conference on Information and Communication Systems (ICICS), 4-6 Avril, Irbid, Jordanie (**IEEE Xplore**)

- 2016 : " Traduction Automatique multilingue : Déconvertisseur UNL-Amazighe ". Communication orale présentée à la semaine de formation de Nooj en collaboration entre l'agence internationale du traitement automatique des langues naturelles et le laboratoire MISC du l'UIT, 7-11 Novembre, Kénitra, Maroc.

- 2015: "Amazigh verb in the Universal Networking Language". International Conference on Computer Systems and Applications, (AICCSA), 17-20 Novembre, Marrakech, Maroc (**Scopus**)

- 2015: "Amazigh Representation in the UNL Framework: Resource Implementation". Procedia Computer Science Volume 73, pp. 234-241. The International Conference on Advanced Wireless, Information, and Communication Technologies, 5-7 Octobre, 2015, Tunisie (**Scopus**)

❖ **Articles de conférences nationales**

- 2016 : "Vers la construction des ressources linguistiques nécessaires pour la génération de la langue amazighe à partir de l'inter-langue UNL ". 7ème Conférence Internationale sur les Technologies d'Information et de Communication pour l'Amazighe TICAM'16
- 2016 : " Corpus multilingues pour les langues peu dotées ". 7ème Conférence Internationale sur les Technologies d'Information et de Communication pour l'Amazighe TICAM'16. IRCAM, Rabat
- 2016 : "Préparation des ressources linguistiques pour la traduction automatique via l'inter-langue UNL ". 5ème édition des Doctoriales FSR, Rabat, Maroc.
- 2015 : " Vers un système de traduction automatique de la langue amazighe via l'UNL ". 3ème édition des journées scientifiques URAC 2015, N°29
- 2014 : "Traduction automatique de la langue Amazighe". 3ème édition des Doctoriales FSR, Rabat, Maroc
- 2014 : "Etude comparative des approches de traduction automatique ". Journées Doctorales en Technologies de l'information et de la Communication (JDTIC'14)

Chapitre 1 : Etat d’art de la traduction automatique

Chapitre 1

Etat de l'art de la Traduction Automatique

Sommaire

1.1. Introduction

La Traduction Automatique (TA) est l'automatisation du processus de traduction d'une langue naturelle source (LNs) vers une autre langue cible (LNc), en utilisant l'ordinateur sans aucune intervention humaine. Elle se distingue de la Traduction Assistée par Ordinateur (TAO), dont son but est d'aider un humain à effectuer une traduction à l'aide, entre autres, de dictionnaires électroniques et de bases de données terminologiques. Dans ce chapitre, nous présentons un bref aperçu sur l'historique de ce domaine. Puis, nous introduisons les différentes approches utilisées dans la mise en place des systèmes de traduction automatique.

1.2. Bref historique de la TA

La TA est un domaine complexe de la linguistique computationnelle, qui a suscité l'intérêt de plusieurs chercheurs depuis le début de la deuxième moitié du vingtième siècle. En juillet 1949, Warren Weaver a écrit un fameux mémorandum sur la TA, qui a stimulé les débuts de la recherche en ce domaine dans plusieurs universités américaines. En 1954, à New York, IBM en collaboration avec l'Université de Georgetown ont réalisé la première démonstration d'un système de traduction automatique russe-anglais (Dostert, 1955). C'était une expérience à petite échelle de 250 mots seulement et six règles « grammaticales » qui pouvait traduire en anglais une soixantaine de phrases russes soigneusement choisies, principalement dans le domaine de la chimie (Hutchins, 2004).

La recherche en TA a été freinée vers 1966, suite à la sortie du rapport du comité consultatif sur le traitement automatique des langues du gouvernement des Etats-Unis (Automatic Language Processing Advisory Committee, ALPAC). Le comité a conclu que la TA était plus chère, moins précise et plus lente que la traduction humaine. La publication du rapport avait un impact profond sur la recherche en TA aux Etats-Unis. Elle a presque été totalement abandonnée pendant presque deux décennies. Par contre, au Canada, en France et en Allemagne, la recherche a toujours continué.

Dans les années 70, le système Systran a été développé et utilisé par la commission européenne. Le système TAUM-Meteo développé à l'Université de Montréal a été installé pour traduire des prévisions météorologiques anglais–français dans les deux sens (de l'anglais vers le français et du français vers l'anglais). Aujourd'hui, il existe des outils de traduction en ligne, tels que le système Systran, Google translate et Babelfish. Bien qu'il n'y ait pas de système qui assure une traduction de qualité pour un domaine large, ces systèmes fournissent

une sortie utile pour l'accès à l'information. Les travaux menés sur la traduction automatique, avant les années 90, ont été généralement focalisés sur une approche linguistique (à base des règles), utilisant des analyseurs syntaxiques et sémantiques. Cependant aujourd'hui, c'est les approches statistiques fondées sur l'apprentissage automatique à partir des corpus bilingues ainsi que les approches hybrides qui sont de plus en plus adoptées par de nombreux chercheurs. La Figure 1.1 donne un aperçu sur l'évolution de la fréquence de l'utilisation des approches de TA au fil du temps. Nous détaillons chaque approche à part dans les sections qui suivent.

Figure 1. 1 Evolution des approches de TA au fil du temps

1.3. Approches linguistiques

Le « triangle de Vauquois » proposé par (Vauquois, 1968) pour la traduction décrit les différentes architectures linguistiques possibles d'un système de TA. Chaque chemin dans le triangle (*cf.* Figure 1.2) correspond à une architecture linguistique.

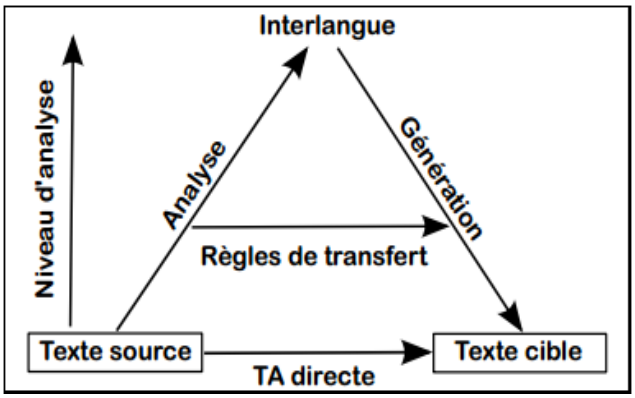

Figure 1. 2 Le « triangle de Vauquois »

Le triangle présente trois architectures linguistiques selon le niveau d'analyse requis. Nous trouvons l'approche directe (ou approche simpliste) en bas du triangle qui ne demande aucune analyse profonde de la phrase source, suivie de l'approche à base de règles de transfert et puis l'approche par l'interlangue au sommet.

1.3.1. Approche simpliste

L'approche simpliste (ou directe) se base seulement sur un dictionnaire bilingue, sans avoir recours à une analyse syntaxique préalable du texte source. Elle se limite à la segmentation du texte source en des mots, puis à les analyser morphologiquement pour obtenir leurs formes de base (lemmes). Ensuite, l'étape du transfert lexical effectue la traduction de chaque lemme vers la langue cible. Finalement, l'étape de génération morphologique, qui pour chaque lemme cible, génère sa forme fléchie correspondante (*cf.* Figure 1.3). Aussi, il peut y avoir des règles de réarrangement, qui permettent de réordonner les mots dans le texte cible (Hutchins et Somers, 1992).

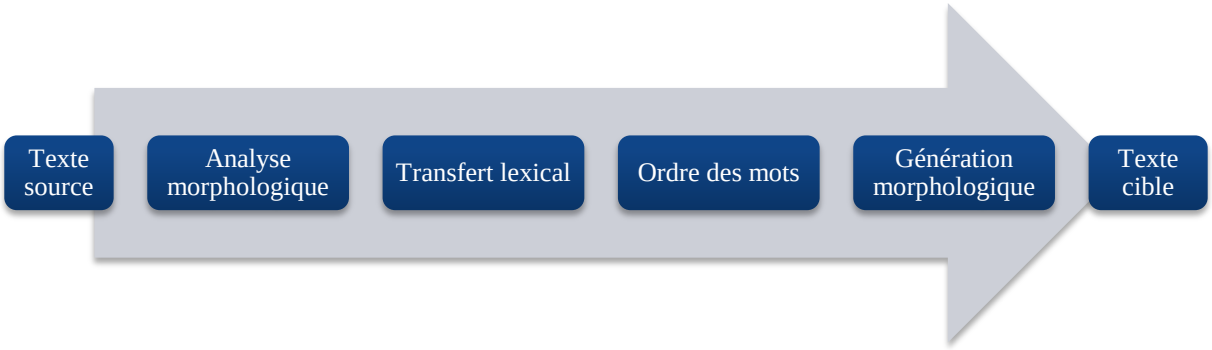

Figure 1. 3 Le processus de la TA suivant l'approche simpliste

Les systèmes de traduction directe sont des systèmes bilingues unidirectionnels. Par ailleurs, le manque d'analyse syntaxique de la phrase source et de connaissance de la construction de la phrase cible est un défaut majeur de ces systèmes, surtout lorsqu'il s'agit de la traduction entre deux langues distantes (par exemple l'arabe et le français). Cependant, ils peuvent donner des résultats acceptables pour le cas de deux langues proches (entre le français et l'anglais par exemple (Arnold *et al.*, 1994)).

1.3.2. Approche par transfert

Le fonctionnement des systèmes par transfert s'articule en trois étapes essentielles : l'analyse, le transfert lexical et de structures, et la génération ou synthèse (*cf.* Figure 1.4).

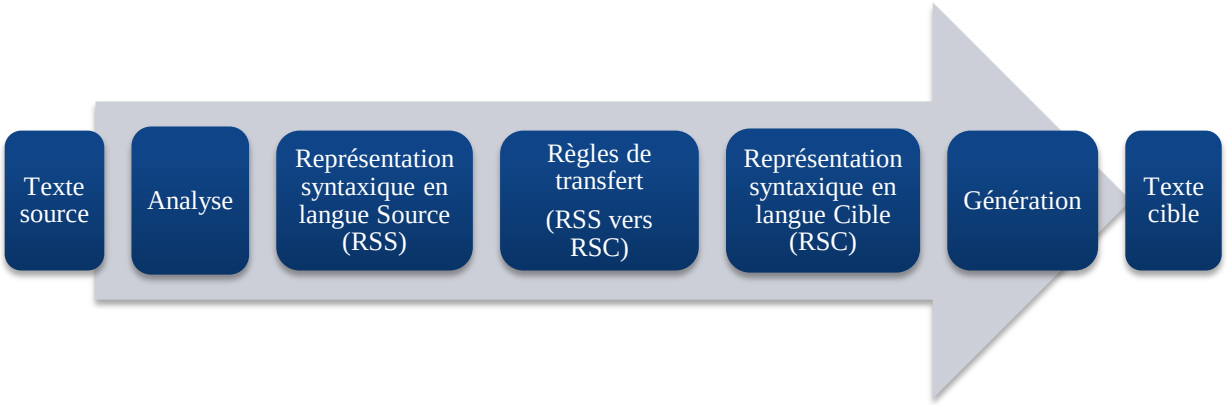

Figure 1. 4 Processus de traduction par transfert

Avec cette approche, le texte source est analysé en se servant d'un dictionnaire et d'une grammaire de la langue source. Pendant cette phase d'analyse, le système procède à des analyses morphologiques et syntaxiques des mots qui donnent lieu à une représentation syntaxique arborescente du texte source. Cette dernière sera à son tour convertie à une autre représentation syntaxique dans la langue cible. Pour assurer ce passage de la représentation syntaxique source (RSS) vers la représentation syntaxique cible (RSC), un ensemble de règles de transfert de structures entre les deux langues sont requises (*cf.* Figure 1.5). Finalement, pendant la phase de génération, le système produit le texte cible correspondant, en se basant sur le dictionnaire et la grammaire de la langue cible.

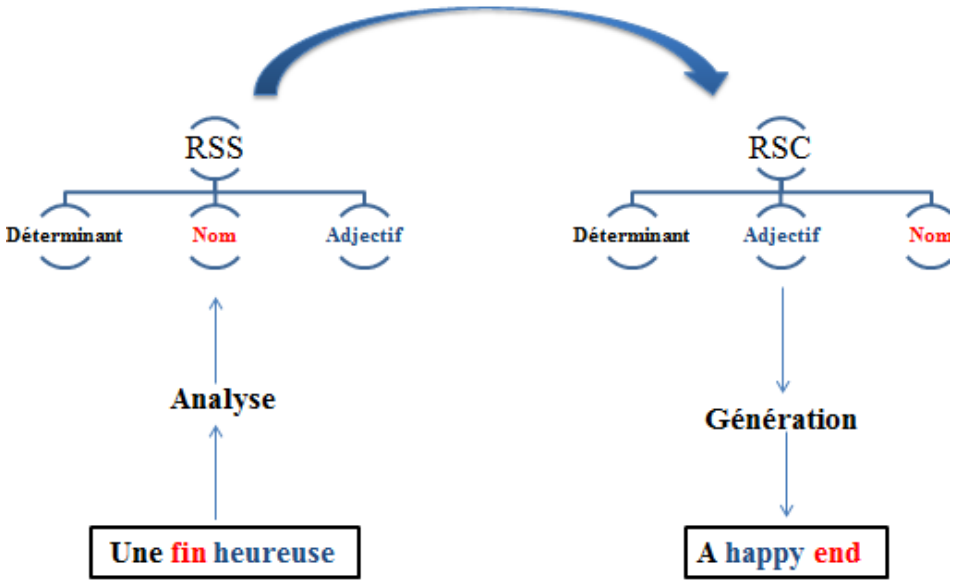

Figure 1. 5 Transfert de structures entre la langue française et la langue anglaise

Parmi les systèmes utilisant l'approche par transfert, nous citons à titre d'exemple le système Systran, fondé par Peter Toma en 1968, qui est considéré parmi les premiers systèmes de traduction automatique utilisant cette approche, le système Apertium (Forcada, 2011) et le système Eurotra, établi et sponsorisé par la commission européenne (Arnold et Des Tombe, 1987 ; Schuts et *al.*, 1991 ; Gehlot *et al.*, 2015).

1.3.3. Approche par interlangue

Dans l'approche par interlangue, le processus de traduction automatique entre une paire de langues se réalise via une représentation sémantique abstraite centrale. Ce processus consiste à analyser le texte source vers une représentation syntactico-sémantique et à générer le texte cible à partir de cette représentation (*cf.* Figure 1.6).

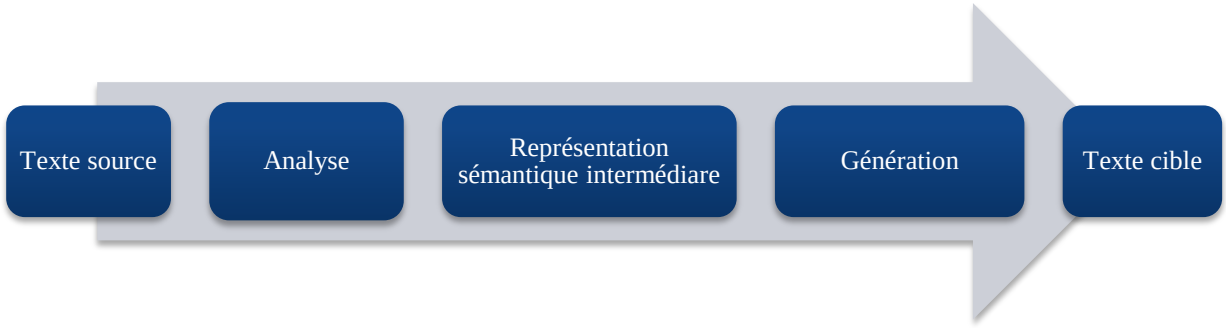

Figure 1. 6 Processus de traduction par interlangue

Par l'utilisation de cette approche, le module de passage de la représentation RSS vers la représentation RSC (*cf.* Figure 1.5) n'aura plus lieu d'être en se comparant avec l'approche par interlangue. La chose qui facilitera le développement des systèmes de TA, notamment les systèmes multilingues, car elle réduit le nombre de modules à développer pour l'ajout d'une nouvelle langue. En effet, pour couvrir tous les sens de traduction entre *n* langues, nous n'avons besoin que de *n* modules d'analyse et de *n* modules de génération (*cf.* Figure 1.6), contrairement à l'approche par transfert qui nécessite le développement de n*(n-1) modules (*cf.* Figure 1.7).

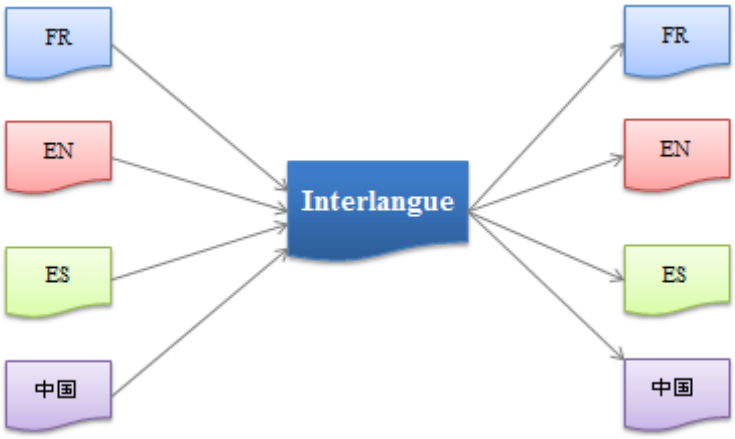

Figure 1. 7 L'approche par interlangue est de complexité 2N

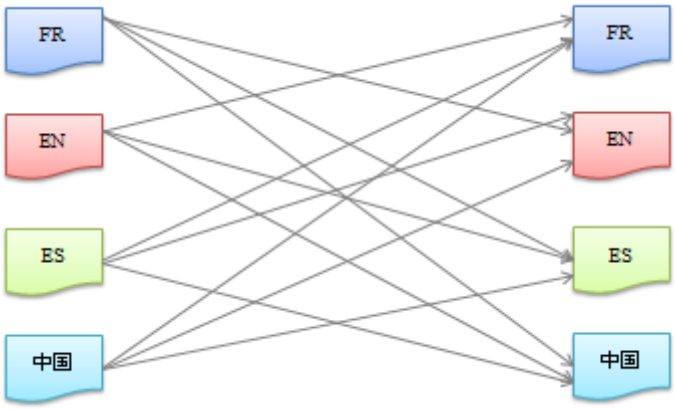

Figure 1. 8 L'approche par transfert est de complexité N*N-1

Cependant, la complexité de cette approche par interlangue réside dans l'obligation de construire un vocabulaire pivot pour représenter tous les concepts possibles de toutes les langues. La construction peut être basée sur une langue artificielle comme l'UNL (Uchida *et al.*, 1999), ou sur une langue auxiliaire « naturelle » (comme l'anglais ou l'espéranto).

Parmi les systèmes utilisant l'approche par interlangue, nous citons à titre d'exemple :

- Le système Pivot : il a été développé par l'entreprise NEC (*Nippon Electric Company*). Il assure une traduction bidirectionnelle entre le japonais et l'anglais (Muraki, 1987).
- Le système KANT (Knowledge-based, Accurate Natural language Translation) : il a été développé à l'Université Carnegie-Mellon (CMU) en Pennsylvanie, USA en 1989

(Nyberg et Mitamura, 1992). C'est un système commercial, utilisé pour traduire les documents techniques anglais vers le français, l'espagnol et l'allemand.

- Le système ATLAS II : il a été développé par l'entreprise japonaise Fujitsu (Uchida, 1989). C'est un système commercial, utilisé pour traduire depuis l'anglais vers le japonais et vice versa.

- Le système DLT (Distributed Language Translation) : il a été développé en 1985 pour la traduction de douze langues européennes (Witkam, 1988).

- Le système UNITRAN (UNIversal TRANslator) : il assure la traduction bidirectionnelle entre l'anglais et l'espagnol (Bonnie, 1987).

- Le système LILY (Language-to-Interlanguage-to-Language System) : il se base sur l'interlangue UNL (Universal Networking Language) pour assurer la traduction multilingue entre les différentes langues participantes au projet UNL (Alansary et Nagi, 2013)

1.4. Approches à base de corpus

Les systèmes à base de règles linguistiques nécessitent l'intervention des spécialistes de la langue, ce qui rend leur développement coûteux en temps et en ressources humaines. Cependant, les approches à base de corpus s'avéraient moins coûteuses en termes d'intervention humaine avec la nécessité de la disponibilité d'une grande quantité de données textuelles. De nos jours, les chercheurs dans le domaine du TALN s'y intéressent de plus en plus et orientent leurs recherches vers ces approches que nous pouvons les distinguer en deux types : un à base des modèles statistiques et l'autre fondé sur les exemples.

1.4.1. Approche à base des statistiques

La Traduction Automatique Statistique (TAS) s'appuie sur l'exploitation mathématique de corpus textuels, dont le volume n'a cessé de croître ces dernières années. Le concept de la traduction automatique statistique a été inventé premièrement au niveau du laboratoire de recherche IBM par (Brown *et al.*,1990). La traduction d'une langue source vers une langue cible à base de l'approche statistique consiste à trouver la phrase cible T qui est la traduction la plus probable de la phrase source S (Brown *et al.*, 1993). Il est à noter que T est une phrase

qui doit répondre à deux critères qui feront d'elle la meilleure traduction de S : T doit être à la fois fidèle au contenu de S et à la fois correcte dans la langue cible.

Dans un premier temps, le problème de la traduction automatique statistique a été modélisé en se basant sur le principe du canal bruité déjà utilisé en reconnaissance automatique de la parole (Shannon, 1948 ; Shannon, 2001) et à l'aide du théorème de Bayes. Ce problème peut être aussi formulé comme suit :

$$T^* = \operatorname{argmax}_T P(T|S) = \operatorname{argmax}_T P(S|T)\, P(T)/P(S) \quad (1.1)$$

Le terme P(S) ne dépend pas de T (il est constant par rapport à T). Il n'a aucune influence sur le calcul de la fonction argmax, donc il peut être écarté du calcul pour obtenir l'équation fondamentale de la traduction statistique suivante :

$$T^* = \operatorname{argmax}_T P(S|T)\, P(T) \quad (1.2)$$

À l'aide de cette équation (1.2), il paraît que la meilleure traduction est déterminée par les probabilités P(S|T) et P(T), qui sont générées indépendamment l'une de l'autre et représentent respectivement le modèle de traduction et le modèle de langue. En effet, l'approche de traduction automatique statistique nécessite trois composants essentiels : *un modèle de langue*, *un modèle de traduction* et *un décodeur*. Ces modèles se basent sur la théorie mathématique de distribution et d'estimation probabiliste de Jelinek (Jelinek, 1969). Le modèle de traduction propose une correspondance en langue cible du texte source. Alors que le modèle de langue valide grammaticalement la traduction dans la langue cible. Finalement, le décodeur statistique (la fonction mathématique argmax) combine ces deux modèles et produit en sortie une traduction T*, qui est la plus probable pour une phrase source S. La Figure 1.9 illustre une architecture générale d'un système de traduction automatique statistique.

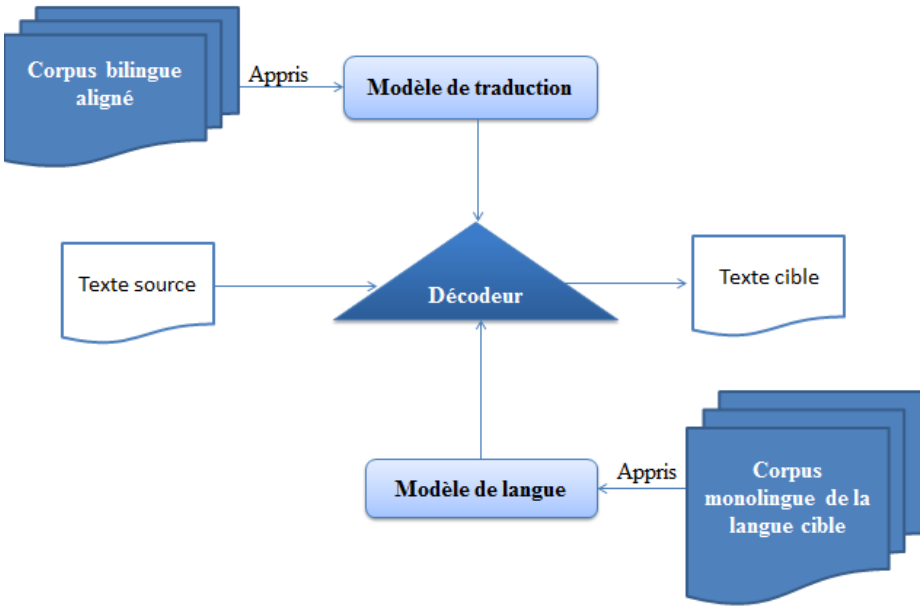

Figure 1. 9 Architecture générale d'un système de traduction automatique statistique

1.4.1.1. Modèle de traduction

La construction d'un modèle de traduction fait appel à la notion d'alignement. Cette notion consiste à mettre en relation les segments des textes sources avec les segments des textes cibles. Dans ce cas, nous parlons d'un corpus parallèle aligné. La prédiction de ces relations de traduction joue un rôle important dans les systèmes de traduction automatique. À partir de ces alignements, une table de traduction est construite pour estimer la probabilité de traduction d'un mot ou d'un groupe de mots de la langue source vers un mot ou un groupe de mots dans la langue cible.

Comme nous avons vu, le modèle de traduction représente la probabilité P(S|T). L'estimation de cette probabilité repose sur l'exploitation de corpus d'apprentissage parallèles alignés. Il existe plusieurs techniques d'apprentissage de modèles de traduction, qui se différencient, notamment, au niveau de l'unité atomique de traduction. Auparavant, les recherches, concernant les modèles de traduction, prenaient le mot comme une unité de traduction. Il s'agit des modèles de traduction à base de mots. Une telle modélisation s'est avérée être limitée, car un mot dans une langue pouvant, par exemple, s'aligner à un ou plusieurs mots dans une autre langue. C'est pour cette raison Koehn, Och, et Marcu ont amélioré ces modèles et ont fondé des modèles de traduction à base de séquences de mots (Koehn *et al.*, 2003). Au plus de ces deux modèles fréquemment utilisés, il existe aussi des

modèles de traduction à base de la syntaxe[1] (Yamada et Knight, 2001) et des modèles dits hiérarchiques[2] (Chiang, 2007).

1.4.1.2. Modèle de langue

Le modèle de langue est largement utilisé dans le domaine du traitement automatique des langues naturelles, nous pouvons le trouver au niveau de plusieurs applications telles que la reconnaissance de l'écriture, la correction d'orthographe, la reconnaissance optique des caractères, la reconnaissance automatique de la parole et bien évidemment la TA. Ce modèle constitue l'une des principales fonctions caractéristiques impliquées dans le processus de traduction automatique statistique. Il s'agit d'un modèle probabiliste qui permet de déterminer la vraisemblance des hypothèses de traduction produites dans la langue cible.

Soit W = $w_1 w_2 \ldots w_L$ une séquence de L mots dans une langue donnée couvrant un vocabulaire fixé V. Les mots qui n'appartiennent pas au vocabulaire V, ils sont considérés comme des inconnus par le modèle. Donc pour modéliser les contraintes d'une langue, les modèles de langue statistiques les plus souvent utilisés attribuent une probabilité à la séquence *w* qui s'exprime comme suit :

$$\mathbf{P(W) = P(w_1)\, P(w_2|w_1)\, P(w_3|w_1w_2) \ldots P\,(w_L|\, w_1w_2\ldots\ldots w_{L-1})} \qquad \mathbf{(1.3)}$$

1.4.1.3. Décodeur

Le processus de traduction qui consiste à transformer une phrase source en une phrase cible est appelé décodage dans le domaine de la traduction automatique probabiliste. Ce terme est inspiré de l'idée de l'ancien cryptographe militaire Warren Weaver qui considérait une phrase en russe comme une phrase en anglais chiffrée, d'où le terme décodage. En traduction automatique probabiliste, l'approche du décodage la plus courante (Wang et Waibel, 1997) est une généralisation de l'algorithme de décodage par piles utilisé en reconnaissance vocale et introduit par Jelinek, en 1969.

[1] Modèle de traduction syntaxique est un modèle construit à partir des corpus alignés et annotés morpho-syntaxiquement pour apprendre les dépendances entre les mots ou bien entre des groupes de mots.

[2] Modèle de traduction hiérarchique est un modèle construit à partir de la combinaison entre le modèle de traduction à base des séquences de mots et le modèle syntaxique.

Le décodeur est la partie centrale de tout système de traduction. Après avoir estimé les modèles de traduction et de langue, la partie décodage est la tâche la plus difficile à effectuer. Elle consiste à chercher la phrase de la langue cible T qui maximise le produit P(S|T)*P(T). Chercher la meilleure traduction d'une phrase source parmi toutes les phrases cibles possibles est un problème NP-complet en traduction automatique (Knight, 1999). Parmi les algorithmes de décodage, il existe l'algorithme de recherche "A*" (Och *et al.*, 2001), l'optimisation linéaire en nombres entiers (Germann *et al.*, 2001), l'algorithme glouton (greedy search en anglais) (Wang et Waibel, 1998) et l'algorithme de recherche en faisceau (beam search en anglais) (Koehn,2004).

1.4.2. Approche à base d'exemples

L'approche de TA fondée sur les exemples a été proposée par Nagao en 1984. L'idée de base de cette approche est de réutiliser des exemples de traductions existantes comme base pour la nouvelle traduction. La base d'exemples collectée contient un ensemble de phrases en langue source associées à leurs traductions en langue cible. Le processus de traduction d'une phrase source est basé sur la correspondance entre cette phrase et la base d'exemples. Le processus de TA fondé sur les exemples se décompose en trois étapes :

- **Etape 1 - Trouver les correspondances** :

Pour chaque phrase source à traduire, les phrases « exemples » les plus proches dans la base d'exemples sont sélectionnées. La fonction la plus importante dans ce type de système est de savoir comment trouver la similitude de la phrase à traduire et une phrase « exemple ». Il existe plusieurs méthodes de sélection basées sur la capacité de remplacer des mots correspondants entre deux phrases, des listes de synonymes, ou basées sur la similarité au niveau syntaxique ou sémantique, ou encore basées sur des informations statistiques. Si aucune phrase « exemple » correspondante à la phrase source entière n'est trouvée, la phrase source à traduire sera segmentée en plusieurs fragments, dont des fragments similaires correspondants existent dans la base d'exemples.

- **Etape 2 - Aligner les phrases** :

La phrase à traduire et les phrases exemples diffèrent sur un ou plusieurs mots. L'alignement est utilisé pour identifier les parties des phrases qui peuvent être réutilisées dans la traduction de la phrase source. Les autres parties qui diffèrent entre la phrase source et les phrases exemples peuvent être traduites en utilisant un dictionnaire bilingue.

- **Etape 3 - Combiner** :

Dans cette dernière étape, les traductions des parties réutilisées et non réutilisées sont assemblées d'une manière logique pour créer la traduction de la phrase source entière. La traduction est extraite à partir de véritables exemples et la qualité de traduction est garantie. Le tableau (*cf.* Tableau 1.1), présente, par un exemple, les étapes suivies pour traduire la phrase française : « Il a acheté des livres chez Fnac » en se basant sur une base d'exemples.

Tableau 1. 1 Traduction français – anglais fondée sur les exemples

Etape 1	• *Il a acheté* une jolie montre. • Il vend *les livres chez fnac*.
Etape 2	• ***Il a acheté*** une jolie montre →***he bought*** a pretty watch • Il vend ***des livres chez fnac*** → He sells ***books at fnac***
Etape 3	***Il a acheté des livres chez Fnac →He bought books at fnac***

1.5. Approche hybride

Récemment, ce type d'approches est un sujet de recherche très actif avec plusieurs types d'hybridation. Les méthodes probabilistes peuvent être intégrées dans le processus de traduction à base de règles linguistiques, et réciproquement. L'intégration peut être simple ou complexe. Par exemple, un système de traduction à base de règles peut utiliser une terminologie extraite empiriquement à partir de corpus parallèles bilingues en même temps que des dictionnaires créés par des experts. Des hypothèses d'un système de traduction à base de règles peuvent être reclassées avec des probabilités lexicales. Par ailleurs, certains modules du système de traduction à base de règles peuvent être remplacés par des modules empiriques (à base de corpus). Par exemple, le module de désambiguïsation lexicale et le module d'extraction de règles de transfert peuvent être construits empiriquement en utilisant un corpus parallèle. Il est aussi possible de prétraiter un corpus parallèle, et d'apprendre les alignements pour un système de traduction statistique à l'aide d'analyseurs linguistiques experts. Parmi les systèmes hybrides, nous citons :

- Le système proposé par la société SYSTRAN, en 2009, basé sur son propre système à base de règles (Dugast, 2007 ; Schwenk, 2009). Ce système a été classé le premier sur la tâche de traduction de l'anglais vers le français lors de la campagne d'évaluation «Workshop on Statistical Machine Translation ».

- Le système développé par (Simard *et al.*, 2007), il se base sur un système de TAS entraîné sur un corpus parallèle qui se compose, en plus, des références humaines des traductions issues d'un système à base de règles.
- Le système développé par (Tinsley *et al.*, 2008), qui combine l'approche statistique et l'approche fondée sur les exemples.

1.6. Evaluation des systèmes de TA

Le processus d'élaboration d'un système de TA inclut nécessairement une phase d'évaluation pour mesurer la qualité des traductions produites. L'objectif derrière cette évaluation est, d'une part, de pouvoir améliorer les traductions générées et d'autre part de comparer la qualité de traduction entre différents systèmes. Par exemple, les traductions proposées en français par Google (qui se base sur l'approche statistique) et Systran (qui se base sur l'approche hybride) pour la citation d'Albert Einstein suivante : *"Life is like riding a bicycle. To keep your balance you must keep moving."*, sont relativement différentes.

- Traduction proposée par Google : " *La vie est comme faire du vélo. Pour garder votre équilibre, vous devez continuer à vous déplacer*".
- Traduction proposée par Systran: "*La vie est comme monter une bicyclette. Pour garder votre équilibre que vous devez continuer le déplacement*.".

Dans le domaine de la TA, nous pouvons distinguer entre deux types d'évaluation : une évaluation humaine et une évaluation automatique.

1.6.1 Evaluation humaine

Ce type d'évaluation repose sur l'intervention des experts humains afin d'évaluer la qualité d'une traduction proposée par le système de TA. Le jugement des experts se base sur deux critères : Adéquation et Intelligibilité. Ces deux critères sont évalués indépendamment, même si les deux sont évidemment corrélées. Le premier critère vérifie si les deux phrases cible et source ont le même sens (Koehn, 2010). Le deuxième critère permet d'évaluer la phrase proposée indépendamment de la phrase source. Il se focalise simplement sur le degré de compréhension. Nous donnons ci-dessous, à titre d'exemple, les métriques utilisées par les évaluateurs experts lors de la campagne d'évaluation DARPA94 (Defense Advanced Research Projects Agency) (White *et al.*, 1994) :

- **A-score :** est un score reflétant le critère d'adéquation, il est compris entre 1 (intégralité du sens conservé) et 5 (rien à voir entre la phrase source et sa traduction). L'attribution du A-score ne tient pas en compte la bonne construction de la traduction dans la langue cible. Toutefois, une mauvaise construction peut influencer sa valeur. Il est plus aisé d'attribuer un bon A-score si la phrase est bien construite plutôt que s'il lui manque un verbe par exemple ou si les mots ne sont pas correctement ordonnés.
- **F-score :** est un score reflétant le critère d'intelligibilité, il est compris entre 1 (la traduction est comparable à une phrase écrite par un natif de la langue cible) et 5 (la phrase est incompréhensible).

Ces deux scores sont assez subjectifs. Une même phrase peut être jugée de façon totalement différente par deux experts humains. De plus, il semble inapproprié de dissocier le critère d'intelligibilité du critère d'adéquation pour évaluer la qualité d'une traduction, car l'un des critères peut largement influencer l'autre. Bien que ce genre d'évaluation fait intervenir l'être humain comme un point fort à la traduction automatique, il présente tout de même de sérieux problèmes, à savoir le coût extrêmement élevé, la subjectivité ainsi que la non reproductibilité de ces évaluations.

1.6.2 Evaluation automatique

Depuis les années 2000 et suite aux problèmes soulevés par l'évaluation humaine, la recherche s'est orientée vers l'utilisation de mesures automatiques qui permettent d'évaluer les systèmes de TA avec un coût faible et une bonne reproductibilité. Ces mesures reflètent la similarité ou la distance entre une traduction automatique produite par le système et une traduction référence. Nous citons dans ce qui suit quelques mesures utilisées dans le domaine de la TA : le score BLEU, la F-mesure, le WER, le TER, et le METEOR en mettant le focus, dans cette section, sur les deux premières métriques.

1.6.2.1. Le score BLEU

BLEU (BiLingual Evaluation Understudy) est la métrique d'évaluation automatique la plus populaire dans le domaine de la traduction automatique, elle a été proposée par (Papineni, 2001). Le principe de cette méthode est de calculer le degré de similitude entre une traduction candidate générée d'un système de TA et une ou plusieurs traductions références en calculant le produit des précisions K-grammes (*cf.* équation 1.4), avec 1<K<n, n est le nombre maximal

à considérer dans une séquence de mots. Généralement, les chercheurs calculent un BLEU pour n=4.

$$\textbf{BLEU} = \mathrm{BP} \prod_{k=1}^{n} p_k^{w_k} = e^{\ln(\mathrm{BP})} e^{\sum_{k=1}^{n} w_k \ln(p_k)} \quad \textbf{(1.4)}$$

avec :

- $\mathrm{BP} = \min\left(1, \frac{Lc}{Lref}\right)$ **(1.5)**
- $\mathbf{L_c}$ = longueur de la traduction candidate
- $\mathbf{L_{ref}}$ = longueur de la traduction référence
- $\mathbf{p_k} = \frac{\text{nombre de k−grammes correctes}}{\text{nombre total de k−grammes}}$ **(1.6)**
- $\mathbf{w_k} = \frac{1}{n}$

1.6.2.2. La F-mesure

La F-mesure est une mesure traditionnelle basée sur le mot. Elle combine la précision et le rappel qui sont deux métriques qui calculent respectivement le nombre de mots corrects générés par rapport au nombre total de mots générés par le système de traduction automatique et le nombre de mots corrects générés par rapport au nombre total de mots dans la référence.

$$\textbf{F-mesure} = 2\,\frac{\textbf{précision} \times \textbf{rappel}}{\textbf{précision} + \textbf{rappel}} \quad , \quad (1.7)$$

avec :

- **Précision** : le nombre de phrases générées correctement divisé par le nombre total de phrases retournées
- **Rappel** : le nombre de phrases correctes divisé par le nombre de phrases qui doivent être retournées

1.6.2.3. Le score WER

Le score WER (Taux d'erreur du mot) est une mesure dédiée à l'évaluation de la performance des systèmes de reconnaissance de la parole. Cette mesure est dérivée de la distance de Levenshtein (Levenshtein, 1966) mais en agissant, dans ce cas, sur des mots et

non pas sur des caractères. Ce score est également utilisé en TA pour évaluer la qualité d'une traduction hypothèse par rapport à une traduction référence. Il se base sur le calcul du nombre minimum d'opérations (insertion, suppression ou substitution des mots) à effectuer sur une traduction automatique pour la rendre identique à la traduction de référence.

1.6.2.4. Le score PER

Le score PER (Taux d'erreur du mot indépendamment de la position) a été proposé par Tillmann, en 1997. Il compare les mots produits par la traduction automatique avec ceux de la référence sans tenir compte de leurs ordonnancements dans la phrase.

1.6.2.5. Le score TER

Le score TER (Taux d'édition de traduction) calcule le nombre minimal de modifications nécessaires pour qu'une traduction automatique soit identique à une traduction référence. Nous citons parmi les opérations élémentaires : l'insertion, la substitution et la suppression, qui sont comptées aussi par le score WER, et en addition à ces opérations il y a aussi le déplacement d'un groupe de mots.

1.6.2.6. Le score METEOR

Le score METEOR (Méthode d'évaluation de la traduction avec le réordonnement explicite) a été proposé par Banerjee et Lavie, en 2005. Il est basé sur l'alignement entre les unigrammes d'une traduction automatique et ceux d'une traduction de référence. Le calcul de cette métrique se déroule en deux étapes : d'abords trouver le meilleur alignement entre l'hypothèse de traduction et sa référence, ensuite utiliser cet alignement pour établir le score. L'alignement se fait successivement sur les unigrammes de même forme orthographique puis les mots de la même racine et enfin les synonymes présents parmi les mots qui ne sont pas encore alignés. À partir de cet alignement, le score est déterminé.

1.7. Conclusion : Quelle approche de TA choisir pour le cas d'une langue peu dotée ?

Dans ce chapitre, nous avons abordé les différentes approches de TA, qui peuvent être catégorisées de manière générale en des approches : à base de règles, à base de corpus ou hybrides. Si les systèmes de traduction automatique à base des statistiques ont l'avantage de ne pas nécessiter la programmation des règles linguistiques, qui est une tâche coûteuse en

temps et en ressources humaines, ils nécessitent, en contrepartie, de grands corpus parallèles afin de produire des résultats satisfaisants. Pourtant, les textes parallèles sont des ressources rares : la taille est souvent limitée. Il existe peu de paires de langues telles que l'anglais, le français, l'espagnol, l'arabe, le chinois et quelques langues européennes pour lesquelles des corpus parallèles de taille raisonnable sont disponibles (Hewavitharana et Vogel, 2011). Cependant, les langues peu dotées manquent de telles ressources. L'approche à base de règles reste la plus appropriée pour construire des systèmes de traduction automatique pour ce type de langues. En particulier, pour les systèmes de TA multilingues, l'approche par interlangue reste celle la plus adéquate parmi les autres approches à base de règles (*cf.* Figure 1.6 et Figure 1.7). Dans ce projet de thèse, nous avons travaillé avec l'interlangue dite UNL pour développer un système de traduction multilingue pour une langue peu dotée qui est l'amazighe. Nous avons opté pour l'UNL parce qu'elle est une interlangue artificielle, c'est-à-dire elle n'est pas destinée à parler avec mais plutôt à être compréhensible par les ordinateurs contrairement aux autres interlangues tels que : Esperanto, Interlingua, Volapuk, Ido, etc qui sont destinées à être des langues humaines. Ainsi, l'utilisation d'une telle approche de TA linguistique nécessitera au préalable une analyse morphologique et syntaxique approfondie de la langue étudiée qui fera l'objet du chapitre suivant.

Chapitre 2 : Présentation de la langue amazighe

Chapitre 2

Présentation de la langue amazighe

Sommaire

2.1. Introduction

La langue amazighe, connue aussi sous le nom du berbère ou Tamazight (ⵜⴰⵎⴰⵣⵉⵖⵜ en tifinaghe), est une langue afro-asiatique ou chamito-sémitique (Kirsty, 2006). Elle constitue la langue des populations autochtones de l'Afrique du Nord : Maroc, Algérie, Tunisie, Libye, et l'Oasis Siwa de l'Egypte. Egalement, elle est parlée par les populations de certaines régions du Niger, du Mali et du Burkina Faso ainsi que par les communautés amazighes immigrées partout dans le monde (*cf.* Figure 2.1). Cependant, c'est au Maroc et à l'Algérie qui comptent, respectivement, les populations amazighophones les plus importantes 40%[3] et 27,4%[4].

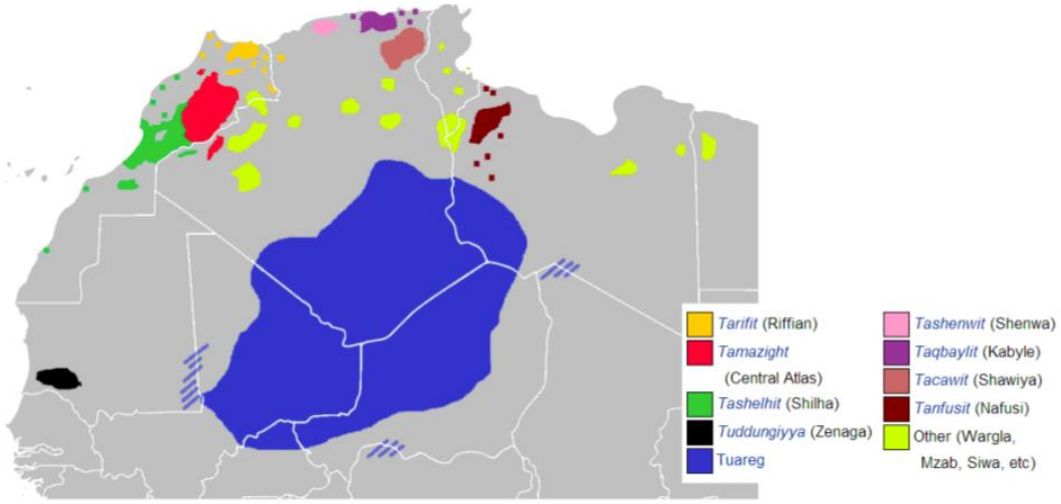

Figure 2. 1 Zones géographiques amazighes

Au Maroc, la langue amazighe se présente sous trois grandes variantes régionales : le Tarifit au Nord, le Tamazight au Maroc central et au Sud-Est, et le Tachelhit au Sud-Ouest et dans le Haut-Atlas. Chacune de ces variantes comprend des parlés locaux (cf. Figure 2.2).

[3] http ://www.axl.cefan.ulaval.ca/afrique/maroc-1demo.htm (visité le 6 octobre 2017).

[4] http ://www.axl.cefan.ulaval.ca/afrique/algerie-1demo.htm (visité le 6 octobre 2017).

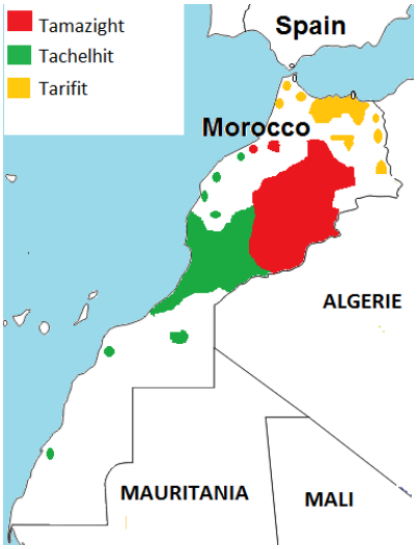

Figure 2. 2 Répartition des variantes de la langue amazighe au Maroc [5]

La langue amazighe est utilisée par une partie importante de la population marocaine dans toutes leurs communications quotidiennes. Cependant, son utilisation n'a pas dépassé le niveau oral jusqu'à la création de l'Institut Royal de la Culture Amazighe (IRCAM), en octobre 2001. L'objectif de cet institut est la promotion de la langue et la culture amazighes. Depuis cette date, l'IRCAM veille à la standardisation de l'amazighe au niveau national pour aménager les variantes de façon à uniformiser les structures morphologiques et grammaticales et en exploitant la variation pour l'enrichissement de la langue. Ce processus de standardisation a abouti à l'élaboration des lexiques, à l'homogénéisation de l'orthographe, à l'élaboration des règles de grammaire (Boukhris et *al.*, 2008) ainsi qu'à le commencement de l'enseignement de la langue amazighe dans plusieurs écoles primaires marocaines. Au niveau de l'enseignement supérieur, des masters et des filières d'études amazighes ont été créés. Au niveau des médias, une chaîne en langue amazighe a été lancée en 2010, par la société Nationale de Radiodiffusion et de Télévision (SNRT). Aussi, la disponibilité de plus en plus de journaux en amazighe tel que le « Monde amazighe »,« Agraw Amazigh » et de « Twiza ».

En outre, depuis 2013 Facebook a adopté la langue amazighe, transcrite par son système d'écriture Tifinaghe, comme langue d'utilisation. Cette initiative vient après l'intégration de l'amazighe dans Windows 8, et par Apple dans son système d'exploitation IOS 9.

[5] https ://fr.wikipedia.org/wiki/Langues berb%C3%A8res(visité le 6 octobre 2017).

En juillet 2011, la langue amazighe est devenue une langue officielle du pays à côté de l'arabe, grâce à la nouvelle constitution dans laquelle il est stipulé, dans son article 5[6], la création d'un conseil national des langues et de la culture marocaine. La mission principale de ce conseil est le développement des langues arabe et amazighe, et les diverses expressions culturelles marocaines.

2.2. Informatisation de la langue amazighe

L'informatisation des langues représente une étape importante et primordiale dans le processus de leur développement. Cependant, cette opération est délicate et coûteuse en termes de temps, notamment dans le cas des langues peu dotées. Elle consiste à doter la langue de toute sorte d'outils et de ressources linguistiques qui garantissent son intégration dans le domaine des technologies de l'information et de la communication, à savoir un système d'écriture, des polices de caractères, un clavier et des ressources linguistiques électroniques (dictionnaires, corpus, ...) ainsi que des applications du TALN, telles que les analyseurs morphologique et syntaxique, les systèmes de recherche d'information et de traduction automatique, etc. Depuis la création de l'IRCAM, et grâce à plusieurs initiatives de recherches scientifiques par certains académiciens marocains, la langue amazighe s'intègre de plus en plus dans le domaine des technologies de l'information et de la communication. En effet, le processus suivi par l'IRCAM dans le cadre de cette intégration a passé par plusieurs étapes, à savoir le codage spécifié par l'ASCII étendu, suivi de la création des polices de caractères tifinaghes, ensuite le codage propre dans le standard Unicode et l'élaboration des normes appropriées concernant la disposition du clavier amazighe, ainsi que le développement des applications du TALN (Zenkouar, 2004 ; Ataa Allah et Boulaknadel, 2014).

2.2.1. Système d'écriture

La langue amazighe a fait ses premiers pas dans le chemin de son informatisation avec l'approbation de l'alphabet tifinaghe, connu sous le nom « tifinaghe-IRCAM » par le

[6] http://www.csefrs.ma/pdf/constitution_2011_FR.pdf (visité le 4 décembre 2017).

consortium Unicode en 2004. Le tifinaghe est l'écriture millénaire de la langue amazighe qui est toujours utilisée chez les amazighes Touaregs[7]. Avec ce système graphique que sont rédigés les anciens écrits au Nord de l'Afrique, depuis la Méditerranée jusqu'au sud du Niger, et depuis les îles Canaris jusqu'aux frontières ouest de l'Egypte. Plusieurs hypothèses, sur l'origine de ce système d'écriture, ont été faites, mais aucune ne peut confirmer son origine exacte. La seule confirmation est que les amazighes « Imazighens » disposaient de leur propre système d'écriture qui remonte à une époque où plusieurs cultures n'en avaient pas encore. Cet alphabet a subi des modifications depuis son origine jusqu'à nos jours, passant du libyque, le tifinaghe saharien et le tifinaghe touareg jusqu'à le néo-tifinaghe. L'alphabet tifinaghe-IRCAM est formé de 33 graphèmes correspondant aux 33 phonèmes de l'amazighe standard qui sont écrits horizontalement et de gauche vers la droite (Ameur et *al.*, 2004).

2.2.2. Etat de l'art des travaux de TALN réalisés

Les recherches menées, au sein de l'IRCAM, ainsi que celles menées au sein d'autres laboratoires universitaires marocains ont montré un changement notable dans le monde de l'informatisation de la langue amazighe. En fait, depuis la création du clavier amazighe, plusieurs projets du TALN ont été réalisés pour l'amazighe, nous pouvons les distinguer entre des projets à vocation la construction des ressources linguistiques électroniques et ceux concernant le développement des applications TALN. Parmi les ressources linguistiques élaborées jusqu'à maintenant, nous pouvons citer à titre d'exemple une base de données terminologique, permettant la compilation et la gestion de la terminologie aménagée par l'IRCAM (El Azrak et El Hamdaoui, 2011) ; un corpus textuel (Boulaknadel et Ataa Allah, 2011) ; un corpus annoté morpho-syntaxiquement (Outahajala et *al.*, 2014); un corpus parallèle aligné par phrase (amazighe-anglais) (Miftah et *al.*, 2017) ; des analyseurs morphologiques (Raiss et Cavalli Sforza, 2012 ; Ataa Allah, 2014 ; Nejme et *al.*, 2016 ; Taghbalout *et al.*, 2015) ; etc. Pour les outils TALN mis en place pour l'amazighe, nous citons le moteur de recherche (Ataa Allah et Boulaknadel, 2010a) ; le pseudo-racineur (Ataa Allah et Boulaknadel, 2010b) ; le concordancier (Ataa Allah et Boulaknadel, 2010c) ; le conjugueur

[7] Les Touaregs, souvent des nomades, sont des habitants du Sahara central et de ses bordures (Algérie, Libye, Niger, Mali, Mauritanie et Burkina Faso). Ils parlent une variante de la langue amazighe et utilisent un alphabet appelé tifinaghe.

des verbes (Ataa Allah et Boulaknadel, 2014) ; le dictionnaire multilingue amazighe-anglais-arabe-espagnol-français (Taghbalout et *al.*, 2017) ; des travaux sur la translittération et la conversion bidirectionnelles des textes écrits en alphabet tifinaghe vers l'alphabet arabe, latin, ou braille (Ataa Allah et Boulaknadel, 2011 ; Ataa Allah et al. 2013 ; Yakoubi, 2016) ; sur la reconnaissance optique des caractères Tifinaghes (Es Saady *et al* ., 2009 ; Es Saady et al ., 2011 ; Aharrane et al., 2015 ; El Gajoui et al., 2015) ; la reconnaissance de la parole (Satori et *al.*, 2014 ; Elouahabi et al. 2016) ; et la reconnaissance des entités nommées amazighes (Talha et *al.*, 2014), etc.

2.3. La morphologie de l'amazighe

L'amazighe est une langue morphologiquement riche avec un système de flexion complexe. En effet, la flexion d'un mot peut faire appel à la fois aux procédés de préfixation, de suffixation et d'infixation. Dans cette section, nous abordons les variations flexionnelles de chacune des catégories grammaticales de l'amazighe. Ataa Allah *et al.* en 2014 ont défini dix catégories, à savoir nom, adjectif, verbe, pronom, numéral, particule, conjonction, interjection, adverbe et préposition. Vu que les cinq dernières catégories sont invariantes, nous faisons le focus sur les autres pour étudier leurs variations morphologiques, en l'occurrence le nom, le verbe, l'adjectif, le pronom, et le numéral.

2.3.1. Nom

Le nom amazighe peut être simple (un seul radical), ou bien composé de plusieurs radicaux. Il varie en genre (masculin/féminin), en nombre (singulier/ pluriel), et en état (libre/état d'annexion).

- **Genre :**

La plupart des noms masculins commencent par ⵄ [a], ⵉ [i], ou ⵓ [u] à l'exception de certains noms, à savoir les noms de parenté comme ⵉⵎⵎⴰ [imma] "maman", qui est un nom féminin commençant par ⵉ [i]. Cependant, les noms féminins sont obtenus par l'affixation du morphème ⵜ [t] au début et à la fin du nom masculin. Par exemple, le mot ⵜⵉⵙⵍⵉⵜ [tislit] "la mariée" est le nom féminin du nom masculin ⵉⵙⵍⵉ [isli] "le marié". Cependant, il existe certains noms féminins qui sont caractérisés seulement par l'ajout d'un morphème ⵜ [t] au début ou à la fin du nom masculin.

- **Nombre :**

Le nom amazighe peut être soit au singulier soit au pluriel. Le pluriel d'un nom amazighe peut prendre plusieurs formes qui sont classés en quatre classes (Boukhris et *al.*, 2008; Oulhaj, 2000):

- Le pluriel externe : est formé par un changement de la première voyelle ⴰ/ⵉ [a/i] accompagné par une suffixation de ⵏ [n] ou bien l'un de ses variants (ⵉⵏ [in], ⴰⵏ [an], ⴰⵢⵏ [ayn], ⵓⵏ [wn], ⴰⵓⵏ [awn], ⵓⴰⵏ [wan], ⵓⵉⵏ [win], ⵜⵏ [tn], ⵢⵉⵏ [yin]).
- Le pluriel brisé : est formé par un changement des voyelles du nom.
- Le pluriel mixte : est formé par un changement de voyelles du nom, accompagné par le procédé de suffixation de ⵏ [n].
- Le pluriel avec ⵉⵏ [id] : est formé par l'ajout du morphème ⵉⵏ [id] avant le nom masculin. Ce type de formation du pluriel concerne un ensemble de noms qui commencent par une consonne, les noms propres, les noms de parentés, les noms composés, les numéraux et les noms d'emprunts.

- **Etat :**

Le nom amazighe est soit dans l'état libre ou dans l'état d'annexion. Un état d'annexion est marqué par un changement de la voyelle initiale. Un nom est dans l'état libre quand il est isolé de tout contexte grammatical alors qu'il est dans l'état d'annexion dans les contextes suivants :

- après le verbe quand il joue le rôle du sujet ;
- une préposition sauf pour le cas de ⴰⵍ / ⴰⵔ [al / ar] et ⵒⵍⴰ [bla] ;
- après un nombre ;
- après ⵉⵏ [id] ;
- après le morphème d'affiliation : ⵓ [u], ⵓⵍⵜ [ult], ⴰⵢⵜ [ayt], ⵉⵔⵜ [ist] (Boukhris et *al.*, 2008).

2.3.2. Verbe

Le verbe, en amazighe, peut être soit simple (un seul radical) ou dérivé (Boukhris et *al.*, 2008). Le verbe simple est composé d'une racine et d'un schème. La racine est une séquence d'une ou plusieurs consonnes, alors que le schème est un modèle de voyelles (V) et de consonnes (C) (Laabdelaoui et *al.*, 2012). Cependant, le verbe dérivé est la concaténation d'un verbe simple et l'un des morphèmes préfixes suivants : ⵙ [s] / ⵙⵙ [ss],

ⵜⵜ [tt] ou ⵎ [m] / ⵎⵎ [mm]. La préfixation de ⵙ [s] / ⵙⵙ [ss] donne la forme (causative) factitive et de ⵜⵜ [tt] donne la forme passive, tandis que le troisième préfixe donne la forme réciproque. Le verbe peut se conjuguer en trois modes : l'indicatif (IND) (*cf.* Tableau 2.1), l'impératif (IMP) (*cf.* Tableau 2.2) et le participatif (PTP) (*cf.* Tableau 2.3). Dans chaque mode, les mêmes indices de personnes sont associés à la forme aspectuelle du verbe (aoriste (AOR), accompli (PFV), accompli négatif (PFV&NEG), inaccompli (NPFV)) obtenue à partir d'un ensemble de procédés morphologiques : alternances vocaliques, préfixation, suffixation, gémination/dégimination de consonnes.

Tableau 2. 1 Indices de personnes du mode indicatif

Nombre	Marques de personnes	Masculin	Féminin
Singulier	1er personne	...ⵖ […gh]	...ⵖ […gh]
	2ème personne	ⵜ…ⴷ [t…d]	ⵜ…ⴷ [t…d]
	3ème personne	ⵉ… [i…]	ⵜ… [t…]
Pluriel	1er personne	ⵏ... [n…]	ⵏ... [n…]
	2ème personne	ⵜ…ⵎ [t…m]	ⵜ…ⵎⵜ [t…mt]
	3ème personne	…ⵏ […n]	…ⵏⵜ […nt]

Tableau 2. 2 Indices de personnes du mode impératif

Nombre	Marques de personnes	Masculin	Féminin
Singulier	2ème personne	Pas de changement	Pas de changement
Pluriel	2ème personne	...ⴰⵜ/ⵜ/ⵎ [.. at/t/m]	…ⵎⵜ [...mt] /…ⴰⵎⵜ [...amt]

Tableau 2. 3 Indices de personnes du mode participatif

Nombre	Singulier	Pluriel
Masculin/féminin	ⵉ…ⵏ [i ... n]	...ⵏⵉⵏ […nin]

2.3.3. Pronom

Le pronom fait référence à tout élément qui pourrait remplacer un nom ou un groupe nominal. Dans la langue amazighe, les pronoms sont soit des :

- **Pronoms possessifs** :

Ils varient selon l'indice de personne, le nombre et le genre. Ils varient aussi selon le genre du possesseur et de l'objet possédé (*cf.* Tableau 2.4).

- **Pronoms démonstratifs** :

Ils varient en genre et en nombre comme le décrit le Tableau 2.5.

- **Pronoms personnels** :

Ils sont divisés en deux : les pronoms autonomes et les pronoms affixes sujet (*cf.* Tableau 2.1, Tableau 2.2, Tableau 2.3), ou objet (*cf.* Tableau 2.6), tous varient en genre et en nombre.

- **Pronoms interrogatifs** :

Ils sont divisés en deux catégories : les pronoms simples invariants (ⴾⵄ [ma], ⵡⵉ [wi], ⵓ [u]), et les pronoms composés, construits par la combinaison entre ⴾⵄ et ⵍⵡⵄ [nwa] ou ⵍⵡⵏ [nwn]. Ils varient en genre et en nombre comme indiqué dans le Tableau 2.7.

- **Pronoms indéfinis** :

Selon (Boukhris et *al.*, 2008), la plupart des pronoms indéfinis sont invariables, seul le pronom ⵡⵄⵣⴹ [wayd] “autres” qui se fléchit en genre et en nombre.

Tableau 2. 4 La flexion des pronoms possessifs

Possesseur	Objet possédé	
	Masculin	**Féminin**
1ère personne du singulier	ⵡⵉⵏⵓ [winu]	ⵜⵉⵏⵓ [tinu]
2ème personne du singulier féminin	ⵡⵉⵏⵏⴾ [winnm]	ⵜⵉⵏⵏⴾ [tinnm]
2ème personne du singulier masculin	ⵡⵉⵏⵏⴽ [wink]	ⵜⵉⵏⵏⴽ [tinnk]
3ème personne du singulier	ⵡⵉⵏⵏⵙ winns]	ⵜⵉⵏⵏⵙ [tinns]
1ère personne du pluriel	ⵡⵉⵏⵏⵖ [winngh]	ⵜⵉⵏⵏⵖ [tinngh]
2ème personne du pluriel masculin	ⵡⵉⵏⵏⵓⵏ [winnun]	ⵜⵉⵏⵏⵓⵏ [tinnun]
2ème personne du pluriel féminin	ⵡⵉⵏⵏⵓⵏⵜ [winnunt]	ⵜⵉⵏⵏⵓⵏⵜ [tinnunt]
3ème personne du pluriel masculin	ⵡⵉⵏⵏⵙⵏ [winnsn]	ⵜⵉⵏⵏⵙⵏ [tinnsn]
3ème personne du pluriel féminin	ⵡⵉⵏⵏⵙⵏⵜ [winnsnt]	ⵜⵉⵏⵏⵙⵏⵜ [tinnsnt]

Tableau 2. 5 Flexion des pronoms démonstratifs

	Masculin		Féminin	
	Singulier	**Pluriel**	**Singulier**	**Pluriel**
Proximité	ⵍⵓⴰ [wad] / ⵍⵓ [wa]	ⵍⵉⵏⴰ [wid] / ⵢⵉⵏⵓ [yina]	ⵜⵓⴰ [tad] / ⵜⵓ [ta]	ⵜⵉⵏⴰ [tid] / ⵜⵉⵏⵓ [tina]
Distance	ⵍⵓⵍⵍ [wann] / ⵍⵉⵏⵍ [winn]	ⵍⵉⵏⵍ [winn] / ⵢⵉⵏⵉⵏⵍ [yininn]	ⵜⵓⵍⵍ [tann] / ⵜⵉⵏⵍ [tinn]	ⵜⵉⵏⵍ [tinn] / ⵜⵉⵏⵉⵏⵍ [tininn]
Absence	ⵍⵓⵎⵉⵏ [walli] / ⵍⵍⵏ [wnni]	ⵍⵉⵏⵎⵉⵏ [willi] / ⵢⵉⵏⵍⵏ [yinni]	ⵜⵓⵎⵉⵏ[talli] / ⵜⵍⵏ [tnni]	ⵜⵉⵏⵎⵉⵏ [tilli]/ ⵜⵉⵏⵍⵏ [tinni]
	ⵍⵓⴰⵓ [wada] / ⵍⵍⵓ [wnna]	ⵍⵉⵏⴰⵓ [wida] / ⵍⵉⵏⵍⵓ [winna]	ⵜⵓⴰⵓ [tada]/ ⵜⵍⵓ [tnna]	ⵜⵉⵏⴰⵓ [tida]/ ⵜⵉⵏⵍⵓ [tinna]

Tableau 2. 6 Flexion des pronoms affixes objets

		Pronoms objet direct		Pronoms objet indirect	
		Masculin	**Féminin**	**Masculin**	**Féminin**
Singulier	1ère	ⵉⵢⵉ [iyi]/ ⵉ [i]	ⵉⵢⵉ [iyi]	ⵉⵢⵉ [iyi]	ⵉⵢⵉ [iyi]
	2ème	ⴽ [k]	ⴽⵎ [km]	ⵓⴽ [ak]	ⵓⵎ [am]
	3ème	ⵜ [t]	ⵜⵜ [tt]	ⵓⵙ [as]	ⵓⵙ [as]
Pluriel	1ère	ⵓⵖ [agh] / ⵓⵍⵖ [angh]	ⵓⵖ [agh] / ⵓⵍⵖ [angh]	ⵓⵖ [agh] / ⵓⵍⵖ [angh]	ⵓⵖ [agh] / ⵓⵍⵖ [angh]
	2ème	ⴽʷⵏ [kʷn]/ ⵍⵏ [wn]	ⴽʷⵏⵜ [kʷnt] / ⵍⵏⵜ [wnt]	ⵓⴽʷⵏ [akʷn] / ⵓⵍⵏ [awn]	ⵓⴽʷⵏⵜ [akʷnt] / ⵓⵍⵏⵜ [awnt]
	3ème	ⵜⵏ [tn]	ⵜⵏⵜ [tnt]	ⵓⵙⵏ [asn]	ⵓⵙⵏⵜ [asnt]

Tableau 2. 7 Flexion des pronoms interrogatifs

Genre	Singulier	Pluriel
Masculin	ⵎⵓⵍⵓ [manwa]/ ⵎⵓⵍⵍ [manwn]	ⵎⵓⵍⵉⵏ [manwi] / ⵎⵓⵢⵏ [manyn]
Féminin	ⵎⵓⵜⵓ [manta] / ⵎⵓⵜⵏ [mantn]	ⵎⵓⵜⵉⵏ [manti] / ⵎⵓⵜⵉⵏⵏ [mantin]

2.3.4. Numéraux

Les numéraux incluent aussi bien les cardinaux que les ordinaux. Les cardinaux se fléchissent en genre à l'exception de ⵜⵉⵎⵉⴷⵉ [timidi] “cent” et de ⵉⵅⵀ [ifd] “mille” qui se fléchissent en nombre et en état. Cependant, les ordinaux se fléchissent seulement en genre à l'exception de ⴰⵎⵣⵡⴰⵔⵓ [amzwaru] “le premier” et ⴰⵎⴳⴳⴰⵔⵓ [amggaru] “le dernier” qui se fléchissent en genre et en nombre. Les autres ordinaux sont des mots composés construits sur la base de ⵡⵉⵙ [wis] ou ⵜⵉⵙ [tis] suivi du cardinal. Par exemple : le féminin de ⵡⵉⵙ ⴽⵕⴰⴷ [wis krad] “le troisième” est ⵜⵉⵙ ⴽⵕⴰⴷⵜ [tis kradt] “la troisième”.

2.4. La syntaxe de l’amazighe

La présente section abordera les principaux aspects de la syntaxe de l’amazighe, à savoir l’ordre des constituants ainsi que les structures de la négation et de l’interrogation en amazighe. En fait, les études menées, sur l’amazighe, ont pratiquement décrit tous ses aspects, mais à des degrés d’importance différents. La morphologie a été beaucoup plus présente dans les travaux de recherche par rapport à la syntaxe et à la sémantique.

2.4.1. Ordre des constituants

La langue amazighe appartient à la famille des langues VSO (Verb Subjet Object), c’est à dire les langues dont le verbe précède le sujet qui, à son tour, précède le complément d’objet direct, en l’occurrence l’arabe, et l’irish (Bentolila, 1981). Les éléments d’une phrase sont arrangés dans cet ordre, ainsi la phrase “mange elle le pain” est une phrase grammaticale correcte pour ces langues. La majorité de ces langues peuvent accepter aussi l’ordre SVO (Subject Verb Object). Dans la langue amazighe, l’alternation entre ces deux constructions VSO et SVO engendre un changement morphologique du nom jouant le rôle du sujet. En effet, son état passe de l’état d’annexion, pour le cas VSO, à l’état libre pour l’ordre SVO. Par exemple, la phrase ⵖⵙⵙ ⴷ ⵙⵙⵍⵎⴰⴷ [yus d uslmad] “L’enseignant est venu ici”, la traduction littérale “est-venu ici l’enseignant” : le verbe est ⵖⵙⵙ [yus] “est-venu”, la particule est ⴷ [d] “ici”, et le sujet est ⵙⵙⵍⵎⴰⴷ [uslmad] “enseignant”. Dans ce cas, le sujet vient après le verbe, ainsi il sera sous la forme de l’état d’annexion, marqué par le changement de la voyelle initiale “ⴰ”, lorsqu’il est dans l’état libre “ⴰⵙⵍⵎⴰⴷ” [aslmad] “enseignant” par “ⵙ”. Cependant, lorsque le sujet ⴰⵙⵍⵎⴰⴷ [aslmad] “enseignant” vient avant le verbe dans la phrase : ⴰⵙⵍⵎⴰⴷ

ⵖⵙⵔ ⴰ, il reste dans son état libre. Cadi (1987) a montré que 78 % des phrases amazighes sont construites avec l'ordre VSO, et seulement 22 % avec la construction SVO.

2.4.2. Structures syntaxiques des phrases déclaratives

Nous présentons le résultat de l'analyse syntaxique que nous avons effectuée sur les constituants de chaque syntagme[8] amazighe, à savoir le syntagme nominal (NP), verbal (VP), adjectival (JP), adverbial (AP) et déterminatif (DP). Pour le syntagme nominal, nous avons identifié 6 règles syntagmatiques, dites aussi les règles de réécritures :

- **NP→ N :** ⵄⴰⵍⵉⵙⵔ [adlis] "livre" (N peut être nom ou pronom).
- **NP→DP NP :** ⵔⵙⵍ ⵙⴰⵍⵉⵙⵔⵍ [sin idlisn] "deux livres".
- **NP→DP PP :** ⵄⵣⵗⵍ ⵏ ⵜⵔⵓⵄⵜ [azgn n tsragt] "demi-heure".
- **NP→ NP PP :** ⵙⵎⵙ ⵏ ⵜⵄⴰⴰⵄⵓⵜ [imi n taddart] "Entrée de la maison".
- **NP→ NP C NP :** ⵄⴰⵍⵉⵙⵔ ⴰ ⵓⵍⵓⵗ [adlis d ulug] "livre et cahier".
- **NP→ NP JP** : ⵜⵉⵀⵉⵔⵜⵓⵍⴽⵉⵜ [tihirit tafulkit] "jolie voiture", sa traduction littérale est : "voiture jolie".
- **NP→ NP VP :** ⵜⵉⵀⵉⵔⵜⵓⵍⴽⵉⵜⵉⵏ ⵉⵣⵉⵍⵏ [tihiritin iziln] "jolies voitures", sa traduction littérale est "voitures jolies".

Pour le syntagme adjectival, nous avons identifié les règles de réécritures suivantes :

- **JP→ J** : ⵄⴼⵓⵍⴽⵉ [afulki] "joli".
- **JP→ JP AP** : ⵉⴼⵓⵍⴽⵉⵏ ⵄⵜⵜⴰⵙ [ifulkin attas] "très beaux", sa traduction littérale est : 'beaux très'.
- **JP→ JP C JP** : ⵜⴰⴼⵓⵍⴽⵉⵜ ⴰ ⵜⴰⵅⴰⵜⴰⵔⵜ [tafulkit d tkhatart] "jolie et grande".

Pour le syntagme verbal, nous avons identifié les règles syntagmatiques suivantes :

- **VP→V** : ⵖⵓⵡⴷ [yuwd] "Il est arrivé ".
- **VP→VP NP** : ⵙⵓⵢⴰ ⵄⵍⵉⵙⵔ [isgha adlis] "Il a achèté le livre".

[8] Un syntagme est un ensemble de mots formant une seule unité catégorielle et fonctionnelle, constituant une unité sémantique, mais dont chaque constituant conserve ses propres signification et syntaxe.

- **VP→VP (NP) AP** : ⴻⵣⵉ ⵣⵉⴽⴽ [ign zikk] “Il a dormi tôt”. Les parenthèses, qui entourent le NP, marquent que son existence est optionnelle. Un exemple de cette structure VP→VP NP AP est ⵢⵉⵡⵛⴰ ⴰⵙ ⵛⵡⴰⵢⵜ [yiwcha as chwayt] “Il lui a donné un peu”. Sa traduction littérale est : “Il-a-donné lui un-peu”.

- **VP→ VP PP** : ⴻⵉⵜⵙ ⵙ ⵜⵣⵔⵔⴻ ⵏⵏⵙ [iftu s tgmmi nns] “Il est allé à sa maison”.
  Sa traduction littérale est : “il-est-allé à maison sa”.
- **VP→ P VP** : ⴰⴷ ⵙⵖⵏ [ad sghn] “Ils achèteront”.

Pour le syntagme adverbial, nous avons identifié les règles de réécritures suivantes :

- **AP→A** : ⵣⵉⴽⴽ [zikk] “tôt”.
- **AP→PP**: ⵙ ⵜⵜⴰⵡⵉⵍ [s ttawil] “doucement”.
- **AP→AP DP** : ⴰⵙⵙⴰ ⴰⵏⵏ [assa ann] “Cette journée-là”.

Pour le syntagme prépositionnel et le syntagme déterminatif, nous avons identifié les règles de syntagmatiques suivantes :

- **PP→P NP** : ⴰⴳⴷ ⵡⵉⵏⵓ [agd winu] “avec le mien”.
- **DP→(NP)D :** ⴰⴷⵏⵉⵙ ⴰⵏⵏ [ann] “Ce livre là”.

2.4.3. Structures syntaxiques des phrases négatives

En amazighe, la phrase négative est caractérisée par l’emploi du morphème de négation ⵓⵔ [ur] “ne... pas”. La structure syntaxique de la phrase négative est :

- Pour la phrase verbale : **ⵓⵔ + VP.** Avec l’antéposition des pronoms personnels objets (Pro. Objet) du verbe dans le syntagme verbal VP. Exemple : ⵓⵔ ⴰⵙⵏ ⵜⵏⵏⵉⵖ [ur asn t nnigh] “je ne le leur ai pas dit”.
- Pour la phrase non verbale : **ⵓⵔ + ⴷ + NP**. Exemple : ⵓⵔ ⴷ ⴰⴷⵏⵉⵙ [ur d adlis] “Ce n’est pas le livre”.

Le morphème ⵓⵔ peut être accompagné d'un deuxième élément de négation, comme ⵓⵔ ⵙⴰⵔ [ur sar] “ne... jamais”, et ⵓⵔ…ⵍⵓⵍⵉ [ur...walu] “ne…rien”, ⵓⵔ…ⵢⴰⵏ [ur...yan] “ne … personne”, ⵓⵔ…ⵢⴰⵜ [ur...yat] “ne …rien”. Nous donnons les phrases suivantes à titre d’exemples : **ⵓⵔ ⵙⴰⵔ ⵙ ⵙⵍⵍⵖ** [ur sar as siwlgh] “je ne lui parlerai jamais”. **ⵓⵔ ⵉⵏⵏⵉ ⵀⵜⵜⴰ ⵢⴰⵏ** [ur iddi Hta yan] “Personne n’est parti”.

2.4.4. Structures syntaxiques des phrases interrogatives

Dans cette section, nous présenterons les différentes structures syntaxiques interrogatives attestées en amazighe. Comme la plupart des langues naturelles, il existe deux types de phrases interrogatives : les questions totales et celles partielles (Boukhris *et al.*, 2008).

2.4.4.1 Questions totales

Les questions totales requièrent soit la réponse oui ou la réponse non. En amazighe, elles sont introduites par l’un des morphèmes suivants : ⵉⵙ [is] “est ce que”, ou ⵎⴰ [ma] “est ce que”. Ils introduisent aussi bien la phrase verbale que la phrase non verbale. Ainsi les structures syntaxiques de l’interrogation totale sont les suivantes :

- ⵉⵙ [is] ou ⵎⴰ [ma] + VP**.** Avec l’antéposition des pronoms personnels objets (Pro. Objet) du verbe dans le syntagme verbal VP (Boukhris *et al.,* 2008). Par exemple : ⵉⵙ ⵜⵏ ⵉⵣⵔⴰ ? [is tn izra ?] “Les a-t-il vus ?”.
- ⵉⵙ [is] ou ⵎⴰ [ma] +ⴷ [d] + NP. Exemple : ⵉⵙ ⴷ ⴱⴰⴱⴰⴽ ? [is d babak ?] “Est-ce ton père ?”.

ⵎⴰ [ma] à la différence de ⵉⵙ [is], lorsqu’il introduit une phrase verbale, il n'entraîne pas l'antéposition des Pro. Objet du verbe. Ainsi, si la structure syntaxique de la question introduite par ⵉⵙ [is] est : **ⵉⵙ + Pro. Objet** + Verbe ? alors celle par ⵎⴰ [ma] est “ⵎⴰ+ Verbe + **Pro. Objet** ? ”.

2.4.4.2. Questions partielles

Les questions partielles requièrent une réponse explicative sur un constituant précis dans la phrase. En amazighe, elles sont introduites par l'un des morphèmes interrogatifs suivants : ⵎⴰⵅ [makh] "pourquoi", ⵎⴰⵏⵉ [mani] "où", ⵎⵉⵍⵎⵉ [milmi] "quand", etc. Ces morphèmes peuvent être des pronoms interrogatifs, des adverbes interrogatifs ou des adjectifs interrogatifs (Boukhris *et al.,* 2008). Peu importe le morphème interrogatif, une question partielle amazighe peut suivre les structures syntaxiques suivantes :

- **Morphème interrogatif + PP.** Par exemple : ⵎⴰⵏⵉ ⵙ ⵉⴷⴷⴰ ? [mani s idda ?] "où est-il parti ?".
- **Morphème interrogatif +VP.** Par exemple : ⵎⴰⵏⵉ ⵉⵇⵇⵉⵎ ? [mani iqqim ?] "où est-il assis ?".

La négation des questions totales et partielles conserve la même structure syntaxique que leurs équivalentes affirmatives, le morphème de négation ⵓⵔ suit directement le morphème interrogatif. Ainsi, la structure de la phrase interrogative négative est la suivante : "**Morphème interrogatif + ⵓⵔ + VP ?** ". Exemple : ⵎⴰ + ⵓⵔ + ⴷ ⵜⵓⵙⵉ ? [ma ur d tusi ?] " N'est-elle pas venue ? ".

2.5. Conclusion

La langue amazighe est une des langues les plus anciennes de l'Afrique du Nord. Dans ce chapitre, nous l'avons présentée brièvement : son historique, son système d'écriture, et sa situation au Maroc par rapport à son intégration dans les technologies d'information. Aussi, nous avons abordé ses aspects linguistiques : morphologiques et syntaxiques. La connaissance de ces deux aspects est nécessaire pour assurer le transfert grammatical entre la phrase source LN_S et la phrase cible LN_C dans un système de TA à base de règles, notamment un système de TA par interlangue qui est l'objet du cas d'étude de cette thèse.

Chapitre 3 : L’interlangue UNL

Chapitre 3

L'interlangue UNL

Sommaire

3.1. Introduction

UNL (Universal Networking Language), qui est traduit en français par « le langage universel de communication sur Internet », est un langage artificiel qui représente et exprime l'information contenue dans un texte sous format d'un graphe sémantique. L'objectif de la création d'un tel langage est de jouer le rôle de médiateur entre les différentes langues naturelles du monde entier, afin de briser les barrières linguistiques et de promouvoir le multilinguisme, le dialogue et la coopération entre les peuples. Cette représentation sémantique est utilisée comme étant un langage interlangue (pivot) dans des systèmes de traduction automatique (TA), où la traduction de n'importe quelle langue source vers n'importe quelle langue cible consiste à « convertir » la phrase source vers la représentation sémantique UNL, puis à « déconvertir » la phrase cible à partir de cette représentation UNL. Le présent chapitre décrira l'UNL en tant que langage, en présentant sa syntaxe, et en tant que système de TA, en présentant ses composantes linguistiques et logicielles.

3.2. Historique du projet UNL

Le projet UNL a été fondé à l'Institut des Etudes Avancées de l'Université des Nations Unies de Tokyo, en 1996, sous les auspices de l'UNESCO (Uchida et *al.*, 1999). Le but de ce projet est de permettre à toute personne du monde entier d'accéder à toutes les informations existantes sur Internet dans sa langue maternelle, favorisant ainsi le multilinguisme et réduisant les contraintes d'accès à l'information dues aux barrières linguistiques. Par exemple, une page web écrite en français peut être lue par un chinois, un arabe, un américain, etc dans leurs propres langues maternelles (*cf.* Figure 3.1).

Figure 3. 1 L'objectif du projet UNL est l'éclosion du multilinguisme

Initialement, le projet a démarré avec la participation de 15 langues : l'arabe, le chinois, l'anglais, le français, l'allemand, le hindi, l'indonésien, l'italien, le japonais, le letton, le mongol, le portugais, le russe, l'espagnol et le thaïlandais (Uchida *et al.*, 1999).

3.3. Présentation de l'UNL en tant que langage

Le langage UNL permet de coder le sens véhiculé par un texte donné sous format d'un réseau sémantique qui se compose d'un ensemble de nœuds appelés *Mots universels (Universal Words, UWs), d'*un ensemble de liens entre les nœuds appelés des *Relations universelles*, et d'un ensemble d'informations rattachées à chaque nœud, appelées des *Attributs universels* (*cf.* Figure 3.2).

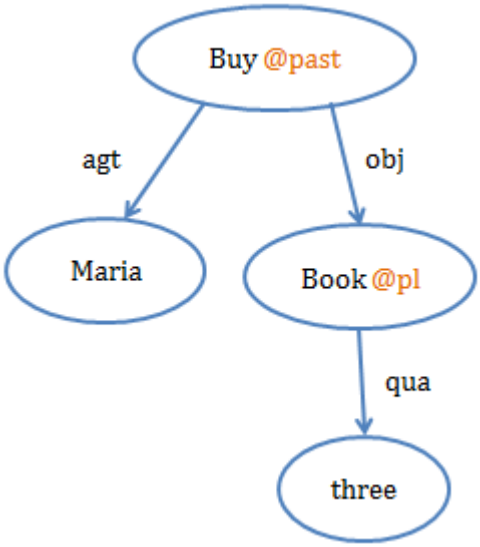

Figure 3. 2 Exemple d'un graphe UNL simplifié correspondant à la phrase : « Maria a acheté trois livres»

3.3.1. Mots Universels (UW)

Les mots universels ou Universal Words (UWs) constituent le vocabulaire du langage UNL. Chaque UW est formé d'un mot et d'une liste de restrictions. Le mot est généralement en anglais, mais il pourrait être dans une autre langue translittérée aux caractères latins dans le cas où il n'existe pas d'équivalent anglais pour un concept local d'une culture donnée. En effet, la notion d'universalité, dans l'UNL, veut dire que ce langage est capable d'être utilisé et compris par tous les humains. Il peut représenter aussi bien les concepts globaux que ceux les plus locaux. Cependant, la liste de restrictions est ajoutée dans le cas où l'UW est ambigu. Ces restrictions sont des relations entre le dit UW et un autre (Bekios *et al.*, 2007). Par exemple, le

mot anglais « bank » est ambigu, il veut dire en même temps « une banque » et « une rive », pour ce cas, l'ajout d'une liste de restrictions est obligatoire pour lever toute ambiguïté possible. Les UWs sont de deux types : les UWs permanents et les UWs temporaires.

- **UWs permanents** : sont les UWs qui correspondent aux concepts lexicalisés par au moins une langue.
- **UWs temporaires** sont les UWs qui représentent les concepts ou des entités qui sont soit encore en cours de lexicalisation ("googlers", "twittered", "Youtubeur",…) ou sont assez spécifiques pour être inclus dans le dictionnaire UNL, ou bien qui ne sont pas traduisibles ("3.14159", "H2O", "www.undlfoundation.org").

3.3.2. Relations Universelles

Les relations constituent la syntaxe de l'UNL. Elles sont symbolisées par trois caractères signifiant le type de la relation sémantique liant deux UWs dans un graphe UNL. Les relations universelles sont représentées comme suit : **<Rel> (<source>, <cible>),** avec

- <Rel> est le nom de la relation ;
- <Source> est l'UW qui attribue la relation <rel> ;
- <Target> est l'UW qui reçoit la relation <rel>.

La fondation UNDL a défini 39 relations universelles, nous présentons quelques-unes figurant dans la représentation UNL correspondante à la phrase suivante : « Marie a vu Peter quand John est arrivé » (*cf.* Figure 3.3). Comme le montre cette figure, les nœuds représentant les UWs : "Mary", "saw", "Peter", "John", "arrived" correspondent, respectivement, aux mots "Marie", "a vu", "Peter", "John", " est arrivé". Cependant les relations universelles sont les arcs liant ces UWs. En l'occurrence les relations :

- Agt (saw; Mary) exprime que Marie est l'agent du verbe voir ;
- obj (saw; Peter) décrit que Peter est l'objet du verbe voir ;
- tim (saw; hyper-noeud), dans cet exemple, relie l'UW "saw" avec un hyper-nœud constitué des deux UWs "John" et "arrived". Elle exprime le temps de l'action du verbe voir.

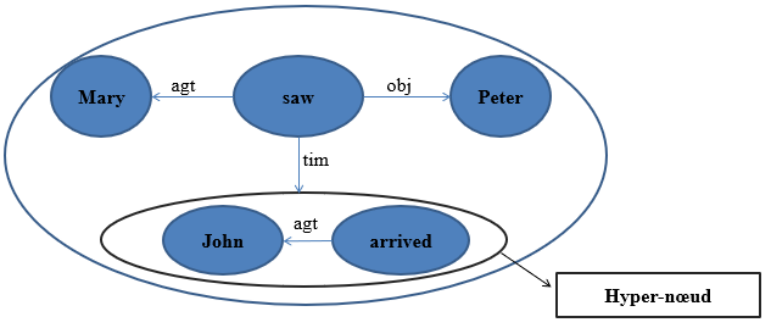

Figure 3. 3 Graphe UNL correspondant à la phrase : « Marie a vu Peter quand John est arrivé »

3.3.3. Attributs Universels

Les attributs universels sont des annotations, précédés par le symbole '@' et ajoutés aux nœuds ou aux hyper-nœuds, dans un graphe UNL, pour indiquer leurs catégories grammaticales, tels que le nombre, le genre, le temps, le mode, l'aspect, ou bien indiquer le contexte dans lequel ils sont utilisés. Tableau 3.1 décrit quelques attributs à titre d'exemple.

Tableau 3. 1 Extrait des attributs UNL

Attributs	Définition
@entry	Cet attribut indique le nœud principal dans un graphe UNL
@past, @present, @future, …	Exemple d'attributs exprimant le temps
@male, @female, @neutral	Liste des attributs exprimant le genre
@surprise, @anger, @hesitation	Exemple d'attributs exprimant les émotions …
@i @singular, **@i @pl,**	Liste des attributs exprimant la personne : • @i : exprime l'indice de personne, avec i appartient à {1, 2, 3} • @singular : singulier • @pl : pluriel
@obligation, @prohibition, @ability	Exemples d'attributs exprimant les modalités

3.3.4. Format des expressions UNL

Dans cette section, nous présentons, la représentation UNL d'une phrase simple et d'une autre relativement complexe.

- « Maria a acheté trois livres.» (2.1)
- « Maria, qui habite à Casablanca, travaille à Rabat. » (2.2)

La représentation UNL graphique et non graphique de la phrase (2.1) est illustrée respectivement dans les Figure 3.2 et Figure 3.4.

```
{unl}

agt(to buy(icl>get).@past, Maria(icl>name))
Obj(to buy(icl>get).@past,book(icl>product),@pl)
qua(book(icl>product),@pl,3)

{/unl}
```

Figure 3.4 Représentation UNL non graphique de la phrase (2.1)

Pour le cas des phrases complexes, leurs représentations en UNL sont réalisées en se servant des hyper-nœuds. Un hyper-nœud est un sous-graphe fonctionnant comme une seule entité sémantique comme l'hyper-nœud, nommé « scope :01 », présenté dans la Figure 3.5. L'expression UNL non graphique de la phrase (2.2) est présentée dans la Figure 3.6.

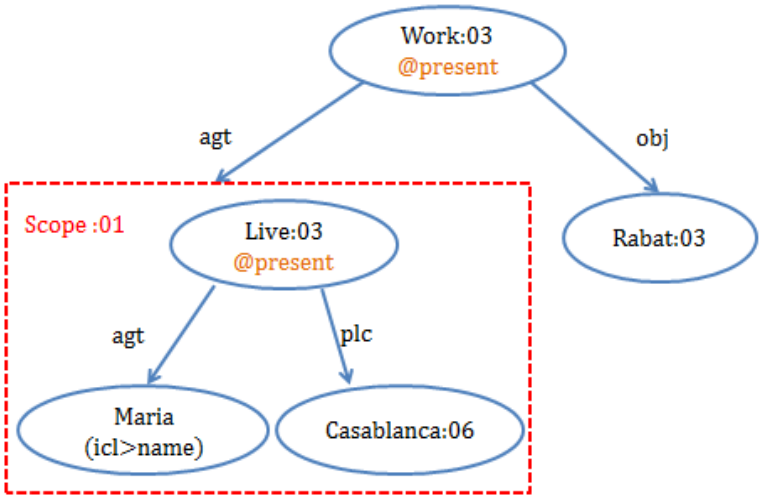

Figure 3. 5 Représentation UNL graphique de la phrase (2.2)

```
{unl}

Agt(work:03.@present, :01)
Obj(work:03.@present, Rabat:02)
Agt:01(live:04.@present, Maria(icl>name):05)
Plc:01(live:04.@present,Casablanca:06)

{/unl}
```

Figure 3. 6 Représentation UNL non-graphique de la phrase (2.2)

3.4. Présentation de l'UNL en tant que système de TA

Le développement d'un système de TA par interlangue se réalise en développant deux modules indépendants l'un de l'autre : le module d'analyse et le module de génération. Le module d'analyse est le module qui se charge de la représentation du sens d'un document, dans une langue naturelle source (LN_S), en un document UNL ; alors que le module de génération est celui qui prend le document UNL obtenu et le transforme en une langue naturelle cible (LNc). La Figure 3.7 présente l'architecture générale de la TA à base de l'UNL. La description de chacune de ses composantes est donnée dans les paragraphes suivants.

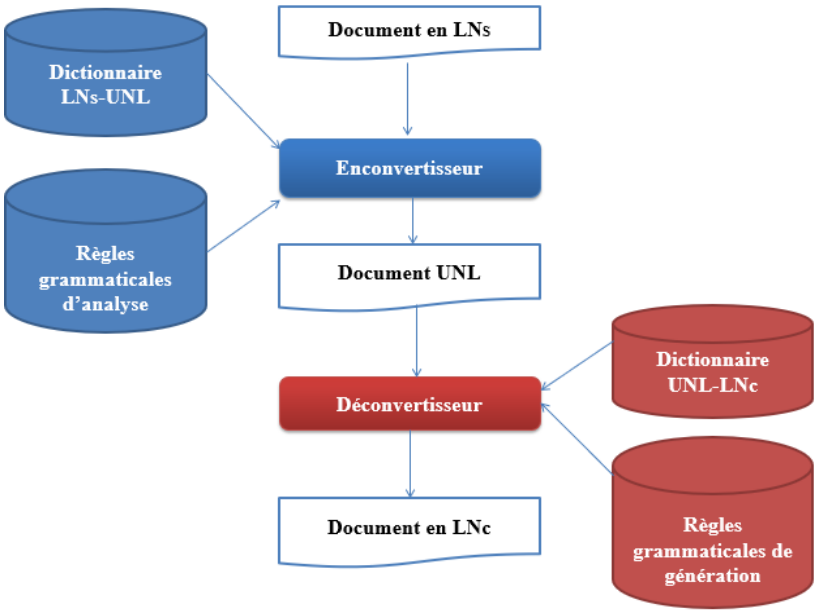

Figure 3. 7 Architecture générale de la TA à base de l'UNL

Comme, le présente cette figure, le système UNL de traduction requière un ensemble d'utilitaires et de ressources linguistiques.

3.4.1. Ressources requises pour le module d'analyse

- **Dictionnaire d'analyse (LN-UNL) :** est un dictionnaire bilingue qui associe chaque mot d'une LN à son équivalent en UNL. Le format de ce dictionnaire est énumératif, c'est-à-dire il apporte toutes les formes fléchies d'un mot en tant que des entrées du dictionnaire (Teixeira Martins et Avetisyan, 2009).
- **Règles grammaticales d'analyse :** sont l'ensemble des règles grammaticales responsables de la transformation d'une phrase en une langue naturelle vers sa représentation en UNL.
- **Enconvertisseur (Analyseur) :** est l'utilitaire qui se charge de la transformation d'une phrase d'une LN_s vers l'expression UNL correspondante.

3.4.2. Ressources requises pour le module de génération

- **Dictionnaire de génération (UNL-LN) :** est un dictionnaire bilingue qui associe chaque mot d'une LN à son équivalent en UNL. Le format de ce dictionnaire est génératif, c'est-à-dire il apporte juste les formes de base (les lemmes) sans leurs formes fléchies (Teixeira Martins et Avetisyan, 2009). Les spécifications requises pour développer un tel dictionnaire sont détaillées dans le Chapitre 4.
- **Règles grammaticales de génération :** sont un ensemble de règles sémantico-syntaxiques et morphologiques responsables de la transformation des graphes UNL en des phrases en langues naturelles.
- **Déconvertisseur (Générateur) :** est l'utilitaire qui se charge de la transformation d'une expression UNL vers son équivalent en une langue naturelle.

3.5. Travaux antérieurs

Ces dernières années, la TA à base de l'UNL a intéressé plusieurs chercheurs dans le domaine de la TA. A titre d'exemple, nous citons les travaux proposés par :

- Thuyen, en 2016, pour la langue vietnamienne ;
- Kumar et Sharma, en 2013, pour la langue punjabie ;
- Dikonov, en 2011, pour le russe et l'anglais ;
- Adli et Alansary, en 2010, pour la langue arabe ;

- Giri et Leena, en 2000, ont développé le module d'analyse de la langue hindi, et Nalawade a poursuivi, en 2007, le développement du module de génération vers cette langue.

Comme nous pouvons le remarquer, la TA à base de l'UNL concerne aussi bien les langues avancées ou moyennement avancées, comme l'anglais et l'arabe, que les langues peu dotées comme le vietnamien, le punjabi.

3.6. Conclusion

Dans ce chapitre, nous avons présenté l'UNL en tant qu'une interlangue, en présentant sa syntaxe et en tant qu'un système de traduction automatique en citant quelques travaux de recherche dans ce domaine de traduction à base de l'UNL. En fait, cette interlangue présente plusieurs avantages, à savoir :

- Elle est basés sur une langue naturelle qui est l'anglais ;
- Elle est conçue aussi bien pour les langues peu dotées et en danger que les langues avancées ;
- Elle offre la possibilité d'ajouter de nouveaux concepts ;
- Elle est capable de résoudre les ambiguïtés des mots anglais par l'ajout des restrictions.

Pour tous ces avantages, nous avons opté pour l'intégration de la langue amazighe en tant qu'une langue peu dotée dans le projet UNL. Pour ce faire, nous avons essayé de préparer les ressources linguistiques nécessaires pour la TA depuis et vers cette langue. Dans les travaux de cette thèse, nous avons abordé dans une première étape la TA vers la langue amazighe.

Chapitre 4 : Construction du dictionnaire UNL-Amazighe

Chapitre 4

Construction du dictionnaire UNL-Amazighe

Sommaire

4.1. Introduction

Le dictionnaire électronique est une ressource linguistique très sollicitée par la plupart des applications du TALN. La construction d'un tel dictionnaire est une tâche laborieuse pour le cas d'une langue peu dotée qui ne possède pas assez de ressources électroniques. C'est le cas confronté pour la langue amazighe.

Ce chapitre présente, dans un premier temps, notre contribution qui consiste en la recherche et la formalisation des paradigmes flexionnels et les cadres de sous-catégorisation de la langue amazighe, nécessaires pour l'élaboration du dictionnaire bilingue UNL-amazighe. Ensuite, dans un deuxième temps, il présente l'exploitation de ce dictionnaire bilingue UNL-amazighe pour mettre en place un dictionnaire bidirectionnel multilingue via l'interlangue UNL. L'avantage de cette approche est la possibilité de construire un dictionnaire multilingue entre n langues, même s'il n'existe pas de dictionnaires bilingues entre eux. Actuellement, le dictionnaire multilingue mis en œuvre comporte les cinq langues les plus parlées au Maroc : l'arabe et l'amazighe, qui sont les langues officielles ; le français, qui est considéré comme la première langue étrangère ; l'espagnol qui est fréquemment parlé par la population du nord du pays ; et l'anglais qui devient de plus en plus utilisé dans les secteurs de l'éducation, des affaires et des sciences au Maroc.

4.2. Format du dictionnaire UNL-LN

Selon les spécifications de l'UNL, chaque entrée du dictionnaire respecte le format suivant (Teixeira Martins et Avetisyan, 2009) :

[NLW] {ID} 'UW' (ATTR …) ;

avec :

- NLW : mot d'une Langue Naturelle (Natural Language Word) ;
- ID : Identifiant de l'entrée ;
- UW : le mot universel (Universal Word) ;
- ATTR : la liste des informations linguistiques, à savoir : la catégorie lexicale (LEX), la catégorie syntaxique (POS), la transitivité (TRA) (dans le cas des verbes), le paradigme flexionnel (PAR) et le cadre de sous catégorisation (FRA).

Nous donnons, à titre d'exemples, les entrées du dictionnaire UNL-amazighe suivantes :

- [ⴰⵎⵓⵔⴰⵖ]{2572}"Kid"(LEX=N, POS=NOU, LST=WRD, GEN=MCL, NUM=SNG, PAR=M50, FRA=Y0) ;
- [ⵙⵖ]{4646}"buy"(LEX=V, POS=VER, LST=WRD, TRA=TST, PAR=M170, FRA=Y0).

Le dictionnaire UNL-LN est appelé un dictionnaire génératif lorsqu'il n'apporte que les formes de base (les lemmes) d'une LN comme entrées ; mais quand il apporte toutes les formes fléchies des mots comme entrées, il est appelé un dictionnaire énumératif. Ce dernier est obtenu en développant le dictionnaire génératif, et en fournissant à chaque entrée l'identificateur de son paradigme flexionnel (PAR). Un paradigme flexionnel est un ensemble de règles flexionnelles qui permettent de générer à partir d'un lemme toutes ses formes fléchies. Selon les spécifications d'UNL, chaque règle flexionnelle s'écrit de la manière suivante : associer à chaque lemme amazighe, son paradigme flexionnel adéquat n'était pas une tâche évidente vu que les ressources linguistiques sur lesquels nous nous sommes basées (Abdellaoui, 2012 ; Agnaou *et al.*, 2011; Ameur, 2009a; Ameur, 2009b; El Azrak, 2009) ne fournissaient pas ce genre d'information. Du coup, nous étions amenés à faire une analyse morphologique de la langue amazighe pour chercher et formaliser ses paradigmes flexionnels.

<*CONDITION*>:= <*ACTION*>; avec:

- <CONDITION>: contient l'ensemble des propriétés grammaticales d'une forme de base (lemme), séparées par le symbole « & » ;
- <ACTION> décrit les changements à effectuer sur la forme de base, à savoir la préfixation, la suffixation ou l'infixation.

Par exemple, la règle flexionnelle française suivante : PLR: = 0> "s" ; se lit comme suit : pour générer la forme plurielle (PLR) d'un mot au singulier, le caractère "s" est ajouté à la fin du mot (garçon> garçons, table> tables).

4.3. Formalisation des paradigmes nominaux et adjectivaux

La construction des paradigmes flexionnels de la catégorie nominale et adjectivale a présenté, pour nous, un vrai challenge, étant donné qu'il y a un manque de travaux sur la classification de ces deux catégories par rapport à leurs flexions et à l'existence de plusieurs formes imprévisibles pour les formes du pluriel et de l'état d'annexion.

Notre approche, pour formaliser ces paradigmes flexionnels nominaux, consistait à collecter les noms amazighes avec leurs formes fléchies (leurs flexions) à partir de quatre différents lexiques (Agnaou et al., 2011; Ameur et al., 2009a; Ameur et al., 2009b; El Azrak et al., 2009), afin de les analyser et étudier leurs variations morphologiques pour pouvoir les classer selon le procédé de formation de leurs flexions. Notre conception de classification des noms était, dans une première étape, de distinguer entre les noms masculins et les noms féminins, puis, dans une deuxième étape, de les classer en des classes morphologiques de telle façon, chaque classe regroupe les noms dont le procédé de formation de leurs formes du pluriel est le même. La troisième étape, quant à elle, consistait de partir de chacune des classes définies dans la deuxième étape pour identifier de nouvelles sous-classes des noms ayant la même forme de l'état d'annexion (*cf.* Figure 4.1). En procédant de cette manière, nous avons pu identifier 100 classes flexionnelles nominales : 53 classes pour les noms masculins et 47 classes pour les noms féminins (Taghbalout *et al.*, 2015).

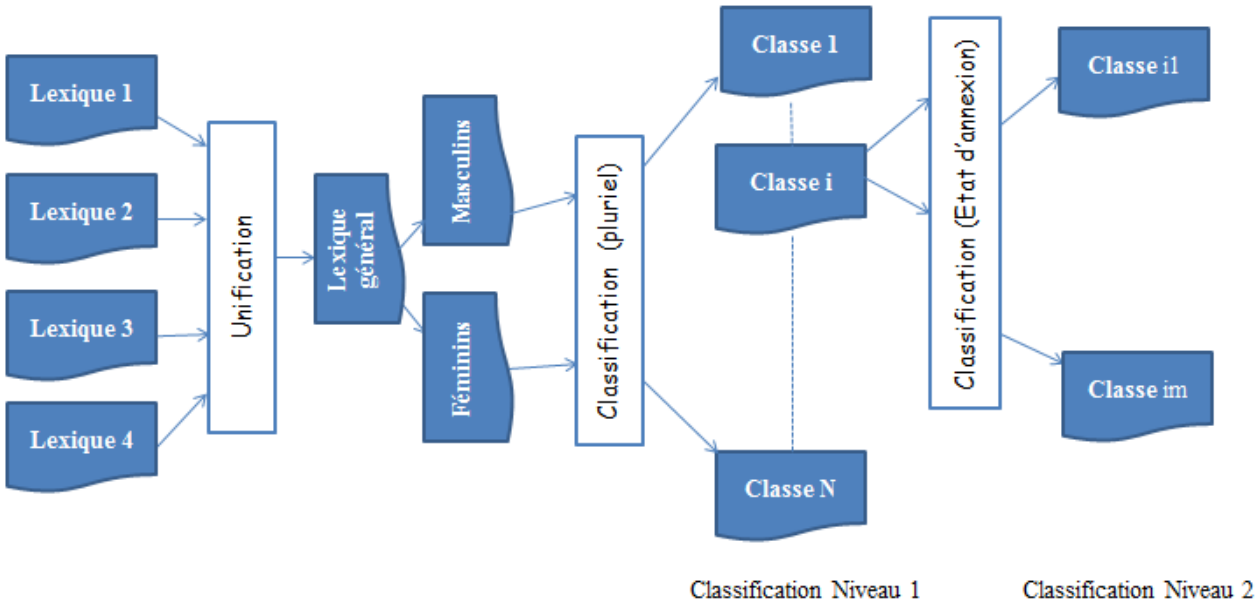

Figure 4. 1 L'approche suivie pour la construction des classes flexionnelles nominales et adjectivales

A ce stade, nous pouvons formuler les paradigmes flexionnels de la catégorie nominale amazighe, en définissant des fonctions implémentant le procédé de formation du pluriel et celui de formation de l'état d'annexion d'une classe morphologique donnée. A titre d'exemple, la Figure 4.2 présente une fonction implémentant le processus de génération de la forme du pluriel à partir de la forme du singulier de la classe nominale N° 29.

```
function class29($sing){//ⴰⵜⵚ(singulier) --> ⵚⵜⵚⴰⵏ (pluriel)

    //Changement de la voyelle initiale par "ⵚ"
    $l=strlen($sing)-3;
    $initial=substr($sing,0,3);
    $sing_ini=substr($sing,3,$l);
    //suffixation de "ⴰⵏ"
    $plr=$initial.$sing_ini.'ⴰⵏ';
    // Retour de la forme du pluriel
    return $plr;
}
```

Figure 4. 2 Exemple de fonction implémentant le procédé de formation du pluriel d'une classe morphologique nominale

Pour les adjectifs, nous avons distingué entre deux types : ceux qui sont des adjectifs (adj.) seulement et ceux qui sont en même temps des adjectifs et des noms (adj. et n.). Le premier type supporte la flexion du genre et du nombre, par contre le deuxième type supporte la flexion du genre, du nombre et de l'état d'annexion. De la même façon avec laquelle nous avons pu construire les classes nominales (*cf.* Figure 4.1), nous avons formalisé 19 classes flexionnelles adjectivales : 6 classes pour le cas des adjectifs et 13 classes pour le cas des adjectifs et noms (Taghbalout et *al.*, 2016a). Pour le cas d'un adjectif qui est aussi un nom, selon les spécifications d'UNL, il est nécessaire de créer un paradigme flexionnel pour le cas du nom et un autre paradigme pour le cas de l'adjectif même si les deux paradigmes partagent les mêmes règles. L'évaluation de la couverture de ces classes proposées est présentée dans la section 4.5.

Notre méthodologie de classification n'a pas procédé par schèmes, parce qu'il existe des noms amazighes ayant le même schème[9] mais ils prennent des marques de pluriel différentes, et aussi il existe des noms ayant des schèmes différents mais ils prennent les mêmes marques du pluriel. En effet, comme le présente le Tableau 4.1, même si le nom ⴰⵊⴰⴳⵓ [abagus] 'ceinture' et ⴰⴱⴰⵟⵓⵍ [abatul] 'terrain plat' partagent le même schème ⴰCⴰCⵓC [aCaCuC], ils prennent des marques du pluriel différentes. Par contre, le mot ⴰⵍⵉⵙⵉ [adlis] 'livre', ayant le schème ⴰCCⵙC [aCCiC], et le mot ⴰⵊⴰⴳⵓ [abagus], ayant le schème ⴰCⴰCⵓC [aCaCuC] qui est différent de celui du mot ⴰⵍⵉⵙⵉ, leurs formes du pluriel se forment de la même façon.

9 Schème : partie du mot complémentaire à la racine.

Tableau 4. 1 Exemples des formes de pluriel

Nom	Schème	Forme du pluriel
ⵄⵖⵄⵅⵙⵔ [abagus]	ⵄCⵄCⵙC	ⵉⵖⵄⵅⵙⵔⵏ [ibagusn]
ⵄⵖⵄⵜⵙⵍ [abatul]	ⵄCⵄCⵙC	ⵉⵖⵙⵜⵄⵍ [ibutal]
ⵄⵍⵍⵉⵙ [adlis]	ⵄCCⵉC	ⵉⵍⵍⵉⵙⵏ [idlisn]

Après avoir formalisé les classes flexionnelles nominales et adjectivales de l'amazighe, nous avons développé les règles flexionnelles adéquates pour générer toutes les formes fléchies d'un lemme appartenant à une classe donnée. A titre d'exemple, le Tableau 4.2 décrit les règles flexionnelles d'une classe nominale.

Tableau 4. 2 Règles flexionnelles du paradigme nominal M49 (Nom modèle = ⵄⵙⵍⵎⴰⴷ [aslmad] 'enseignant')

Règles flexionnelles	Explication	Formes fléchies
MCL&SNG&NOM: =0>"";	Pas de changement dans le cas où le nom est au masculin, singulier, et à l'état libre	ⵄⵙⵍⵎⴰⴷ [aslmad]
MCL&PLR&NOM:="ⵉ"<1,0>"ⵏ";	Changement de la première lettre par "ⵉ" et suffixation de "ⵏ" lorsque le nom est au masculin, pluriel, et à l'état libre	ⵉⵙⵍⵎⴰⴷⵏ [islmadn]
MCL&SNG&CTS:= "ⵓ"<1;	Changement de la première lettre par "ⵓ" lorsque le nom est au masculin, singulier, et à l'état d'annexion	ⵓⵙⵍⵎⴰⴷ [uslmad]
MCL&PLR&CTS:="ⵉ"<1, 0>"ⵏ";	Changement de la première lettre par "ⵉ" et suffixation de "ⵏ" lorsque le nom est au masculin, pluriel, et à l'état d'annexion	ⵉⵙⵍⵎⴰⴷⵏ [islmadn]
FEM&SNG&NOM:="ⵜ"<0,0>"ⵜ";	Préfixation de la lettre "ⵜ" et suffixation de la lettre "ⵜ" lorsque le nom est à l'état libre, au féminin, et au singulier	ⵜⵄⵙⵍⵎⴰⴷⵜ [taslmadt]
FEM&SNG&CTS:="ⵜ"<1, 0>"ⵜ";	Changement de la première lettre par "ⵜ" et suffixation de "ⵜ" lorsque le nom est à l'état d'annexion, au féminin, et au singulier	ⵜⵙⵍⵎⴰⴷⵜ [tslmadt]
FEM&PLR&NOM:="ⵜⵉ"<1,0>"ⵉⵏ";	Changement de la première lettre par "ⵜⵉ" et suffixation de "ⵉⵏ" lorsque le nom est à l'état libre, au féminin, et au pluriel	ⵜⵉⵙⵍⵎⴰⴷⵉⵏ [tislmadin]

4.4. Formalisation des paradigmes flexionnels verbaux

L'élaboration des paradigmes flexionnels, pour la catégorie verbale, s'est basée sur la classification proposée par (Abdellaoui et al., 2012). Cette classification consiste à diviser les verbes en trente classes suivant la formation de l'aspect accompli et de l'aspect inaccompli. Cependant, notre approche de classification consiste à analyser chaque classe, parmi les trente classes, et formaliser de nouvelles sous-classes qui en plus prennent en compte le procédé de formation de l'accompli négatif. En procédant ainsi, chaque sous-classe contiendra juste les verbes ayant les mêmes règles flexionnelles pour générer ses aspects. Finalement, nous nous sommes arrivés à formaliser 67 classes flexionnelles pour les verbes amazighes (Taghbalout *et al.*, 2016b). L'évaluation de la couverture de ces classes est présentée dans la Section 4.5. Il faut noter, qu'un même verbe amazighe peut appartenir à plusieurs classes flexionnelles à la fois suivant le sens qu'il porte et sa variété régionale.

La formalisation des règles flexionnelles verbales se fait par l'expression, dans un premier temps, des règles morphotactiques responsables de la génération de la forme aspectuelle, suivies par les règles morphotactiques relatives aux désinences des indices de personnes qui sont marquées par la couleur bleue dans le Tableau 4.3. Cependant, les marques de l'aspect sont marquées en rouge. Les formes fléchies, d'un verbe amazighe, sont déterminées selon :

- le mode (indicatif (IND), impératif (IMP), participial (PTP)) ;
- le genre (masculin (MCL), féminin (FEM)) ;
- le nombre (singulier (SNG), pluriel(PLR)) ;
- l'indice de personne (1ère personne du singulier (1PS), 2ème personne du singulier (2PS), 3ème personne du singulier (3PS), 1ère personne du pluriel (1PP),…).

Toutes ces règles flexionnelles sont combinées, une après l'autre, d'une manière linéaire. Tableau 4.3 décrit un extrait des règles flexionnelles crées pour la classe verbale 17-1.

Tableau 4. 3 Règles flexionnelles pour générer les formes de l'inaccompli (NPFV) au mode (IND)

Classe 17-1	Règles flexionnelles	Formes fléchies
"ⵙⴰⵙ" [bdd] 'se lever'	1PS&NPFV&IND:="ⵜⵜ"<0, 0>"ⵔ",0>"ⵖ";	ⵜⵜⵙⴰⴰⵔⵖ [ttbddagh]
	2PS&NPFV&IND:="ⵜⵜ"<0,0>"ⵔ","ⵜ"<0,0>"ⴰ","ⵜⵜⵜ:"ⵜⵛⵜⵜ";	ⵜⵛⵜⵜⵙⴰⴰⵔⴰ [tttbddad]
	3PS&MCL&NPFV&IND:="ⵜⵜ"<0,0>"ⵔ","ⵣ"<0;	ⵣⵜⵜⵙⴰⴰⵔ [ittbdda]
	3PS&FEM&NPFV&IND:="ⵜⵜ"<0,0>"ⵔ","ⵜ"<0, "ⵜⵜⵜ":"ⵜⵛⵜⵜ";	ⵜⵛⵜⵜⵙⴰⴰⵔ [tttbdda]
	1PP&NPFV&IND:="ⵜⵜ"<0,0>"ⵔ","ⵏ"<0;	ⵏⵜⵜⵙⴰⴰⵔ [nttbdda]
	2PP&MCL&NPFV&IND:="ⵜⵜ"<0,0>"ⵔ","ⵜ"<0,>"ⵆ" ,"ⵜⵜⵜ":"ⵜⵛⵜⵜ";	ⵜⵛⵜⵜⵙⴰⴰⵔⵆ [tttbddam]
	2PP&FEM&NPFV&IND:="ⵜⵜ"<0,0>"ⵔ","ⵜ"<0, >"ⵆⵜ", ⵜⵜⵜ" : "ⵜⵛⵜⵜ";	ⵜⵛⵜⵜⵙⴰⴰⵔⵆⵜ [tttbddamt]
	3PP&MCL&NPFV&IND:="ⵜⵜ"<0,0>"ⵔ",0>"ⵏ ";	ⵜⵜⵙⴰⴰⵔⵏ [ttbddan]
	3PP&FEM&NPFV&IND:="ⵜⵜ"<0,0>"ⵔ", 0>"ⵏⵜ";	ⵜⵜⵙⴰⴰⵔⵏⵜ [ttbddant]

4.5. Evaluation des classes flexionnelles formalisées

Dans les sections précédentes 4.3 et 4.4, nous avons décrit la méthodologie suivie pour la formalisation des classes flexionnelles des noms, des adjectifs et des verbes amazighes. Dans cette section, nous évaluerons la couverture de ces classes. Sachant que l'étude morphologique, que nous avons entreprise, s'est basée sur le dictionnaire «D1», de taille de 8728 lemmes, construit à partir de cinq ouvrages (Abdellaoui et al., 2012 ; Agnaou et al., 2011; Ameur et al., 2009a; Ameur et al., 2009b; El Azrak, 2009), nous avons cherché de nouveaux lemmes qui n'appartiennent pas à "D1", et nous avons essayé d'appliquer sur eux les paradigmes formalisés. En effet, nous avons extrait de nouveaux noms, adjectifs et verbes d'un large

dictionnaire récent de taille de 14 693 lemmes "D2" (Ameur et *al.*, 2017). Le tableau 4.4 présente les résultats des taux de couverture des classes flexionnelles formalisées pour chaque catégorie grammaticale.

Tableau 4. 4 Evaluation de la couverture des classes flexionnelles formalisées

	Nom	**Verbe**	**Adjectif (type 2)**
Lemmes appartenant à " D2" et non à "D1"	4 803	1 645	618
Lemmes non couverts	201	160	9
Couverture (%)	**96%**	**91%**	**99%**

Comme illustré dans le Tableau 4.4, les résultats obtenus pour la catégorie nominale sont satisfaisants. Le taux des noms couverts atteint 96%. L'analyse des 4% restant non couverts sont des noms empruntés ou des noms irréguliers ayant des formes de pluriel imprévisibles tel que le nom ⵉⵡⵉ [iwi] "mon fils", qui a un pluriel irrégulier ⵜⴰⵔⵡⴰ [tarwa] et le mot ⵙⵙⵓⵇ [ssuq] "marché", qui est un nom emprunté de l'arabe et dont la forme du pluriel est [laswaq] "marchés". Pour la catégorie verbale, le taux des verbes couverts est de 91%. Les 9% restant non couverts sont des verbes irréguliers ou des verbes qui se mettent, généralement, à un seul aspect, par exemple le verbe ⵙⴽⵓⵀⵀⵓ [skuhhu] "tousser" qui se conjugue souvent à l'aspect inaccompli. Pour les adjectifs, comme le montre le Tableau 4.4, sont bien couverts, les 1% qui restent non couverts sont des adjectifs irréguliers prenant des marques de pluriel imprévisibles.

Nous envisageons améliorer la couverture, des classes que nous avons proposées, notamment pour le cas des verbes, en formalisant de nouvelles classes flexionnelles, afin de prendre en compte aussi les variations morphologiques des entrées non couvertes du dictionnaire "D2".

4.6. Formalisation des cadres de sous catégorisation

La sous-catégorisation est la définition du nombre et des types d'arguments syntaxiques nécessaires qu'exige un mot pour avoir un sens. Par exemple, certains verbes, dits transitifs, nécessitent la présence de deux syntagmes nominaux : l'un en tant qu'un spécificateur (sujet) et l'autre en tant qu'un complément d'objet direct afin que la phrase ait un sens. Pour la langue amazighe, nous avons formalisé, quinze cadres de sous-catégorisation pour les noms et verbes.

Tableau 4.5, ci-dessous, donne quelques cadres de sous-catégorisation que nous avons formalisés.

Tableau 4. 5 Cadres de sous-catégorisation formalisés pour la langue amazighe

Cadre de sous-catégorisation	Description	Exemple
VS(NP)	Ce cadre décrit les verbes intransitifs exigeant seulement la présence d'un syntagme nominal (NP) comme spécificateur (VS).	ⵙⵅⵍ [ign] " il dort"
VS(NP)VC(NP);	Ce cadre décrit les verbes transitifs directs exigeant la présence d'un syntagme nominal (NP) comme spécificateur (VS), et d'un autre NP comme complément d'objet direct.	ⵙⵙⵢⴰ ⵎⵙⵎⵓⵏ ⵢⵓⵜ ⵜⴰⵏⵏⴰⵔⵜ [isgha mimoun yut taddart] " Mimoun a acheté une maison"
VS(NP)VC(NP)VC(PH([ⵉ]));	Ce cadre décrit les verbes di-transitifs exigeant la présence d'un syntagme nominal (NP) comme spécificateur (VS), d'un autre NP comme complément d'objet direct et d'un troisième NP comme complément d'objet indirect précédé par la préposition 'ⵉ' [i].	ⵙⵙⵙⵣⵣ ⵎⵙⵎⵓⵏ ⵜⴰⴱⵔⴰⵜ ⵉ ⵎⵎⵉⵙ [issk mimoun tabrat i mmis] " Mimoun a envoyé une lettre à son fils"
NC (PH([ⵏ]));	Ce cadre décrit les noms dont le complément est introduit par la préposition'ⵏ' [n].	ⵙⵎⵉ ⵏ ⵜⴰⵏⵏⴰⵔⵜ [imi n taddart] " Entrée d'une maison"

Chaque cadre de sous catégorisation est représenté par un identifiant qui le caractérise au niveau du dictionnaire UNL-amazighe. Une fois que nous avons formalisé les paradigmes flexionnels et les cadres de sous-catégorisation amazighes, il nous restera l'élaboration du mapping entre les lemmes amazighes et les UWs que nous présenterons dans la section suivante.

4.7. Réalisation du mapping entre les lemmes amazighes et les UWs

Chaque lemme amazighe de notre dictionnaire a été associé à l'identifiant de l'UW qui est le concept portant le sens de ce lemme. Les UWs sont divisés en quatre classes principales : les concepts verbaux, les concepts nominaux, les concepts adjectivaux, et les concepts adverbiaux.

4.7.1. Les étapes du mapping lexical

Nous avons effectué le mapping lexical entre les mots amazighes et leurs UWs équivalents, en procédant comme suit :

- identification du concept derrière le mot amazighe ;
- mise en correspondance entre le mot amazighe et l'UW approprié ;
- création de nouveaux UWs pour les mots amazighes qui n'ont pas d'équivalents en UNL, exprimant exactement le vrai sens de ces mots, notamment pour le cas des concepts locaux propres à la culture amazighe. Dans cette situation, ces mêmes mots amazighes translittérés en anglais peuvent être considérés comme des nouveaux UWs.

4.7.2. Les défis relevés lors de la phase du mapping lexical

Lors de la phase du mapping lexical entre les mots amazighes et leurs UWs correspondants, un ensemble de défis ont été relevés. Comme nous l'avons cité à la section 4.7, les UWs sont classés en des catégories morphosyntaxiques selon la langue anglaise. Cependant, il existe des mots amazighes qui peuvent ne pas partager les mêmes catégories morphosyntaxiques avec leurs équivalents en UNL. Par exemple, le mot ⵙⵓ [bu], qui est considéré comme déterminant en amazighe, il correspond au UW "owner" qui est un nom. Pour cette raison, nous ne pouvons pas trouver l'UW correspondant au mot ⵙⵓ [bu]. Un autre problème rencontré est le manque des UWs adéquats. En fait, certains mots amazighes n'ont pas d'équivalents en UNL qui expriment leurs significations exactes. Par exemple les mots amazighes ⴰⵎⵣⵔⵣⵔ [amzrzr], qui signifie le cheval blanc, et le mot ⴰⵀⵉⵏⵓ [ahidous], qui désigne une danse traditionnelle. Devant cette situation, nous avons créé de nouveaux UWs pour ces concepts.

Nous avons présenté dans les sections précédentes, les différents travaux menés en vue de l'élaboration du dictionnaire bilingue UNL-amazighe. Actuellement, nous sommes arrivés à l'alimenter par 8 728 lemmes. Dans les prochaines sections, nous présenterons notre deuxième

contribution qui consistera à l'exploitation de ce dictionnaire bilingue pour la construction d'un dictionnaire multilingue.

4.8. Implémentation du dictionnaire multilingue amazighe

Dans cette section, nous rapportons notre conception de développement d'un dictionnaire électronique bidirectionnel multilingue pour n langues. Ensuite, nous présenterons l'implémentation réalisée pour le cas de cinq langues : l'amazighe, l'anglais, l'arabe, l'espagnol et le français.

4.8.1. Analyse et conception

Notre méthodologie de développement d'un dictionnaire multilingue pour n langues, via l'interlangue UNL, consiste à faire fusionner n dictionnaires bilingues LN-UNL, en créant une matrice (a_{ij}) de taille m*n avec m est le nombre d'UWs (lignes) et n est le nombre de langues (colonnes). Comme décrit dans le Tableau 4.6, chaque UW est accompagné par un ensemble d'attributs morphosyntaxiques et sémantiques (Taghbalout *et al.*, 2017).

Tableau 4. 6 Structure proposée du dictionnaire multilingue

UWs	Description de l'UW	Amazighe	Anglais	Arabe	Espagnol
crippled(aoj>thing); LEX=J	Décrit une personne ayant un handicap au niveau des pieds (Boiteux). • La catégorie lexicale (LEX=J) désigne un adjectif	ⵄⴰⵣⵉⵣⵓⵏ [ahizun], ⵄⴰⵎⵓⵛⴰⵍ [amuchal], ⵄⴰⵣⵉⴷⴰⵔ [ahidar]	Lame	Cojo Lisiado	Mutilé
child(icl>person); LEX=N, ABN=CCT, ALY=ALI, ANI=ANM, CAR=CTB, SEM=HUM	Désigne une personne immature, infantile. • La catégorie lexicale (LEX=N) désigne un nom.	ⵣⵉⵣⵉⵣ [ijiji], ⵄⵛⴱⴱⵓⵍ [achbbul], ⵄⵣⵉⵣⵉⵣ [ahziz], ⵄⵀⵔⴰⵎ [ahram], ⵄⵀⵔⵎⵓⵛ [ahrmouch]	Child	niño	enfant

	• L'abstraction (ABN) est concrète (CCT). • L'aliénabilité (ALY) est aliénable (ALI). • L'animalité (ANI) est animée (ANM) • La cardinalité (CAR) est dénombrable (CTB) • La classe sémantique (SEM) est une personne (HUM)				

4.8.2. Implémentation

Le dictionnaire multilingue électronique, que nous avons développé, baptisé AMuD (Amazigh Multilingual Dictionary), est doté d'une interface graphique simple et intuitive. Le diagramme d'activité d'AMuD est décrit dans la Figure 4.3.

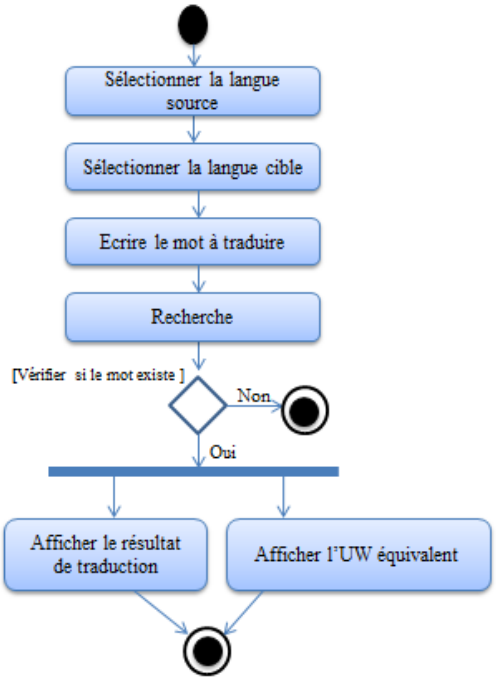

Figure 4. 3 Diagramme d'activité de l'application AMuD

Le développement du dictionnaire AMuD s'est basé sur six dictionnaires :

- Le dictionnaire UNL-amazighe, dont nous avons décrit le processus de son élaboration dans les sections précédentes. Il est de taille de 8 728 entrées.
- Le dictionnaire UNL-français, que nous avons participé à son enrichissement par 4 050 entrées, en collaboration avec la fondation UNDL. Il est de l'ordre de 140 000 entrées.
- Le dictionnaire UNL élaboré par la fondation UNDL. Il est de taille de 107 548 entrées.
- Le dictionnaire UNL-anglais élaboré par la fondation UNDL. Il est de taille de 130 000 entrées.
- Le dictionnaire UNL-arabe élaboré par la fondation UNDL. Il est de taille de 144 449 entrées
- Le dictionnaire UNL-espagnole élaboré par la fondation UNDL. Il est de taille de 73 653 entrées.

L'interface graphique d'AMuD est constituée de deux listes déroulantes : une pour choisir la langue source et l'autre pour choisir la langue cible à partir de la liste des langues {amazighe, anglais, arabe, espagnole, français} ; d'un bouton de traduction et de quatre zones (*cf.* Figure 4.4).

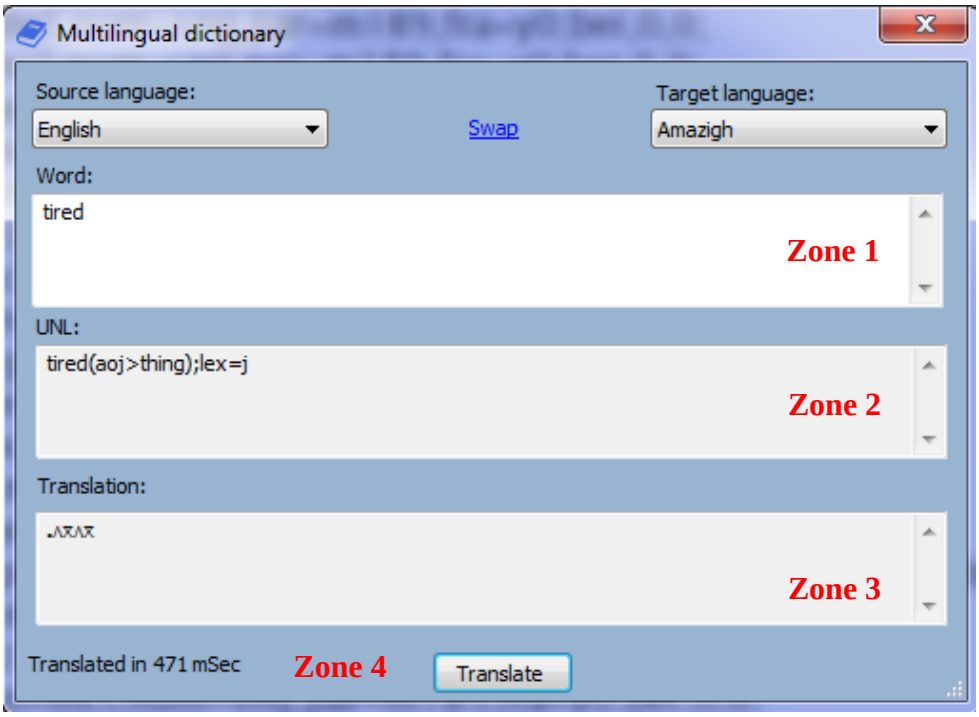

Figure. 4 Capture d'écran de l'application

La figure 4.4 montre un exemple de traduction de la langue anglaise vers l'amazighe. Chaque zone, dans la figure, est dédiée à une information spécifique :

- **Zone 1** :

Après avoir choisi la langue source et la langue cible, l'utilisateur saisi le mot anglais "tired" (fatigué en français) pour chercher son équivalent en amazighe. L'utilisateur a la possibilité de chercher les mots soit en les écrivant en minuscules, ou en majuscules, ou bien en combinant les deux.

- **Zone 2** :

Dans cette zone, le système affiche l'UW correspondant au mot anglais saisi, qui est *tired (aoj>thing).* L'UW est accompagné de sa catégorie lexicale (LEX=J) qui est dans ce cas un adjectif.

- **Zone 3** :

Cette zone est dédiée à l'affichage de la traduction amazighe, qui est dans ce cas ⵄⴰⵅⴰⵅ [adgdg].

- **Zone 4** :

Cette zone affiche le temps de réponse du système pour chercher la traduction du mot. Nous avons ajouté cette information juste pour donner une idée sur la performance de l'opération de recherche.

Concernant l'environnement de développement, opté pour la mise en œuvre de l'application AMuD, nous avons choisi de travailler avec le moteur de base de données SQLite, parce qu'il est open source, rapide, et compact vu la grande taille des dictionnaires utilisés. Pour l'interface graphique, nous l'avons développé avec le language C# sous Visual Studio.

Certes la langue amazighe est une langue peu dotée, qui manque de dictionnaires bilingues, mais nous avons pu la doter d'un dictionnaire électronique multilingue amazighe-anglais-arabe-espagnol-français grâce à l'interlangue UNL. Nous pourrons, également, intégrer de nouvelles langues dans le dictionnaire AMuD. En fait, il y a plus de 17 langues qui disposent de leurs dictionnaires selon les spécifications UNL comme le russe, l'arménien, le panjabi, l'hindi, etc. Ainsi, nous pourrons intégrer ces langues dans AMud et traduire les mots de ces langues en amazighe et vice versa.

4.9. Conclusion

Dans ce chapitre, nous avons présenté, dans un premier temps, notre première contribution qui a consisté en la recherche et la formalisation des paradigmes flexionnels de la langue amazighe en vue de l'élaboration du dictionnaire bilingue UNL-amazighe. Ensuite, dans un deuxième temps, nous avons exploité ce dictionnaire bilingue pour mettre en place un dictionnaire multilingue au profit de cette langue. Un tel dictionnaire UNL-amazighe est une ressource indispensable pour réaliser la traduction automatique vers l'amazighe via l'interlangue UNL, qui fera l'objet du chapitre suivant.

Chapitre 5 : Traduction automatique multilingue vers l'amazighe

Chapitre 5

Traduction automatique vers l'amazighe

Sommaire

5.1. Introduction

Ce chapitre présentera notre contribution à la traduction automatique vers la langue amazighe via l'interlangue UNL. Nous rappelons que la traduction automatique par interlangue UNL, entre la langue source LN_S et la langue cible LN_C, consiste en l'élaboration de deux modules indépendants : le module d'analyse du texte de LN_S vers sa représentation sémantique en UNL, et le module de génération de LN_C depuis sa représentation en UNL.

Les ressources linguistiques nécessaires pour implémenter le module de génération vers l'amazighe sont : le dictionnaire UNL-Amazighe que nous avons développé dans le chapitre 4 et la mise en œuvre des règles grammaticales de transformation UNL-amazighe responsables de la transformation des graphes UNL en une structure de liste de mots, qui fera l'objet du présent chapitre.

5.2. Processus de traduction des expressions UNL vers l'amazighe

Le processus de génération d'une expression UNL vers un texte amazighe, se déroule en trois étapes : Segmentation, Tokénisation et Transformation.

5.2.1. Segmentation

La segmentation est la division du document UNL (*cf.* Figure 5.1) en une série de graphes isolés. Elle est effectuée à l'aide des balises du document UNL :

- La balise [S] définit le début d'une phrase, et [/ S] définit sa fin.
- La balise {org} définit le début de la phrase source et {/ org} définit sa fin.
- La balise {unl} définit le début du graphe UNL et {/ unl} définit sa fin.

```
[S:S#2]
  {org}
    He is arriving today
  {/org}
  {unl}
    tim(arrive:05.@present.@progressive,today:07)
    agt(arrive:05.@present.@progressive,00:01.@3.@male)
  {/unl}
[/S]
```

Figure 5. 1 Exemple d'une phrase UNL

5.2.2. Tokénisation

La tokénisation est le processus de l'identification des UWs, des relations et des attributs universels dans un graphe UNL. Par exemple, dans le graphe *tim(arrive,today)agt(arrive,00.@3.@male)* présenté dans la figure (*cf.* Figure 5.1), il y a deux relations universelles : *tim* et *agt*, quatre attributs *@present, @progressive, @male, @3* et trois UWs « arrive », « today » et l'UW vide « 00 » réservé pour les pronoms.

5.2.3. Transformation

La transformation correspond au processus permettant le passage de la représentation sémantique UNL, en tant qu'une structure arborescente, en un texte amazighe, en tant qu'une structure de liste de mots (*cf.* Figure 5.2)

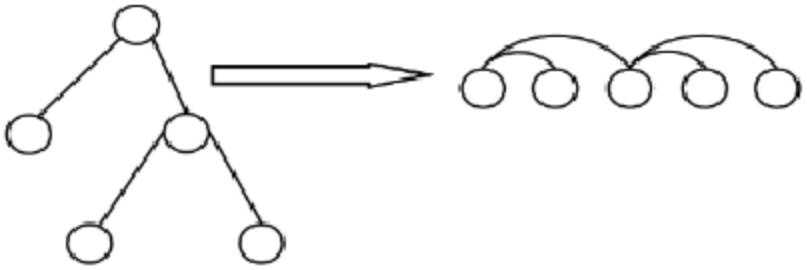

Figure 5. 2 Transformation de structures lors du processus de déconversion

Le passage entre ces deux structures se fait à travers l'application d'un ensemble de règles grammaticales de transformation sur les nœuds du graphe. Ces règles sont de trois types[10] :

- Les règles de formation de la Structure Profonde (SP), c'est à dire les règles pour l'interprétation sémantique du graphe qui transforment les relations sémantiques en des relations syntaxiques.
- Les règles du traitement syntaxique qui gèrent l'ordre et l'organisation des mots au sein d'une phrase dans la langue cible.

[10] http://www.unlweb.net/wiki/Transformation_grammar

- Les règles de formation de la Structure de Surface (SS) qui sont les règles de génération morphologique, c'est à dire les règles responsables de générer les formes fléchies adéquates selon les contextes syntaxique et lexical.

Chaque règle de transformation a le format suivant : **A:= B ;** le côté gauche A est une condition et le côté droit B est une action à effectuer sur A. L'implémentation de ces règles de transformation se base sur la théorie X-barre de la grammaire des constituants qui sera abordée dans la section suivante (*cf.* section 5.3).

5.3. La théorie X-barre

La théorie X-barre est une implémentation spécifique de la « grammaire de constituants ». C'est une méthode d'analyse syntaxique, qui divise la phrase en des groupes de mots appelés syntagmes, qui sont, à leur tour, divisés en des groupes de mots plus petits d'une manière récursive jusqu'à atteindre des constituants irréductibles (Chomsky 1970). Cette théorie affirme que tous les syntagmes sont composés d'un noyau appelé aussi une tête lexicale X qui peut être un nom (N), un verbe (V), un adjectif (J), un adverbe (A), une préposition (P) ou un déterminant (D) (*cf.* Figure 5.3).

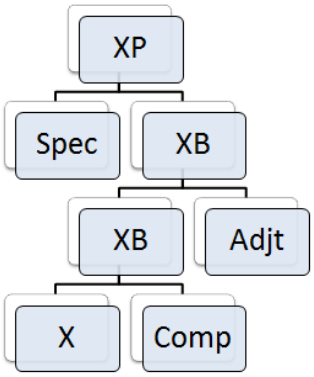

Figure 5. 3 Structure X-barre

- "**Spec**" : indique le spécificateur, qui désigne l'argument externe qui qualifie la tête du syntagme. Les spécificateurs comprennent les articles, les déterminants possessifs et démonstratifs, etc.
- "**Comp**" : indique le complément, qui désigne l'argument interne qui complète le sens introduit par la tête.

- "**Adjt**" : indique le modificateur, qui modifie la tête. Il comprend les adjectifs, les adverbes de manière, etc.
- **XB** : représente tous les constituants intermédiaires dérivés de X.
- **XP** : c'est le syntagme, la projection maximale de X (pour X= nom, il s'agit du syntagme nominal NP, pour X=verbe, il s'agit du syntagme verbal VP, etc.)

Comme le présente la figure (*cf.* Figure 5.3), le complément "Comp" se combine avec la tête X pour former la projection intermédiaire XB, le modificateur "Adjt" se combine avec XB pour former un autre XB, et le spécificateur "Spec" se combine avec le constituant intermédiaire XB le plus élevé pour construire la projection maximale XP. Par exemple, la représentation X-barre du syntagme nominal "Une belle histoire" est montrée sur la figure (*cf.* Figure 5.4). Ce syntagme peut contenir un seul spécificateur et autant de modificateurs et de compléments nécessaires.

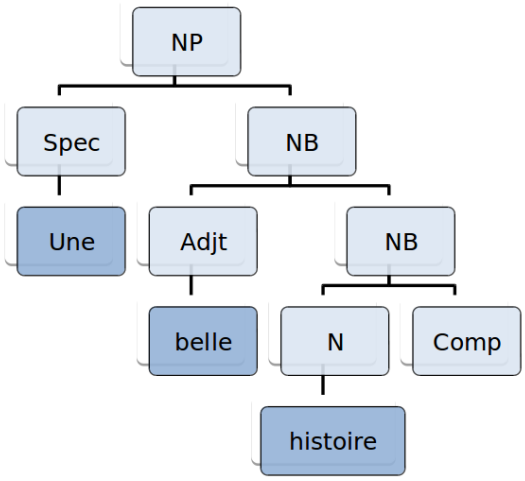

Figure 5. 4 La représentation X-barre du syntagme nominal "Une belle histoire"

De la même façon les autres syntagmes sont représentés, à savoir: le syntagme verbal (VP), le syntagme adjectival (JP), le syntagme adverbiale (AP), le syntagme prépositionnel (PP), le syntagme déterminatif (DP), le syntagme complétif (CP), qu'est la projection maximale d'une conjonction, le syntagme flexionnel (IP), qui constitue la projection maximale d'un auxiliaire.
La théorie de X-barre présente de nombreux avantages par rapport à la grammaire traditionnelle, principalement :

- L'universalité, elle peut être utilisée par différentes langues, car cette théorie prétend que toutes les langues partagent la même structure syntaxique ;
- La représentation des constituants inférieurs à XP et supérieurs à X
- La résolution des ambiguïtés structurelles syntaxiques (Carnie, 2002).

5.4. Formalisation des règles de transformation UNL-amazighe

Pour concevoir et formaliser les règles de transformation UNL-amazighe, nous nous sommes basés sur un corpus UNL, fourni par la fondation UNDL, appelé UC-A1, de taille de 50 structures UNL couvrant les phénomènes linguistiques de base : les groupes nominaux et verbaux, les cardinaux, les ordinaux, les pronoms, les déterminants, la négation, l'interrogation, etc. Nous avons utilisé l'utilitaire de déconversion EUGENE "dEep-to-sUrafce GENErator". La figure 5.5 résume la méthodologie que nous avons suivie pour élaborer la liste des règles de transformation qui seront capables de couvrir la transformation de toute structure UNL vers un texte amazighe. Finalement, nous avons pu formaliser 628 règles grammaticales de transformation.

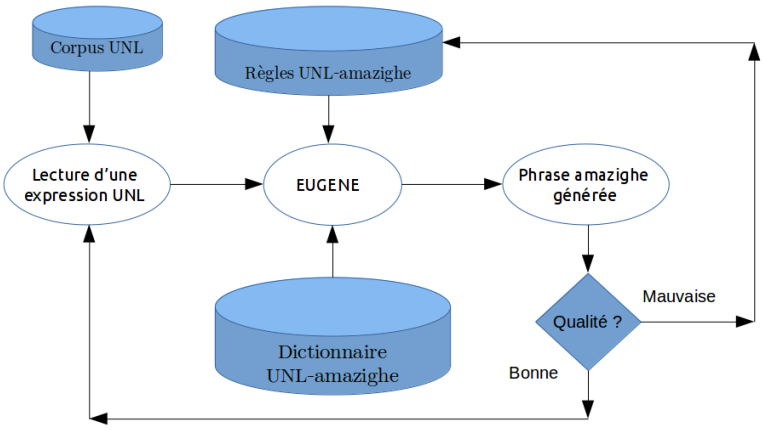

Figure 5. 5 Méthodologie de création des règles de transformation

Quelques symboles fréquemment utilisés dans l'expression d'une règle de transformation donnée sont présentés dans le tableau (*cf.* Tableau 5.1).

Tableau 5. 1 Liste des symboles de base utilisés dans le framework UNL

Symbole	Définition	Exemple
()	Nœud	(%a)
[]	Entrée du dictionnaire	[aller]
rel(x;y)	Relation	Agt (kill ; Peter)= relation qui définit que Peter est l'agent de l'UW "kill" correspondant au verbe "tuer"
^	Non	^a= non a
{|}	Ou	{a|b}=a ou b
%	Index des nœuds	%x
?	Opérateur de recherche dans le dictionnaire	?[a]

A travers un exemple d'un graphe UNL (*cf.* Figure 5.6), nous décrivons le processus de déconversion de ce graphe et nous nous arrêtons à chaque type de règle de transformation pour le présenter en détail.

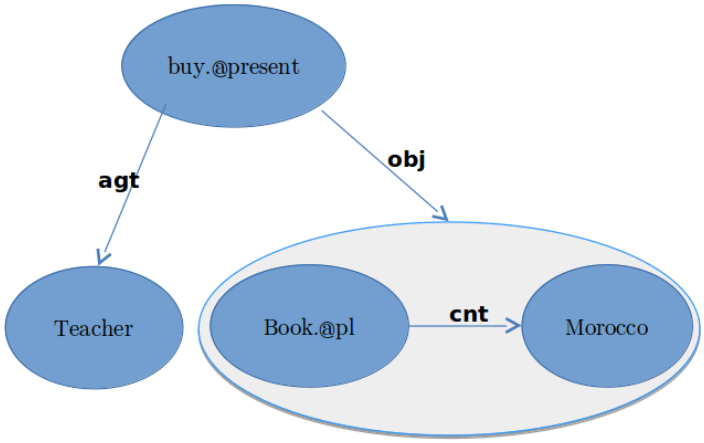

Figure 5. 6 Un exemple de graphe UNL

Dans la figure (*cf.* Figure 5.6), le graphe UNL exprime le sens porté par la phrase : "L'enseignant achète des livres sur le Maroc". En effet, l'UW "book", l'équivalent du mot

"livre", est accompagné de l'attribut universel *@pl* pour marquer que le nom est au pluriel. Cet UW est relié à l'UW "Morocco" par la relation universelle *Content (cnt)* pour exprimer que "Maroc" est le thème de l'UW "livre". L'hyper-nœud, composé des nœuds "book" et "Morocco", constitue l'objet de l'UW "buy", comme étant une entité sémantique unique. L'UW "teacher" est l'agent de l'UW "buy", qui est accompagné par l'attribut *@present* pour caractériser le temps de l'action qui est le présent.

Nous décrivons, dans ce qui suit, les étapes suivies pour transformer le graphe UNL de la (*cf.* Figure 5.6), en une liste de mots amazighes. La traduction en amazighe du sens porté par le graphe est : "ⵙⵓⵢⴰ ⵜⵓⵎⵛⴰⵍ ⵙⵍⵎⵙⵓⵏ ⵅⵜ ⵎⵛⵢⵓⵙⵏ" [isgha uslmad idlisn khf lmghrib]

5.4.1. Règles de formation de la structure profonde

Les règles de formation des structures profondes réorganisent la structure graphique sémantique en une structure arborescente syntaxique profonde. Ce sont les règles qui font le mapping des relations sémantiques vers les relations syntaxiques. Par exemple, la génération du texte amazighe correspondant au graphe sémantique dans la figure (cf. Figure 5.6) requiert les règles suivantes :

- **agt(%a,V;%b,N):=VS(%a;%b)**: cette règle transforme la relation universelle sémantique *agt* entre le verbe ⵓⵢ [sgh] "acheter", l'équivalent de l'UW "buy" indexé par %a et le nom ⴰⵓⵎⵛⴰⵍ [aslmad] "enseignant", l'équivalent de l'UW nominal "teacher" indexé par %b en la relation syntaxique *spécifieur du verbe* (VS) entre %b et %a.
- **obj(%a,V;%b):=VC(%a;%b)**: cette règle transforme la relation universelle sémantique *obj* entre ⵓⵢ [sgh] "acheter", l'équivalent de l'UW verbal "buy" indexé par %a et le syntagme nominal ⵙⵍⵎⵙⵓⵏ ⵅⵜ ⵎⵛⵢⵓⵙⵏ [idlisn khf lmghrib] "livres sur le Maroc" indexé par %b en la relation syntaxique *complément du verbe* (VC) entre %b et %a.
- **cnt(%a,N;%b,N):=NA(%a;PC(%c,[ⵅⵜ];%b)**: cette règle transforme la relation universelle sémantique *cnt* entre ⴰⵍⵎⵙⵓ [adlis] "livre", l'équivalent de l'UW "book" et ⵎⵛⵢⵓⵙⵏ [lmghrib] "Maroc", l'équivalent de l'UW "Morocco" en la relation syntaxique *modificateur du nom* (NA) entre %b, qui est introduit par la particule ⵅⵜ [khf] et %a.

5.4.2. Règles du traitement syntaxique

Les règles du traitement syntaxique sont considérées comme les plus délicates parmi les autres règles de transformation. Elles sont utilisées pour construire les projections intermédiaires (XB) qui sont combinées pour former les projections maximales (XPs). Par exemple, les relations syntaxiques VS, VC et NA obtenues à partir de l'application des règles de la première étape (*cf.* section 5.4.1), seront combinées pour former la projection maximale VP selon la théorie de X-barre. En fait, la relation NA entre ⴰⵍⵉⵙ [adlis] "livre" et ⵍⵎⵖⵔⵉⴱ [lmghrib] "Maroc", sera progressivement transformée en la projection maximale NP (syntagme nominal) passant par la projection intermédiaire NB comme il est décrit dans la figure (*cf.* Figure 5.7).

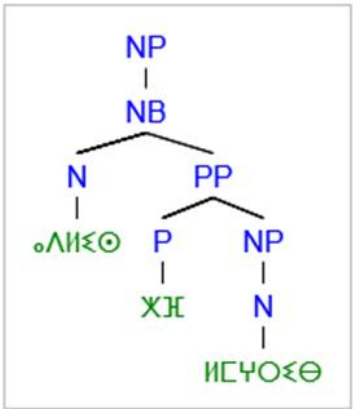

Figure 5. 7 Représentation d'un syntagme nominal

La projection maximale NP dans la figure (*cf.* Figure 5.7) constitue le complément du verbe ⵙⵖ [sgh] "acheter", comme le montre la figure (*cf.* Figure 5.8). Il sera joint au verbe pour former la projection intermédiaire VB ⵙⵖ ⴰⵍⵉⵙ ⵅⵀ ⵍⵎⵖⵔⵉⴱ [sgh adlis khf lmghrib]. Ensuite, le VB résultant sera combiné avec le spécifieur VS, qui est le nom ⴰⵙⵍⵎⴰⴷ [aslmad] "enseignant", pour former la projection maximale VP comme le présente la figure (*cf.* Figure 5.8).

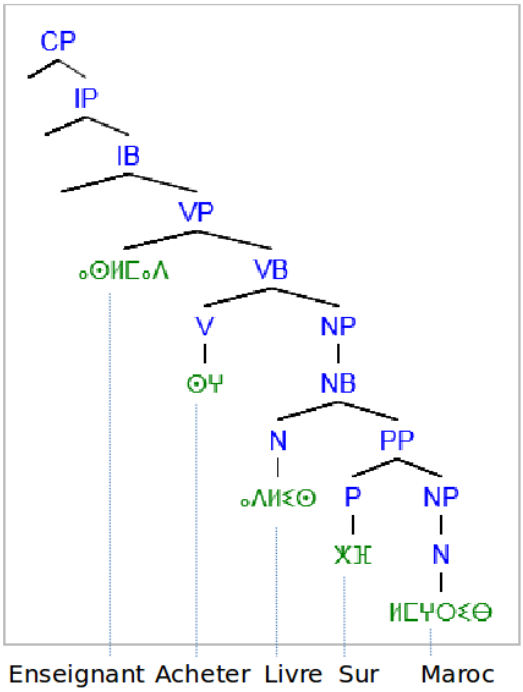

Figure 5. 8 Représentation de l'arbre syntaxique profond de la phrase

Selon la théorie X-barre, aucun spécifieur ne peut être trouvé entre le complément et la tête X. C'est-à-dire, aucun sujet ne peut intervenir entre l'objet et le verbe. En fait, l'arbre X-barre suit l'ordre des langues Sujet-Verbe-Objet (SVO). C'est vrai que la présentation, ci-dessous, de la structure syntaxique profonde donne une phrase amazighe qui est grammaticale mais cet ordre SVO reste moins fréquent par rapport à l'ordre Verbe-Sujet-Objet (VSO). Par conséquent, il est impossible de dessiner des arbres X-barre pour certaines phrases des langues VSO. Pour résoudre cette limitation, Chomsky a proposé un ensemble de règles, à savoir les règles du *mouvement tête-à-tête*[11] qui permettent le déplacement d'une tête d'une position à une autre dans l'arbre (Chomsky 1975, Carnie 2002). Par conséquent, afin d'obtenir l'ordre des mots de base VSO de l'arbre syntaxique profond présenté dans la figure (*cf.* Figure 5.8), nous déplaçons le verbe ⵙⵖ [sgh] "acheter" à la position vide de la tête **I** de IP laissant la trace t_i dans sa position initiale comme le montre la figure (*cf.* Figure 5.9), qui donne l'arbre syntaxique de surface de la phrase amazighe.

11 Mouvement tête à tête est la règle de transformation qui fait déplacer un élément, dans une phrase, de sa position d'origine vers une nouvelle position. Par exemple l'inversion du sujet par rapport au verbe, dans le français, pour exprimer l'interrogation.

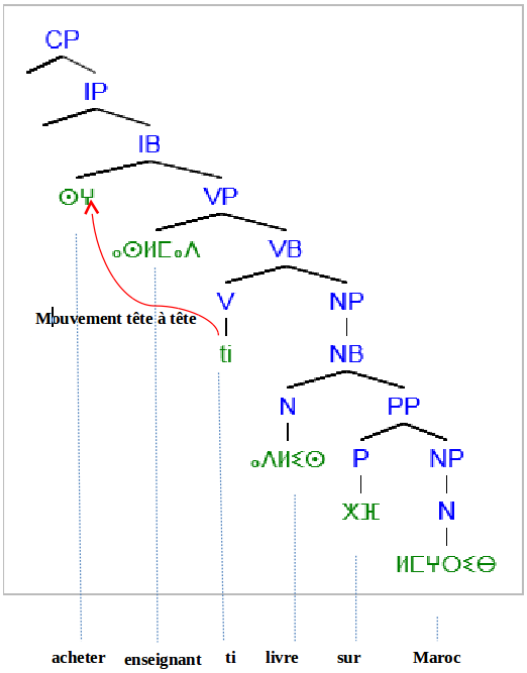

Figure 5. 9 Arbre syntaxique de surface

5.4.3. Règles de génération morphologique

Les règles, utilisées dans cette étape, ont pour rôle la linéarisation de la structure syntaxique arborescente en une structure de liste de mots. Ainsi, l'arbre syntaxique de la figure (*cf.* Figure 5.9) sera transformé en une liste de mots ⵙⵓⵢⴰ ⵙⵓⵎⵍⴰⵙ ⵙⵍⵎⵉⵙⵓⵏ ⵅⵔ ⵎⵕⵢⵓⵛⴹ [isgha uslmad idlisn khf lmghrib] "l'enseignant achète des livres sur le Maroc". ⵙⵓⵢⴰ [isgha] "achète", ⵙⵓⵎⵍⴰⵙ [uslmad] "enseignant", ⵙⵍⵎⵉⵙⵓⵏ [idlisn] "livres" ont été générés suite à l'application des règles morphologiques nécessaires respectivement sur les lemmes ⵙⵢ, ⴰⵙⵎⵍⴰⵙ, ⴰⵏⵉⵙ. Par exemple, le nom ⵙⵍⵎⵉⵙⵓⵏ [idlisn] "livres" a été généré au pluriel, car l'UW " book ", correspondant à ⴰⵏⵉⵙ portait l'attribut universel marquant le pluriel "@pl".

5.5. Etudes de cas de la traduction de l'UNL vers l'amazighe

Après avoir élaboré les deux ressources linguistiques requises pour traduire les expressions UNL vers des phrases amazighes : le dictionnaire bilingue UNL-Amazighe (8728 entrées) et les règles de transformations UNL-amazighes (628 règles grammaticales), nous avons réussi à générer les syntagmes nominaux, verbaux, les cardinaux, les phrases interrogatives, les

phrases négatives, *etc*. Dans cette section, nous discutons, en détail, la génération de quelques structures amazighes.

5.5.1. Exemple de génération d'un syntagme nominal

Dans un syntagme nominal amazighe, l'adjectif suit souvent le nom. Dans l'exemple: ⵜⵉⵥⵉⵔⵉⵜ ⵜⴰⴼⵓⵍⴽⵉⵜ [tihirit tafulkit] "belle voiture", le mot ⵜⴰⴼⵓⵍⴽⵉⵜ [tafulkit] "belle"est un adjectif amazighe qui s'accorde en genre, et en nombre avec le nom ⵜⵉⵥⵉⵔⵉⵜ[tihirit] "voiture". La figure 5.9 présente le résultat de génération d'un syntagme nominal amazighe (bloc 2) à partir de l'expression UNL (bloc 3). Le bloc 1 est la langue d'origine de l'expression UNL qui est l'anglais.

```
[S:NP#1]
   {org}
      beautiful car                    1
   {/org}
   {ber}
      ⵜⵉⵥⵉⵔⵉⵜ ⵜⴰⴼⵓⵍⴽⵉⵜ              2
   {/ber}
   {unl}
      mod(car:02,beautiful:01)         3
   {/unl}
[/S]
```

Figure 5. 10 Exemple de génération d'un syntagme nominal

Le processus de génération, de l'expression UNL, dans la figure (*cf.* Figure 5.10) passe par deux étapes:

- **Etape 1: Recherche dans le dictionnaire**

Les UWs, identifiés dans l'expression UNL, (*cf.* Figure 5.10) sont remplacés par les entrées amazighes du dictionnaire qui leur correspondent :

- [ⵜⴰⴼⵓⵍⴽⵉⵜ] "beautiful"(LEX=J,GEN=MCL,CAS=NOM,NUM=SNG,PAR=M190)
- [ⵜⵉⵥⵉⵔⵉⵜ ["car"(LEX=N,GEN=FEM,CAS=NOM, NUM=SNG, PAR=M43)

Avec:

- « LEX » désigne la catégorie lexicale. « J » pour les adjectifs, « N » pour les noms, ...etc ;
- « GEN » désigne le genre (masculin (MCL), ou féminin (FEM))
- « NUM » désigne le nombre (singulier (SNG), pluriel (PLR)),

- « CAS » désigne l'état (état libre (NOM), état d'annexion (CTS))
- « PAR » désigne l'identifiant du paradigme flexionnel. M190 et M43 spécifient respectivement quel paradigme flexionnel doit être sélectionné pour générer les formes fléchies de l'adjectif ⵓⵀⵍⵅⵣ et du nom ⵜⵣⵄⵣⵓⵣⵜ

- **Etape 2 : les principales règles de transformation appliquées**

Le tableau 5.2 décrit les principales règles appliquées pour générer la phrase amazighe de la (*cf.* Figure 5.10).

Tableau 5. 2 Principales règles appliquées pour générer le syntagme nominal de la (*cf.* Figure 5.10).

Règles appliquées	Description
mod(%noun,N;%adj, J) := NA(%noun;%adj,+NUM=%noun, +GEN=%noun ,+CAS=%noun) ;	Cette règle transforme la relation sémantique *mod* qui fait le lien entre le nœud indexé par %nom et le nœud %adj en une relation syntaxique *NA* (Noun Adjunct) qui caractérise le modificateur du nom. Comme le montre la règle, l'adjectif s'accorde avec le nom en genre, en nombre et en état d'annexion.
(%x, M190) := (%x, -M190, +FLX(MCL&NOM&SNG:=0>""; MCL&NOM&PLR:="ⵣ"<1,0>"ⵖⵍ"; MCL&CTS&SNG:="ⵙ"<1; MCL&CTS&PLR:="ⵣ"<1,0>"ⵖⵍ"; NOM&FEM&SNG:="ⵜ"<0,0>"ⵜ"; CTS&FEM&SNG:= "ⵜ"<1,0>"ⵜ"; NOM&FEM&PLR:="ⵜⵣ"<1,0>"ⵜⵣⵍ"; CTS&FEM&PLR:= "ⵜ"<1,0>"ⵜⵣⵍ";)) ;	Cette règle fait appel au paradigme flexionnel M190 pour générer la forme fléchie adéquate de l'adjectif ⵓⵀⵍⵅⵣ [afulki] 'beau' suivant les traits linguistiques rattachés à son UW à savoir le nombre, le genre,..etc
(%x,M43):=(%x,-M43,+FLX(NOM&SNG :=0> ""; NOM & PLR:=>"ⵣⵍ"; CTS&SNG:=[2]:""; CTS&PLR:= [2]:"", 0>"ⵣⵍ";)) ;	Cette règle fait appel au paradigme flexionnel M43 pour générer la forme fléchie adéquate de l'adjectif ⵜⵣⵄⵣⵓⵣⵜ [tihirit] 'voiture' suivant les traits linguistiques rattachés à son UW à savoir le nombre, le genre, etc

5.5.2. Exemple de génération d'un numéral

Les numéraux incluent les cardinaux, les ordinaux, les partitifs (quart, demi…) et les multiplicatifs (double, triple…). Au lieu d'inclure tous ces numéraux dans le dictionnaire, nous implémentons des règles grammaticales pour générer automatiquement tous ces numéraux. Par exemple, la Figure 5.11 présente le résultat de génération en amazighe du nombre 1255.

```
[S:CDN#65]
   {org}
      one thousand two hundred and fifty-five
   {/org}
   {ber}
      ⵉⴼⴹ ⴷ ⵙⵏⴰⵜ ⵜⵎⴰⴹ ⴷ ⵙⵎⵎⵓⵙ ⵉⴷ ⵎⵔⴰⵡ ⴷ ⵙⵎⵎⵓⵙ
   {/ber}
   {unl}
      [W]
         "1255":0G
      [/W]
   {/unl}
[/S]
```

Figure 5. 11 Exemple de génération d'un cardinal

- **Etape 1 : Recherche dans le dictionnaire**

Pendant la recherche dans le dictionnaire, la chaine "1255" n'a pas été trouvé, ainsi elle sera considérée comme une entrée temporaire (TEMP).

- **Etape 2 : Application des principales règles de transformation**

Le tableau 5.3 décrit les principales règles appliquées pour générer "1255" en amazighe comme le montre la (*cf.* Figure 5.11)

Tableau 5. 3 Principales règles de transformation appliquées pour générer "1255" en amazighe

Règles appliquées	Description
(%x , "1" , TEMP) := (%x ,"ⵢⴰⵏ", tous +ONE , +UNITE, -TEMP,+FLX(MCL := 0>"";FEM := 1>"ⵜ")) ;	Cette règle assigne la chaîne ⵢⴰⵏ au chiffre "1". La règle contient aussi les règles de flexion du mot ⵢⴰⵏ par rapport au genre (masculin (MCL) ou féminin (FEM))
(%x , "2" , TEMP) := (%x , "ⵙⵉⵏ" ,	Cette règle assigne la chaîne ⵙⵉⵏ au chiffre "2". La

+TWO , +UNITE , -TEMP , +GEN = MCL , +FLX(MCL := 0>"";FEM := [2]:"", 0>"ⵜ")) ;	règle contient aussi les règles de flexion du mot ⵙⵏ par rapport au genre (masculin (MCL) ou féminin (FEM))
(%x , "5" , TEMP) := (%x , "ⵙⵎⵎⵙⵙ" , +GEN = MCL , +FIVE , +UNITE , -TEMP , +PAR = M216) ;	Cette règle assigne la chaîne ⵙⵎⵎⵙⵙ au chiffre "5". La règle appelle le paradigme flexionnel M216 pour générer la forme fléchie de ⵙⵎⵎⵙⵙ par rapport au genre (masculin (MCL) ou féminin (FEM))
(%x , ONE , UNITE) (%y , UNITE) (%z , UNITE) (%w , UNITE) := (%x , "ⵣⵀⴻ") (" " , %space) ("ⴰ " , %and) (%y) (%z) (%w) ;	Cette règle transforme le premier digit "1" par ⵣⵀⴻ [ifd] 'millier', et insère la particule ⴰ entre ⵣⵀⴻ et le reste des nœuds (digits).
(%x , UNITE,^ONE) (%y ,UNITE) (%z , , UNITE) := (%x , -GEN , +GEN = FEM) (" " , %space) (%w , "ⵜⵎⵙⴻ") ("ⴰ" , %and) (%y) (%z) ;	Cette règle modifie le genre du premier digit (le nœud %x) du masculin au féminin et insère le mot ⵜⵎⵙⴻ [tmaD] 'cent' et la particule ⴰ [d] avant les nœuds %y et %z.
(%x, M216) := (%x, -M216, +FLX (MCL := ""<0; FEM := 0>"ⵜ")) ;	Cette règle concerne les cardinaux supérieurs ou égaux à « 3 ». Elle génère la forme féminine en ajoutant ⵜ[t] à la fin du mot
(%x , UNITE) (%y , UNITE) := (%x) (" " , %space) (%z , "ⵣⴰ ⵎⵙⵍ") (" " , %space) ("ⴰ" , %and) (" " , %space) (%y) ;	Cette règle génère les décimaux en insérant ⵣⴰ ⵎⵙⵍ [id mraw] après le premier digit %x and ⴰ [d] avant le dernier digit %y

Dans le tableau ci-dessus +UNIT,-UNIT, ^ONE, +ONE, +TWO, … sont des traits ajoutés au règles pour restreindre leur application.

5.5.3. Exemple de génération d'une phrase interrogative

Nous abordons maintenant la génération d'une structure plus délicate qui est la génération de la phrase interrogative. En amazighe, il existe deux types de questions :

- Les questions directes qui acceptent comme réponse oui/non. Elles sont introduites par le morphème interrogatif ⵉⵙ [is] contrairement à l'anglais et le français qui expriment les questions par l'inversion du sujet

- Les questions indirectes qui commencent souvent par les morphèmes interrogatifs comme ⵎⴰⵅ [makh] "pourquoi", ⵎⴰⵏⵉ [mani] "où", ⵎⵉⵍⵎⵉ [milmi] "quand"

"Figure 5.12" affiche le résultat de génération, en amazighe, d'une question indirecte exprimée en UNL.

```
[S:S#16]
   {org}
      Where did the boy buy the ball?
   {/org}
   {ber}
      ⵎⴰⵏⵉ ⵅ ⵉⵙⵖⴰ ⵓⵎⵊⵊⴰⵔⴷⵓ ⵜⴰⴽⵯⵔⴰ?
   {/ber}
   {unl}
      obj(buy:03.@past.@interrogative,ball:05)
      agt(buy:03.@past.@interrogative,boy:01)
      plc(buy:03.@past.@interrogative,00:02.@wh)
   {/unl}
[/S]
```

Figure 5.12 Exemple de génération d'une phrase interrogative

- **Etape 1 : Recherche dans le dictionnaire**

Les UWs, identifiés dans l'expression UNL "Fig. 5.12", sont remplacés par les entrées du dictionnaire qui leur correspondent :

- [ⴰⵎⵊⵊⴰⵔⴷⵓ]{2572}"boy"(LEX=N,POS=NOU,LST=WRD,GEN=MCL,NUM=SNG, PAR=M50, FRA=Y0)
- [ⵙⵖ]{4646}"buy"(LEX=V,POS=VER,LST=WRD,TRA=TST,PAR=M170, FRA=Y0)
- [ⵜⴰⴽⵯⵔⴰ]{2458}"ball"(LEX=N,POS=NOU,GEN=FEM,CAS=NOM,NUM=SNG, PAR=M239)

- **Etape 2 : les principales règles de transformation appliquées**

Le tableau 5.4 décrit les principales règles appliquées pour générer la phrase interrogative amazighe.

Tableau 5.4 : Principales règles de transformation appliquées pour générer la phrase interrogative présentée dans "Figure 5.12"

Règles appliquées	Description
Plc (%verb, VER, @interrogative ; %place) := VA (%01 ; %02 , +RPLC) ;	Cette règles transforme la relation sémantique "plc" en la relation syntaxique VA entre le verbe (%verb) et le modificateur (%place). RPLC est un trait ajouté au nœud %place pour le marquer afin de l'utiliser par d'autres règles.
VA(%verb,@interrogative;%place) := VB(%place; VB(%verb)) ;	Cette règle projette la projection intermédiaire (VB)
obj (%verb, V; %object, N) := VC(%verb ; %object) ;	Cette règle transforme la relation sémantique "obj" en une relation syntaxique VC entre le verbe (%verb) et son complément (%object)
VC (%verb ; %object) := VB(%verb ; %object)	Cette règle projette la première projection intermédiaire (VB).
agt (%verb , V ; %subject , N) := VS(%verb, +PER=3PS , +GEN=%subject ; %subject, -CAS , +CAS=CTS) ;	Cette règle transforme la relation sémantique "agt" en une relation syntaxique VS entre le verbe (%verb) et le nom (% subject), en assignant au sujet (%subject) l'état d'annexion et en passant au verbe (%verb) le genre et la

	personne du nœud %subject
VS (%verb ; %agent) := VP(%agent; VB(VB(% verb)));	Cette règle projette la projection maximale (VP) entre le verbe et son spécificateur %agent
VP (%agent ; VB (VB (%verb , @interrogative , ^MOVED))) VB (%verb , ^MOVED , @interrogative ; %object) VB (%place , RPLC ; VB (%verb , ^MOVED , @interrogative) , %06) := CP(%q , "ⵛⵉⵖ ⵅ" ; CB(IP(IB(%verb , -@interrogative , +MOVED ; VP(%agent ; VB(VB(%ti ; %object , +@interrogative))))))) ;	Cette règle insère le morphème interrogatif adéquat au début de la phrase et réalise l'ordre VSO en déplaçant le verbe ⵙⵖ[sgh] "acheter" vers la position vide de [I,IP] en laissant la trace t_i comme il est décrit dans (*cf.* Figure 5.9)
(%01, @interrogative) := (%01 , -@interrogative) (%02 , "?" , +QMARK) ;	Cette règle génère le "?"
(%x,M170):= (%x,-M170,+FLX ({3PS&MCL&PFV&AFM&IND} := 0>"ⴰ", "ⵖ"<0; … etc.	La règle morphologique suivante est appelée pour générer la forme conjuguée adéquate du verbe %verbe

5.6. Evaluation du système de génération UNL-amazighe

Afin de mesurer l'exactitude et la couverture des règles de transformation élaborées pour la génération de l'amazighe à partir des graphes UNL, nous nous sommes servis d'un corpus de test UNL-anglais, contenant 275 expressions UNL, extraites du corpus parallèle UC-A2[12], afin de comparer les phrases amazighes générées à partir de ces expressions UNL avec les phrases amazighes références, que nous avons traduites manuellement. La comparaison s'est

[12] UC-A2 est un corpus parallèle UNL-anglais, disponible dans le site officiel de la fondation UNDL. Il contient les différents phénomènes linguistiques

faite en calculant respectivement les métriques d'évaluation automatique « F-mesure » et « Score BLEU » (*cf.* équation 1.4 et 1.7 du premier chapitre). Les résultats obtenus sont : F-mesure=97,1 % et Score BLEU= 80%. Les phrases incorrectes concerne la génération des grands nombres supérieurs au trillion, et la synonymie qui n'est pas prise en considération. En effet, même si le dictionnaire contient les synonymes de chaque sens, mais lors de la génération, c'est le mot qui apparaît le premier dans le dictionnaire qui est sélectionné, et les autres sont négligés. Les valeurs trouvées des deux métriques montre une forte similitude entre le résultat réel et celui prévu, ce qui reflète la performance de notre système de génération UNL-amazighe. Ainsi, en principe, il est capable de traduire depuis n'importe quelle langue disposant de son propre système d'analyse tel que le cas de l'arabe, le russe,…vers la l'amazighe.

5.7. Traduction de l'arabe vers l'amazighe

Nous exploitons notre système de génération UNL-amazighe pour traduire des phrases arabes simples et autres relativement complexes vers l'amazighe en faisant appel au système d'analyse arabe-UNL (Adly et Al Ansary, 2010). La figure 5.13 présente un échantillon de traductions amazighes de quatre phrases arabes, respectivement, une phrase verbale négative, une phrase verbale affirmative, un cardinal, et une phrase nominale.

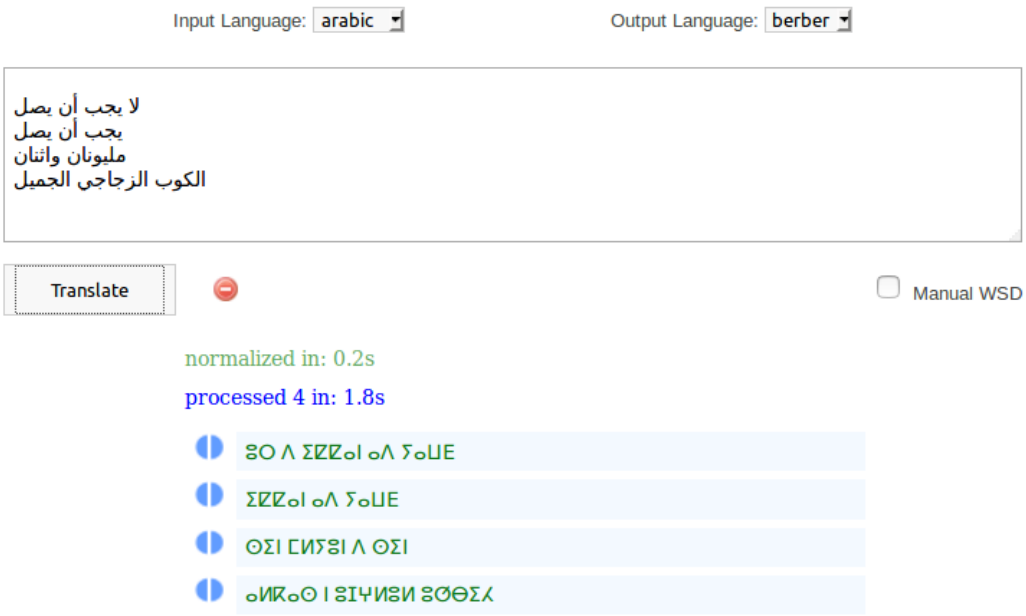

Figure 5.13 Exemple de génération d'une phrase interrogative

Pour déterminer la qualité de traduction depuis l'arabe vers l'amazighe, nous avons utilisé les deux métriques d'évaluation automatique « le score BLEU » (*cf.* équation 1.4) et la « F-mesure » (*cf.* équation 1.7). La première métrique est la plus populaire et la plus adoptée pour l'évaluation des systèmes de traduction automatique comme mentionné dans (Euro Matrix, 2007). C'est la mesure qui donne la meilleure corrélation avec l'évaluation humaine (Amahasees, 2017). C'est pour cette raison, nous l'avons utilisée pour évaluer notre système de traduction arabe-amazighe. Ainsi, le score BLEU et la F-mesure obtenus pour la traduction de 240 phrases arabes vers l'amazighe sont , respectivement, égaux à 87.94% , et 89,1% qui représentent des valeurs satisfaisantes et acceptables parmi les chercheurs dans le domaine de la TA (Euro Matrix, 2007). Pour les phrases qui ne sont pas correctement traduites, c'est dû généralement à :

- l'absence de certains mots dans le dictionnaire d'analyse arabe
- Expressions UNL mal générées par l'analyseur arabe-UNL.
- Phrases complexes contenant plusieurs clauses.

5.8. Conclusion

Dans ce chapitre, nous avons présenté notre contribution qui concerne la réalisation du module de génération UNL-amazighe. Ce travail effectué nous a permis, d'une part, de générer un corpus parallèle anglais-amazighe aligné par phrases à partir d'un corpus anglais-UNL. En d'autres parts, il nous a permis de traduire toute phrase en une langue disposant de son module d'analyse vers la langue amazighe. Dans cette thèse, nous avons fait une étude de cas pour la traduction depuis l'arabe vers l'amazighe. Actuellement, les règles grammaticales de génération UNL-amazighe implémentées sont capables de traiter des expressions UNL contenant une seule clause indépendante, et deux clauses indépendantes combinées par des conjonctions de coordination. Dans les travaux futurs, nous enrichissons la base des règles grammaticales pour prendre en considération des structures de phrases plus complexes.

Chapitre 6 : Traduction automatique Amazighe-anglais

Chapitre 6

Traduction automatique amazighe-anglais

Sommaire

6.1. Introduction

Le chapitre précédent a présenté le travail effectué pour générer des traductions amazighes depuis n'importe quelle langue, ayant son propre enconvertisseur UNL. En particulier, nous avons travaillé sur la traduction unidirectionnelle d'un corpus anglais UCA2 vers l'amazighe. Dans ce chapitre, nous présentons notre conception d'un système de traduction automatique bidirectionnel amazighe-anglais en se basant sur une approche hybride entre l'approche linguistique (par interlangue) et l'approche statistique. En effet, nous avons exploité le corpus parallèle anglais-amazighe que nous avons obtenu à partir de l'approche de traduction par interlangue pour apprendre des modèles probabilistes en vue de réalisation d'un système de traduction automatique statistique (TAS). Généralement, la TAS requière des corpus parallèles de taille importante pour assurer une bonne qualité de traduction. Cependant, la taille de notre corpus initial est relativement petite. Pour remédier à ce problème, nous avons procédé à son enrichissement par le biais d'une méthode de génération de corpus à base de dictionnaire.

6.2. L'approche hybride de la traduction automatique

Ces dernières années, plusieurs recherches ont été menées dans le domaine de la traduction automatique, principalement dans le domaine de la TA hybride. Les travaux de (Thurmair, 2005) ont montré que les erreurs faites par les différents types de systèmes de traductions sont complémentaires. En effet, si les systèmes à base de règles linguistiques (TAR) présentent une faiblesse dans la sélection lexicale, ils donnent une bonne qualité de traduction même s'il ne s'agit pas d'un domaine spécifique. Cependant, les systèmes à base des statistiques (TAS) sont plus robustes et produisent toujours de bons résultats. Ils sont meilleurs dans la sélection lexicale grâce à l'utilisation du modèle de langue extrait à partir d'un corpus spécifique, mais ils ont des difficultés pour faire face à des phénomènes qui nécessitent des connaissances linguistiques, comme la morphologie, la syntaxe, et l'ordre des mots. En outre, ils perdent l'adéquation en raison de traductions manquantes ou fausses (Vilar *et al.*, 2006). Ainsi, pour tirer profit de ces deux approches, un système hybride combinant les avantages des approches statistiques et celles fondées sur les règles linguistiques présentera la meilleure solution. Il existe plusieurs architectures d'hybridation telles que l'hybridation par couplage et l'hybridation par extension.

6.2.1. Hybridation par couplage

Une hybridation par couplage utilise différents systèmes existants TAR et / ou TAS, soit en série, ou en parallèle pour produire une sortie de traduction automatique améliorée. Dans le cas du couplage en série, l'approche la plus étudiée étant la post-édition statistique d'un système basé sur des règles. Cependant, pour le couplage en parallèle, il consiste à sélectionner la meilleure traduction produite à partir de la sortie de plusieurs systèmes (Chen, 2009).

6.2.2. Hybridation par extension

Une hybridation par extension utilise l'approche TAS ou TAR comme architecture de base et l'étends par des composants et ressources linguistiques ou statistiques. Si l'architecture des systèmes participants à l'approche d'hybridation par couplage n'est pas modifiée, dans l'hybridation par extension, l'architecture du système suit fondamentalement l'approche TAR ou TAS qui est modifiée en incluant des ressources de l'autre approche. Ces modifications peuvent se produire lors de la pré-édition, c'est-à-dire que les données du système sont prétraitées par une autre approche que celle utilisée par le système principal (Eisele, 2008).

6.3. Système de traduction automatique hybride amazighe-anglais

Dans cette section, nous proposons une conception d'un système de TA hybride anglais-amazighe qui adopte l'architecture d'hybridation par extension. Il s'agit d'un système TAS qui exploite un corpus parallèle produit par l'approche TAR pour entrainer ses modèles probabilistes : le modèle de langue et le modèle de traduction. L'architecture générale du système de traduction automatique hybride proposée est illustrée par la Figure 6.1.

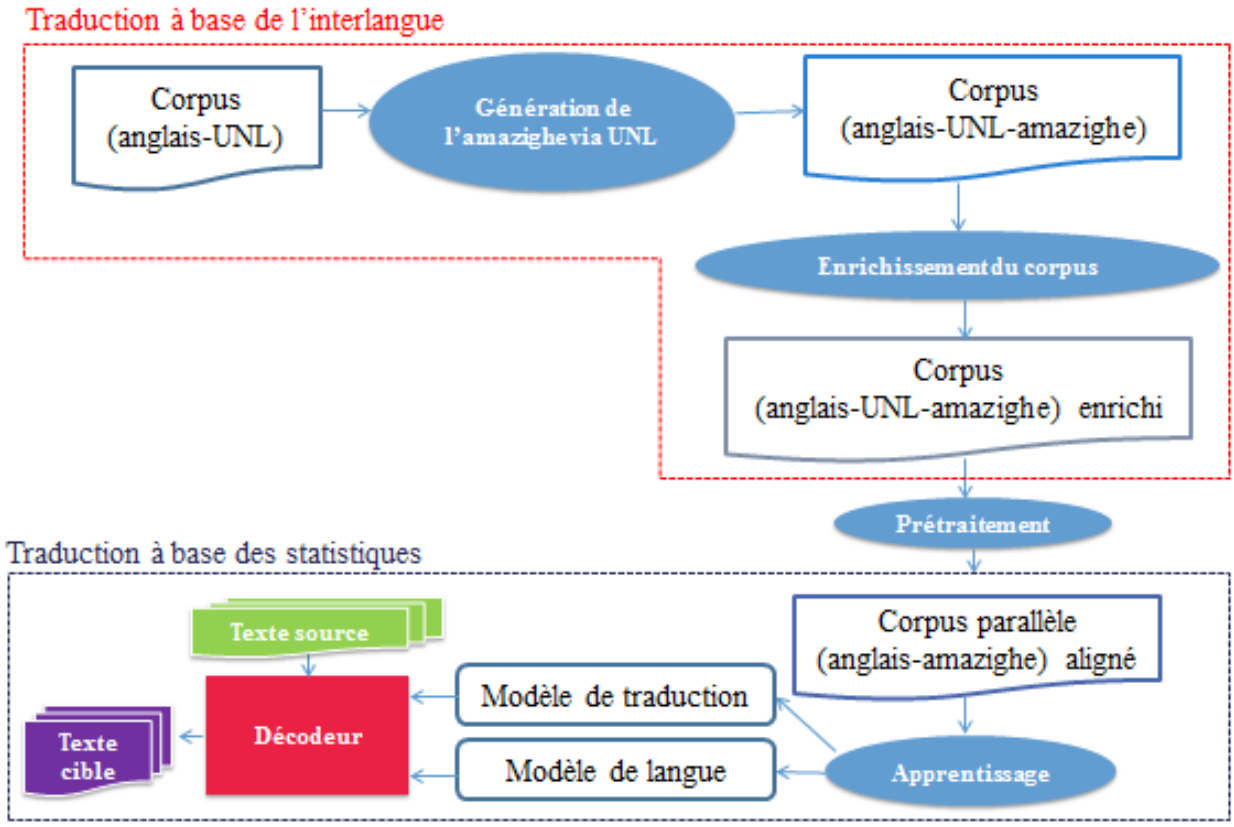

Figure 6. 1 Architecture générale du système de traduction proposé

Comme le présente la Figure 6.1, le processus de « Génération de l'amazighe via UNL » a été décrit dans la Section 5.3. Cependant, celui de « l'enrichissement du corpus », il sera abordé dans la section suivante. Pour le processus du prétraitement, il concerne l'ensemble des opérations effectuées pour nettoyer le corpus résultant (anglais-UNL-amazighe) des expressions UNL pour ressortir un corpus parallèle (anglais-amazighe), prêt pour constituer la base d'apprentissage des modèles probabilistes (cf. section 6.4). Comme nous l'avons vu dans le premier chapitre (cf. section 1.4.1.3), le décodeur est la composante qui permet de trouver la meilleure traduction candidate parmi l'ensemble des traductions candidates générées.

6.4. Enrichissement du corpus

La non-disponibilité de corpus constitue un problème crucial pour les langues peu dotées disposant de peu de ressources électroniques. En fait, la construction de corpus est une étape capitale pour la réalisation d'applications du TALN, notamment pour la TA. Dans cette section, nous présentons la méthode proposée pour enrichir le corpus parallèle (anglais-amazighe-UNL) produit par le déconvertisseur UNL-amazighe, dont la taille est de 275 phrases. La Figure 6.2 présente le format de la structure des phrases constituant ce corpus. Les

balises [S] et [\S] indiquent, respectivement, le début et la fin de la phrase UNL. La balise {org} indique le début de la phrase d'origine (la langue source), pour ce cas, l'anglais ; et {\org} indique sa fin. {unl} et {\unl}indiquent le début et la fin de l'expression UNL encodant le sens de la phrase source. {ber} et {\ber} délimitent la traduction en amazighe générée à partir de l'expression UNL.

[S:VER#1]
{org}
He used to arrive early.
{/org}
{ber}
ⵉⵍⵍⵉⵎ ⴰⵇ ⵉⵜⴰⵡⵉⴹ ⵙⵉⵖⵔⵔ
{/ber}
{unl}
agt(arrive(icl>reach):05.@habitual.@past, 00:01.@3.@male)
tim(arrive(icl>reach):05.@habitual.@past, early(equ>too soon):01)
{/unl}
[/S]

Figure 6. 2 Format d'organisation des phrases du corpus anglais-amazighe-UNL

Nous proposons, dans cette section, une méthode pour l'enrichissement du corpus (anglais-amazighe-UNL), en se basant sur le dictionnaire UNL-amazighe que nous avons implémenté (*cf.* Section 4.2 du premier chapitre). La méthode consiste à parcourir ce corpus et générer, à partir de chaque expression UNL toutes les expressions UNL qui peuvent en découler en changeant, d'une manière itérative un UW par d'autres nouveaux UWs partageant la même classe sémantique (SEM) et la même catégorie morphosyntaxique (POS) avec l'originaire. Par exemple, si nous prenons l'expression UNL suivante présentée dans la Figure 6.2 :

{unl}

agt(**arrive (icl>reach)**:05.@habitual.@past, 00:01.@3.@male)

tim(**arrive(icl>reach)**:05.@habitual.@past, **early(equ>too soon)**)

{/unl}

Elle contient l'UW *arrive (icl>reach)* qui est un concept verbal, appartenant à la classe sémantique des verbes du mouvement (SEM=MOT), l'UW *early(equ>too soon)* qui est un concept adverbial appartenant à la classe sémantique des adverbes du temps (SEM=TIME) et l'UW vide *00* qui représente les pronoms personnels sujets. A partir de ces informations sémantiques et syntaxique, nous pouvons générer (n+1)×(m+1)×9 expressions UNL avec n est le nombre d'UWs ayant SEM=MOT, m est le nombre d'UWs ayant SEM=TIME et 9 est le nombre de pronoms de personne en amazighe. Comme le présente le tableau 6.1, à partir de

l'UW *to arrive (icl>reach)*, nous récupérons 3 autres UWs et à partir de l'UW *early(equ>too soon)* nous récupérons un autre de plus. Ainsi nous avons pu à partir d'une expression régulière générer 4×2×9= 72 nouvelles expressions UNL.

Tableau 6. 1 UWs partageant les mêmes SEM et POS que ceux figurant dans l'expression UNL de la Figure 6.2

UWs	UWs ayant le même SEM et LEX	Equivalent en amazighe
To arrive (icl>reach); SEM=MOT, POS=V	**To access(icl>reach)**	ⵍⴽⵎ [lkm] "Atteindre ou accéder à"
	to travel (icl>travel)	ⵎⵎⵓⴷⴷⵓ [mmuddu]"voyager"
	to backtrack(icl>return)	ⵓⵔⵔⵉ [urri] "retourner"
early(equ>too soon); SEM=TIME, POS=AAV	**on time(man>time)**	ⵙ ⵍⵓⵇⵜ [s luqt] " A l'heure prévue"

Les nouvelles expressions UNL générées vont être soumises au déconvertisseur UNL-amazighe pour produire leurs traductions en amazighe. De cette manière, nous avons pu augmenter la taille de notre corpus d'origine de **275** à **5559** phrases. Ainsi, la taille du corpus est relativement suffisante pour apprendre des modèles probabilistes.

6.5. Mise en œuvre de la traduction amazighe-anglais

Dans cette section, nous présentons le deuxième volet de notre système de traduction automatique hybride qui est la traduction automatique statistique. Comme le montre la Figure 6.1, nous avons besoin de trois composantes essentielles pour pouvoir traduire une phrase source **S** vers une phrase cible **T,** à savoir le modèle de langue qui calcule la probabilité P(T), le modèle de traduction, qui calcule la probabilité P(S|T) et le décodeur, qui prend une phrase S et produit la phrase la plus probable **T**.

Après avoir construit le corpus parallèle anglais-amazighe, nous procédons à la mise en œuvre de notre système de traduction statistique à base de séquences de mots. Pour ce faire, plusieurs étapes sont nécessaires :

- **Etape 1 :**

La préparation des corpus concerne aussi bien le corpus monolingue pour apprendre le modèle de la langue cible que le corpus parallèle pour apprendre le modèle de traduction.

Cette préparation consiste en la segmentation lexicale du corpus, en insérant des espaces entre les mots de la phrase et la ponctuation, et en le nettoyage du corpus des lignes vides.

- **Etape 2 :**

Après la phase du prétraitement des corpus vient la phase d'apprentissage. Cette étape commence par l'alignement des textes parallèles. Chaque paire de phrases doit être alignée mot-à-mot. Cet alignement est réalisé dans les deux sens de traduction, afin de pouvoir en extraire les paires de séquences nécessaires à l'estimation du modèle de traduction. L'ensemble des alignements obtenus forment la table de traduction (cf. section 6.4.1).

- **Etape 3 :**

Construction des modèles de langues, un pour l'anglais pour assurer la traduction vers l'anglais et un autre pour l'amazighe pour assurer la traduction vers celle-ci.

La figure 6.3 résume les différentes composantes mises en jeu avec les outils que nous avons utilisés pour la TAS.

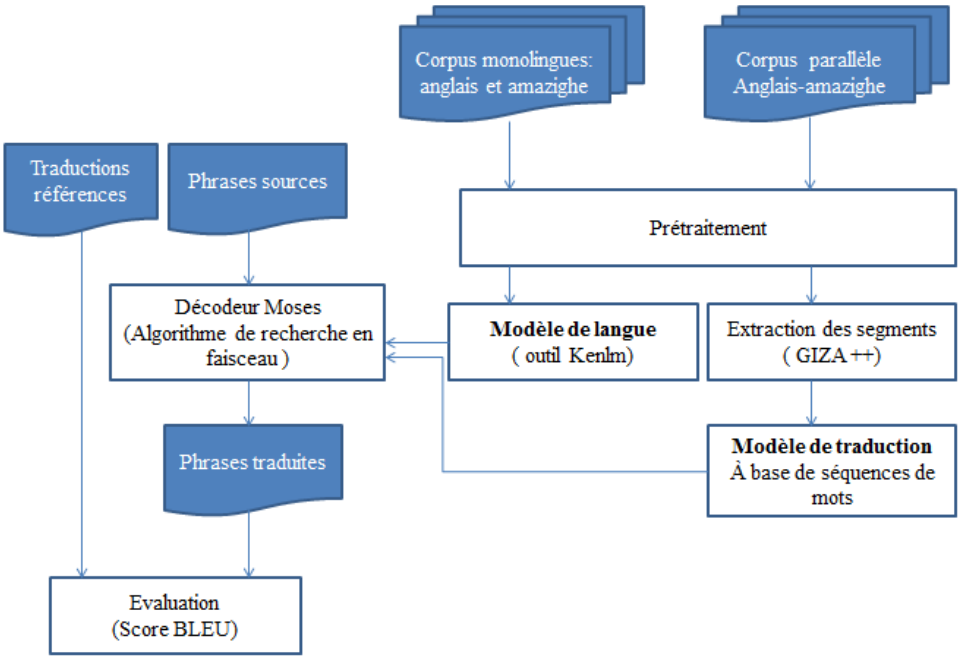

Figure 6. 3 Architecture générale de la traduction automatique statistique à base de séquences de mots

6.5.1. Construction du modèle de traduction à base de séquences de mots

La description détaillée de chacune des étapes mises en jeu dans le processus d'entrainement du modèle de traduction à base de séquences de mots est présentée dans les sous-sections suivantes. Nous avons opté pour le modèle à base de séquences de mots parce

qu'il présente le meilleur compromis qualité et rapidité par rapport aux autres modèles de traduction (Barroug, 2017).

6.5.1.1. Alignement mot à mot

L'alignement mot à mot est l'opération qui consiste à aligner le corpus d'apprentissage, déjà aligné par phrases, en alignant chaque mot du texte source avec son correspondant dans la langue cible, et vice-versa. Un alignement peut donc être vu comme un tableau où chaque indice correspond à un indice de mot dans le texte cible. Nous avons effectué l'alignement mot-à-mot en utilisant l'aligneur GIZA++ (Och et Ney, 2003), qui est un outil statistique très populaire dans la communauté TALN. Il permet d'aligner mot-à-mot les phrases correspondantes dans un corpus parallèle bilingue. Il sert notamment à apprendre des modèles probabilistes pour la traduction automatique. Il se base sur les modèles IBM de 1 à 5 (Brown *et al.*, 1993).

6.5.1.2. Extraction des paires de séquences de mots

Le résultat de l'étape d'alignement mot à mot donne lieu à deux tables : la première représente l'alignement depuis l'anglais vers l'amazighe et la deuxième représente le sens inverse (*cf.* Figure 6.4). A partir de ces deux tables, nous allons extraire des paires de séquences de mots en suivant l'heuristique proposée par (Och et *al.*, 1999). La performance de cette dernière par rapport aux autres méthodes d'extraction de paires de séquences de mots a été démontrée par (Koehn *et al.*, 2003). Cette heuristique consiste à prendre l'intersection des deux tables et l'étendre itérativement en y ajoutant, sous certaines conditions, les points d'alignement qui sont présents dans la table d'union et manquants dans la table d'intersection. En effet, un point d'alignement est ajouté s'il est adjacent d'un autre déjà présent dans l'alignement de l'intersection et si le mot source ou le mot cible constituant le point d'alignement n'est pas aligné avec aucun autre mot dans la table d'intersection. Aussi, s'il existe des points d'alignement de l'union pour lesquels ni le mot source, ni le mot cible n'est aligné dans l'intersection alors ils sont également ajoutés.

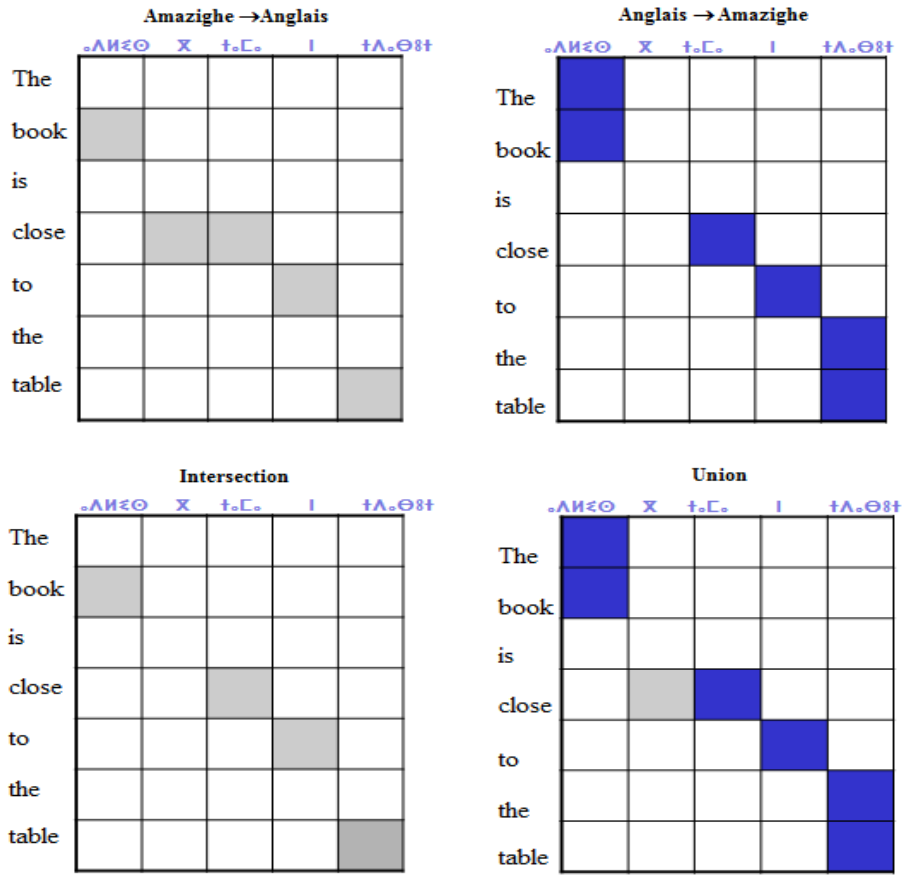

Figure 6. 4 Processus d'extraction des paires de séquences de mots selon (Ochet *al.*, 1999)

La dernière étape consiste à extraire les paires de séquences qui sont consistantes (*cf.* Figure 6.5). Une paire de séquences est dite consistante si les mots sources au sein de la paire sont alignés uniquement avec les mots cibles au sein de cette même paire et inversement.

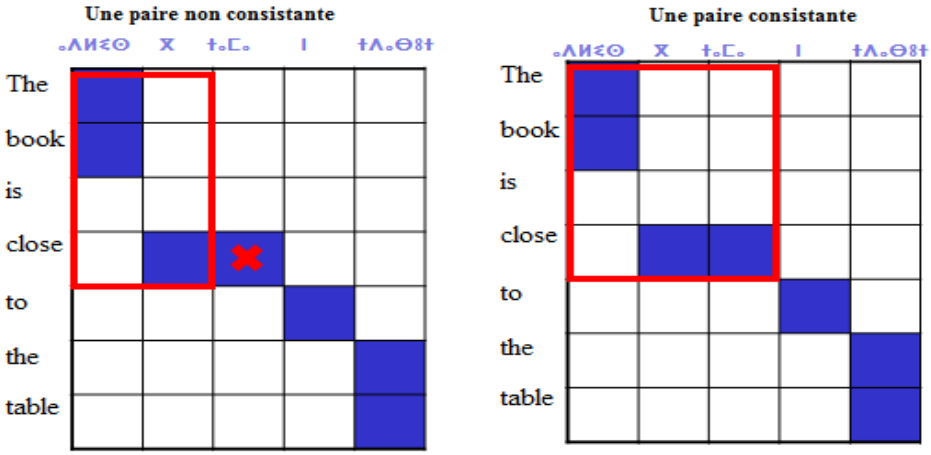

Figure 6. 5 Exemples de paires de séquences consistantes et non consistantes

La recherche des paires de séquences consistantes a induit les séquences suivantes (*cf.* Figure 6.6):

- (the book, ⴰⵍⵉⵙ[adlis]),
- (close, ⴳ ⵜⴰⵎⴰ [g tama]),
- (to, ⵏ [n]),
- (the table, ⵜⴰⴱⴷⵓⵜ[tdabut])
- (the book is close, ⴰⵍⵉⵙ ⴳ ⵜⴰⵎⴰ [adlis g tama]),
- (close to, ⴳ ⵜⴰⵎⴰ ⵏ [g tama n]),
- (to the table, ⵏ ⵜⴰⴱⴷⵓⵜ [n tdabut])
- (the book is close to, ⴰⵍⵉⵙ ⴳ ⵜⴰⵎⴰ ⵏ [adlis g tama n])
- (the book is close to the table, ⴰⵍⵉⵙ ⴳ ⵜⴰⵎⴰ ⵏ ⵜⴰⴱⴷⵓⵜ [adlis g tama n tdabut])

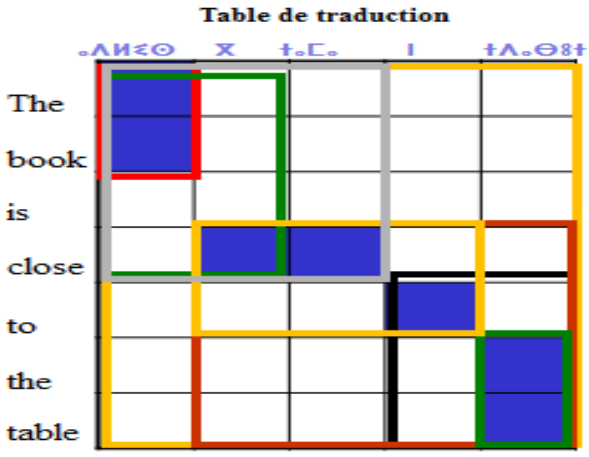

Figure 6. 6 Les differentes paires de séquences consistantes extraites

Après l'extraction des paires de séquences alignées, vient l'étape de l'estimation de leurs probabilités avec la fréquence relative à l'aide de l'équation (6.1) qui calcule le rapport entre le nombre de fois où **s** (la séquence de mots sources) et **t** (la séquence de mots cibles) sont alignés dans le corpus d'apprentissage, et le nombre de fois le segment t apparaît dans une séquence alignée à n'importe quel segment **s**

$$\mathbf{P(s,t)} = \frac{\boldsymbol{count(s,t)}}{\sum_s \boldsymbol{count(s,t)}} \quad (6.1)$$

6.5.2. Construction du modèle de langue n-gramme

Le modèle n-gramme repose sur une hypothèse qui stipule que les séquences de mots peuvent être modélisées par un processus de Markov d'ordre **n–1**. C'est-à-dire que la probabilité d'un mot dépend uniquement des **n−1** mots précédents. Avec **n** est l'ordre du modèle. La probabilité d'une séquence W de mot $\mathbf{w_i}$ dans un modèle n-gramme est calculée suivant l'équation suivante :

$$P(W) = \prod_{i=1}^{m} P(w_i | w_{i-(n-1)} \dots w_{i-1}) \quad (6.2)$$

Le modèle uni-gramme (n=1) revient à la probabilité du mot lui-même, le modèle bi-gramme (n=2) considère un historique d'un mot, le modèle trigramme (n=3) prend en compte un historique de deux mots etc. Lors de la construction d'un tel modèle, l'apprentissage consiste à estimer un ensemble de probabilités à partir d'un corpus monolingue. Généralement, l'estimation de ces probabilités se fait par l'estimateur statistique du maximum de vraisemblance; c'est-à-dire la distribution des probabilités du modèle est celle qui maximise la vraisemblance du corpus d'apprentissage. Mathématiquement, l'estimation d'un modèle N-gramme se calcule comme suit :

$$P(w_i | h_i) = \frac{f(h_i w_i)}{f(h_i)} \quad (6.3)$$

Il s'agit d'un rapport entre la fréquence d'apparition du N-gramme ($h_i w_i$) dans le corpus d'apprentissage et le nombre d'occurrences de l'historique (h_i) de ce n-gramme dans le même corpus. Il faut noter que la qualité d'un modèle de langage N-gramme dépend fortement de la quantité des données textuelles à notre disposition, mais malgré l'existence des corpus de grande taille, le problème de l'insuffisance de données reste toujours un défi pour les chercheurs. En effet, lors de l'apprentissage, nous attribuons une probabilité nulle à tout n-gramme qui n'apparaîtra pas dans le corpus d'apprentissage, même si ce n-gramme est une suite de mots possible dans la langue considérée. Pour remédier à cette problématique, il existe des techniques de lissage qui permettent d'éviter l'assignation de la probabilité nulle au n-gramme non observé, autrement dit, elles vont améliorer l'estimation des probabilités des

mots moins fréquents. Parmi ces techniques, nous pouvons citer : le lissage de Laplace (ou de Lidstone) : qui est la technique la plus utilisée dans la pratique (Chen et Goodman, 1998). Cependant, elle reste inefficace vu qu'elle surestime les probabilités des n-grammes absents dans le corpus d'entrainement (Gale, Church, 1994 ; Channoufi et Mazroui, 2014), et le lissage absolu qui fait partie des méthodes de lissage interpolé, qui est obtenu en prélevant une constante D comprise entre 0 et 1 de chaque probabilité non nulle (Ney et *al.*, 1994). Pour l'élaboration de nos modèles de langue anglais et amazighe, nous avons choisi de se servir des techniques de lissages pour contourner le problème des séquences de mots non observés et ainsi améliorer les résultats. Nous avons opté pour la deuxième méthode parce qu'elle est plus performante que la première (Channoufi et Mazroui, 2014).

6.5.3. Décodeur

Une fois que le modèle de langue et le modèle de traduction sont construits, ces deux modèles sont combinés pour trouver, pour chaque phrase source, la meilleure hypothèse de traduction. La recherche de la meilleure hypothèse de traduction consiste à choisir la phrase cible qui maximise la probabilité conditionnelle P(t|s) définie dans l'équation 6.4.

$$T^* = \operatorname{argmax}_T P(T|S) = \operatorname{argmax}_T P(S|T)\, P(T)/P(S) \qquad (6.4)$$

La recherche de la meilleure hypothèse de traduction revient donc à résoudre un problème d'optimisation combinatoire. Le décodage des modèles de traduction automatique est donc un problème NP-complet (Knight, 1999). Le choix de l'algorithme de recherche adéquat pour faire le décodage dépend essentiellement du type du modèle de traduction. Pour les modèles à base de séquences de mots, c'est l'algorithme de recherche en faisceau (beam search) qui est le plus utilisé vu sa rapidité et son efficacité par rapport aux autres algorithmes de recherche (Rush *et al.*, 2013). Cependant, il est recommandé pour le décodage des modèles à base de mots d'utiliser l'algorithme A* en raison de l'ordre de complexité important de ce modèle, vu les réorganisations possibles des mots individuels (Jelinek, 1969). Ainsi, pour décoder notre modèle de traduction à base de séquences de mots, nous avons choisi d'utiliser l'algorithme de recherche en faisceau. Pour cela, nous nous sommes servis du décodeur Moses (Koehn, 2007) qui implémente cet algorithme, c'est le décodeur de référence dans la communauté de la TA probabiliste.

6.5.4. Evaluation du système de traduction automatique

Nous avons entrainé le modèle de traduction avec 4559 couples (anglais-amazighe) de phrases. Pour les modèles de langue amazighe et anglais, nous les avons entrainés, respectivement avec un corpus monolingue amazighe de taille 4559 et un autre anglais de la même taille. Pour les deux cas, nous avons entrainé des modèles de langues 2-grammes, 3-grammes, 4-grammes, et 5-grammes pour trouver celui qui donnera les meilleures traductions.

Nous avons évalué automatiquement notre système, en traduisant un corpus de test de taille 1000 couples de phrases anglais-amazighes n'appartenant pas au corpus d'apprentissage. Le résultat de la traduction a été comparé avec les traductions références en utilisant la métrique d'évaluation le score BLEU et la F-mesure (*cf.* équation 1.4, et 1.7). Le tableau 6.2 présente les résultats obtenus, pour les deux sens de traduction : depuis l'anglais vers l'amazighe (En→Ber) et vice versa (Ber→En).

Tableau 6. 2 Variation de la qualité de traduction suivant le n-gramme utilisé pour le modèle de langue cible

	N-gramme	**2-gramme**	**3-gramme**	**4-gramme**	**5-gramme**
(Ber→En)	**Score Bleu**	90%	93%	93%	93%
	F-mesure	94%	97%	97%	97%
(En→Ber)	**Score Bleu**	82%	82%	92%	92%
	F-mesure	91%	91%	96%	96%

Les valeurs du score BLEU pour les deux directions de traduction sont satisfaisantes. La richesse morphologique de la langue amazighe et le fait que ces deux langues ne sont pas proches, et ne partageant pas les mêmes structures syntaxiques, augmentent le taux d'erreur de traduction. En effet, les variations morphologiques d'un mot amazighe donné peuvent induire en erreur. Par exemple, le même mot anglais "table" est aligné à la fois au mot amazighe ⵜⴰⵟⴰⴱⵓⵜ [tadabut] "table", dans son état morphologique libre, et au même mot ⵟⴰⴱⵓⵜ [tdabut] "table" dans son état morphologique d'annexion (*cf.* Section 3.3). La table de traduction donne P(table|ⵟⴰⴱⵓⵜ)= 0.9 et P(table|ⵜⴰⵟⴰⴱⵓⵜ)=0.09, ainsi le système pourrait afficher des phrases amazighes incorrectes morphologiquement.

6.6. Conclusion

Dans ce chapitre nous avons présenté notre approche hybride de traduction automatique anglais-amazighe. Elle consiste en la génération d'un corpus parallèle anglais-amazighe, à partir de l'approche linguistique par interlangue, pour constituer un corpus d'entrainement de taille suffisante pour apprendre des modèles statistiques de traduction à base de séquences de mots. La qualité de traduction était satisfaisante pour des couples de phrases relativement simples (non complexes). Dans les travaux futurs, nous allons travailler encore sur l'enrichissement de tel corpus avec de nouveaux couples de phrases plus longues en utilisant une des méthodes de génération automatique du texte.

Chapitre 7 : Conclusion et Perspectives

Chapitre 7

Conclusion et Perspectives

Sommaire

7.1. Conclusion générale

L'objectif principal de cette thèse était la réalisation d'un système de traduction automatique multilingue au profit de la langue amazighe. De nos jours, certes l'approche de traduction automatique à base des statistiques est celle la plus utilisée. Elle présente beaucoup d'avantages par rapport aux autres approches mais pour l'adopter, il faut se munir d'un corpus parallèle important afin d'utiliser des modèles probabilistes. Cependant, la langue amazighe est une langue peu dotée qui ne dispose que de peu de données parallèles. Partant de ce constat, nous avons proposé une approche hybride de traduction automatique qui consiste en deux volets. Le premier procède à construire des corpus parallèles en utilisant l'approche linguistique par l'interlangue UNL, et le deuxième exploite les corpus produits pour estimer des modèles de langues et de traduction capables de fournir des traductions de bonne qualité.
En effet, dans un premier temps, nous avons élaboré un module de génération UNL-amazighe qui permet de traduire depuis n'importe quelle langue LN_s, ayant son propre enconvertisseur vers la langue amazighe. En l'occurrence, la possibilité de construire des corpus parallèles LN_s-amazighe qui peuvent être exploités pour apprendre les modèles statistiques et ainsi assurer la traduction automatique statistique dans les deux sens : depuis LN_s vers l'amazighe et depuis l'amazighe vers LN_s. L'élaboration de ce module de génération a nécessité la construction d'un ensemble de ressources linguistiques, à savoir :

- Le développement des paradigmes flexionnels amazighes et la construction d'un dictionnaire bilingue UNL-amazighe ;
- La formalisation des règles grammaticales de transformation des graphes sémantiques UNL vers le texte amazighe.

Nous avons exploité le développement du dictionnaire bilingue UNL-amazighe pour mettre en œuvre un dictionnaire électronique multilingue bidirectionnel amazighe-anglais-arabe-espagnol-français. En outre, nous avons mené une étude de cas de la TA de l'anglais et de l'arabe vers l'amazighe à base de l'UNL. Par ailleurs, nous avons adopté un système de TA hybride, où nous avons adopté l'approche de traduction à base de séquences de mots pour le modèle de traduction et le modèle n-gramme pour les modèles de langue. L'évaluation a montré des résultats satisfaisants pour les deux directions de traduction à base de cette approche hybride.

7.2. Perspectives

Dans les futurs travaux, nous envisageons plusieurs perspectives à court et à moyen termes. A court terme, nous visons :

- L'enrichissement du dictionnaire électronique multilingue amazighe-anglais-arabe-espagnol-français par de nouvelles entrées.
- L'amélioration du déconvertisseur UNL-amazighe en augmentant la taille de la base des règles grammaticales par de nouvelles règles pour couvrir des structures plus complexes, c'est-à-dire des phrases à plus de trois clauses.
- La construction de nouveaux corpus français-amazighe et arabe-amazighe via la génération automatique depuis l'UNL vers l'amazighe.
- L'élaboration du module d'analyse, c'est à dire l'enconvertisseur amazighe-UNL pour pouvoir aussi effectuer l'autre sens de traduction depuis l'amazighe vers d'autres langues via l'UNL.

A moyen terme, nous proposons :

- d'étudier d'autres approches d'enrichissement de corpus parallèles
- d'utiliser le modèle statistique de traduction en se basant sur la syntaxe.
- de développer un site web dédié à la TA multilingue pour l'amazighe qui offre les fonctionnalités suivantes :
    - La TA bidirectionnelle amazighe-français et amazighe-arabe en se basant sur les statistiques.
    - La TA de l'arabe vers l'amazighe via l'interlangue UNL.

Bibliographie

Adly, N. et Al Ansary, S., (2010), "Evaluation of Arabic Machine Translation System Based on the Universal Networking Language", Natural Language Processing and Information Systems: 14th International Conference on Applications of Natural Language to Information Systems, pages 243-257.

Alansary, S. et Nagi, M., (2013), "LILY: Language-to-Interlanguage-to-Language System Based on UNL*"* , 13th International Conference on Language Engineering, Ain Shams University, Cairo, Egypt, December 11 – 12.

Agnaou, F., Bouzandag, A., El Baghdadi, M., El Gholb, H., Khalafi, A., Ouqua, K. Sghir, M., (2011), "Lexique scolaire", IRCAM, Rabat, Morocco.

Aharrane, N., El Moutaouakil, K. et Satori, K. (2015). "Recognition of handwritten Amazigh characters based on zoning methods and MLP", WSEAS Transactions on Computers, 14, pages 178-185

Ameur, M., Ansar, K., Boumalek, A., El Azrak, N., Laabdelaoui, R., Souifi, H., (2017), 1"Dictionnaire générale de la langue amazighe". IRCAM, Rabat, Morocco

Ameur, M., Bouhjar, A., Boukhris, F., Boukouss, A., Boumalk, A., Elmedlaoui, M., Iazzi, E.M., Souifi, H., (2004), "Initiation à la langue Amazighe". Rabat, Maroc: IRCAM.

Ameur, M., Bouhjar, A., Boumalk, A., El Azrak, N., Laabdelaoui, R., (2009a), "Vocabulaire des médias (Français-Amazighe-Anglais-Arabe)", IRCAM, Rabat, Morocco.

Ameur, M., Bouhjar, A., Boumalk, A., Naït-Zerrad, K., (2009b), "Amawal n tjrrumt - Vocabulaire grammatical" IRCAM, Rabat, Morocco.

Arnold, D. et Des Tombe, L., (1987). "Basic theory and methodology in EUROTRA" in Nirenburg, pages 114–135.

Arnold, D., Balkan, L., Meijer, S., Humphreys, R. L., Sadler, L., (1994), "Machine translation: An Introductory Guide". Published in the UK By NCC Blackwell Ltd., 108 Cowley Rd, Oxford OX4 IJF, and In the USA By Blackwell Publishers, 238 Main St. Cambridge, Mass. 02142. Blackwells-NCC, London.

Ataa Allah, F. et Boulaknadel, S., (2011), "Convertisseur pour la langue amazighe : script arabe -latin – tifinaghe". 2ème Symposium International sur le Traitement Automatique de la Culture Amazighe, SITACAM, Agadir, Morocco, pp. 15-23

Ataa Allah, F. et Frain, J., (2013), "Amazigh Converter based on WordprocessingML". 6th Language & Technology Conference, LTC, Poznan, Polond.

Ataa Allah, F., Boulaknadel, S., (2010a), "Amazigh Search Engine: Tifinaghe Character Based Approach". International Conference on Information and Knowledge Engineering, pages 255-

Ataa Allah, F., Boulaknadel, S., (2010b), "Pseudo-racinisation de la langue amazighe". TALN 2010, Montréal, 19-23 juillet,

Ataa Allah, F., Boulaknadel, S., (2010c). "Online Amazigh Concordancer". International Symposium on Image Video Communications and Mobile Networks. Rabat, Maroc, pages 1-4

Banerjee, S. et Lavie, A. (2005). "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". ACL Workshop on Intrinic and Extrinsic Evaluation Measures for MT and/or Summarization, pages 65-72.

Barroug, M., (2017). "Traduction automatique des textes à base des approches statistiques". Mémoire de Master, Université Mohammed V, Faculté de sciences de Rabat.

Bekios, J., Boguslavsky, I., Cardeñosa, J. , Gallardo, C., (2007), "Using Wordnet for building an Interlingua Dictionary", 5th International Conference on Information Research and Applications, (I. TECH). Varna, Bulgaria, Vol.1, pages 39-46.

Bentolila, F., (1981), "Grammaire fonctionnelle d'un parler berbère : Ayt Seghrouchen d'Oum Jeniba (Maroc)" (Paris: Selaf).

Bonnie, D. (1987). "Unitran: An interlingua approach to machine translation". In Proceedings of the 6th Conference of the American Association of Artificial Intelligence, Seattle, Washington. pages 534–539

Boukhris, F., Boumalk, A., El Moujahid, E.H., Souifi, H., (2008), "La nouvelle grammaire de l'amazighe", IRCAM, Rabat, Morocco.

Boulaknadel, S., Ataa Allah, S., (2011) "Building a Standard Amazigh Corpus", Advances in Intelligent Systems and Computing: Proceeding of the International Conference on Intelligent Human Computer Interaction (IHCI 2011), Prague, Czech Republic, August 29-31, Vol. 179/2013, pages 91-98.

Brown P.F., Della Pietra, V.J., Della Pietra, S.A., Mercer, R.L., (1993), "The mathematics of statistical machine translation: parameter estimation", Computational Linguistics. Vol. 19, no. 2.

Brown, P.F., Cocke, J., Della Pietra, S.A., Della Pietra, V.J., Jelinek, F., Lafferty, J.D., Mercer, R.L., Roossin, P.S., (1990), "A Statistical Approach to Machine Translation". Computational Linguistics 16(2), pages 79–85

Cadi, K., (1987), "Système verbal rifain : Forme et sens linguistique tamaziqht (nord marocain) " (Peeters).

Carnie, A., (2002). "Syntax: A Generative Introduction". Oxford: Blackwell Publishers.

Chen S. F., Goodman J., (1998), "An empirical study of smoothing techniques for language modeling, Harvard University, Computer Science Group, Cambridge, MA, Tech".

Chen, Y., Jellinghaus, M., Eisele, A., Zhang, Y., Hunsicker, S., Theison , S., Federmann, Hans Uszkoreit , C., (2009), "Combining Multi-Engine Translations with Moses". 4th Workshop on SMT, Athens.

Chennoufi A., Mazroui A., (2014), "Méthodes de lissage d'une approche morpho-statistique pour la voyellation automatique des textes arabes", 21ème Traitement Automatique des Langues Naturelles, Marseille, pages 443-448

Chiang, D., (2007), "Hierarchical phrase-based translation", Computational Linguistics, 33(2), pages 201-228.

Chomsky, N., (1970), "Remarks on nominalization". In R. Jacobs and P. Rosenbaum, editors, Readings in English Transformational Grammar, pages 184–221.

Dikonov, V., (2011). "English/Russian to unl enconverter". 5th International Conference on Meaning-Text, Barcelona, Spain, pages 48–58.

Dostert, L., (1955), "The Georgetown-IBM experiment". Machine translation of languages, pages 124–135

Dugast, L., Senellart, J., Koehn, P., (2007), "Statistical post-editing on systran's rule based translation system", 2nd Workshop on Statistical Machine Translation, Association for Computational Linguistics, pages 220–223.

Eisele, A., Federmann, C., Uszkoreit, H., Saint-Amand, H., Kay, M., Jellinghaus, M., Hunsicker, S., Herrmann, T., Chen, Y., (2008), "Hybrid Machine Translation Architectures within and beyond the EuroMatrix project". 12th EAMT, Hamburg

El Azrak, N., (2009), "معجم اللغة الأمازيغية", IRCAM, Rabat, Morocco.

EL Azrak, N., EL Hamdaoui, A., (2011). "Référentiel de la Terminologie Amazighe : Outil d'aide à l'aménagement linguistique". 4 ème atelier international sur l'amazighe et les TICs. Rabat, Morocco.

El Gajoui, K., Ataa Allah, F. et Oumsis, M., (2015). "Diacritical Language OCR Based on Neural Network: Case of Amazigh Language", Procedia Computer Science, 73, pages 298–305

Elouahabi, S., Atounti, M., Bellouki, M., (2016), "Amazigh Isolated-Word speech recognition system using Hidden Markov Model toolkit (HTK) ". International Conference on Information Technology for Organizations Development (IT4OD), pages 1-7

Es Saady, Y., Ait Ouguengay, Y., Rachidi, A., Elyassa, M., Mammass, D., (2009). "Adaptation d'un correcteur orthographique existant à la langue Amazighe: cas du correcteur Hunspell", Actes du 1er symposium international sur le traitement automatique de la culture amazighe, SITACAM, Maroc.

Es Saady, Y., Rachidi, A., El Yassa, M., Mammass, D., (2011), "Amazigh Handwritten Character Recognition based on Horizontal and Vertical Centerline of Character", International Journal of Advanced Science and Technology, vol.33, pages 33-50.

EuroMatrix. (2007), "Survey of Machine Translation Evaluation. Statistical and Hybrid Machine Translation between all European Languages"

Forcada, M.L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, JA., Sánchez-Martínez, F., (2011), "Apertium: a free/open-source platform for rule-based machine translation. Machine translation", Vol. 25, no. 2, pages 127– 144.

Gehlot, A., Sharma, V., Singh, S.P., Kumar, A., (2015), "Hindi to English Transfer Based Machine Translation System", International Journal of Advanced Computer Research, Vol. 5, no. 19

Germann, U., (2001), "Fast decoding and optimal decoding for machine translation". 39th Annual Meeting on Association for Computational Linguistics, ACL '01, pages 228–235.

Giri, L., (2000). "Semantic Net Like Knowledge Structure Generation from Natural Languages". IIT Bombay, Dissertation

Hewavitharana, S. et Vogel, S., (2011), "Extracting parallel phrases from comparable data". 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web, BUCC '11, pages 61–68.

Hutchins, J., (2004), "Machine translation: from real users to research". 6th conference of the Association for Machine Translation in the Americas, AMTA, Washington, DC, USA, pages 102–114.

Hutchins, W. J. et Somers H. L., (1992), "An introduction to Machine Translation", London, Academic Press.

Jelinek, F., (1969), "Fast sequential decoding algorithm using a stack", j-IBM-JRD 13(6), pages 675–685.

Kirsty, R., (2006), "Meroitic – an Afroasiatic language?", SOAS Working Papers in Linguistics, Vol. 14, pages 169-206.

Knight, K., (1999), "Decoding complexity in word-replacement translation models". Computational linguistics, Vol. 25, no. 4, pages 607-615

Koehn, P., (2003), "Statistical phrase-based translation". In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), pages 48–54.

Koehn, P., (2004), "Pharaoh: a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models", Conference of the Association for Machine Translation in the Americas, pages 115-124.

Koehn, P., Hoang, H. , Birch, A., Callison-Burch, C., Federico, M. Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst ,E., (2007), "Moses: open source toolkit for statistical machine translation". In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (ACL '07). Association for Computational Linguistics, Stroudsburg, PA, USA, pages 177-180.

Koehn, P., (2010). "Statistical Machine Translation". Cambridge: Cambridge University Press

Kumar, P. et Rajendra K.S., (2013), "Punjabi deconverter for generating punjabi from universal networking language". Journal of Zhejiang University SCIENCE C, Vol. 14, no. 3, pages 179–196.

Laabdelaoui, R., Boumalk, A., Iazzi, M., Souifi, H., Ansar K., (2012), "Manuel de conjugaison de l'Amazighe". IRCAM, Rabat, Morocco.

Martins, R., et Vahan A., (2009), "Generative and enumerative lexicons in the unl framework". 7th International Conference on Computer Science and Information Technologies, pages. 756–,

Miftah, N., Ataa Allah, F., Taghbalout, I., (2017), "Sentence-aligned parallel corpus Amazigh-English". International Conference on Information and Communication Systems (ICICS), Irbid, Jordanie

Muraki, K., (1987), "PIVOT: A Two-Phase Machine Translation System", Machine Translation Summit- Manuscripts and Program, Japan, pages 81-83.

Nagao, M., (1984), "A framework of a mechanical translation between japanese and english by analogy principle". Artificial and human intelligence, Elithorn, A. and Banerji, R. (Eds.), Elsevier Science Publishers, B.V.

Nalawade, A., (2007), "Natural Language Generation from Universal Networking Language". IIT Bombay, thèse de doctorat IIT Bombay.

Nejme, F.Z., Boulaknadel, S. et Aboutajdine, D., (2016), "Développement de ressources pour la langue amazighe : Le Lexique Morphologique ElAmaLex". Actes de la conférence conjointe JEP-TALN-RECITAL, volume 11 : TALAF.

Ney H., Essen U., Kneser R., (1994), "On structuring probabilistic dependences in stochastic language modeling".Computer, Speech, and Language, vol. 8, pages 1-38.

Nyberg, E. H. et Mitamura, T., (1992), "The KANT system: Fast, accurate, high-quality translation in practical domains". International Conference on Computation Linguistics, pages 1069–1073.

Och, F. J., Tillmann, C., and Ney, H., (1999), "Improved alignment models for statistical machine translation". In Proceeding Of the Joint Conf. of Empirical Methods in Natural Language Processing and Very Large Corpora, pages 20–28.

Och, F.J., Ueffing, N., Ney, H., (2001), "An efficient A* search algorithm for statistical machine translation". In Data-Driven Machine Translation Workshop, pages 55–62.

Och, F.J., Ney, H., (2003), "A Systematic Comparison of Various Statistical Alignment Models". Computational Linguistics, volume 29, number 1, pages 19-51.

Oulhaj, L., (2000). "Grammaire du Tamazight", Centre Tarik ibn Zyad center for studies and research.

Papineni, K., (2001), "BLEU : a Method for Automatic Evaluation of Machine Translation". 40th Annual of the Association for Computational linguistics, pages 311–318.

Raiss, H. et Cavalli-Sforza, V., (2012), "ANMorph: Amazigh nouns morphological analyzer"; Press in Proceedings of the 5th Int. Conf. on Amazigh and ICT

Rush, A., Chang, Y., Michael Collins, M., (2013), "Optimal Beam Search for Machine Translation". Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 210–221,

Satori, H., et Elhaoussi, F., (2014). "Investigation Amazigh speech recognition using CMU

tools", International Journal of Speech Technology- Springer Journals. 17 (3), pages 235.

Schwenk H., (2009), "SMT and SPE machine translation systems for WMT'09", 4th Workshop on Statistical Machine Translation, pages 130–134.

Shannon, C. E., (1948). "A mathematical theory of communication". Bell System Technical Journal, 27, pages 379–423, 623–656.

Shannon, C. E., (2001). "A mathematical theory of communication". ACM SIGMOBILE Mobile Computing and Communications Review Vol. 5 no. 1, pages 3-55

Simard, M., (2007), "Rule-based translation with statistical phrase-based post-editing", 2nd Workshop on Statistical Machine Translation, Association for Computational Linguistics, pages 203–206

Taghbalout, I., Ataa Allah, F., El Marraki, M., (2015), "Amazigh representation in the UNL framework: Resource implementation", The International Conference on Advanced Wireless, Information, and Communication Technologies, Octobre 5-7, Tunisie, Procedia Computer Science, Vol. 73, pages 234–241.

Taghbalout, I., Ataa Allah, F., El Marraki, M., (2016a), "Towards UNL based machine translation for Amazigh language". International Journal of Computational Science and Engineering. In press.

Taghbalout, I., Ataa Allah, F., El Marraki, M., (2016b), "Amazigh verb in the Universal Networking Language". IEEE/ACS International Conference on Computer Systems and Applications, (AICCSA), Novembre 17-20, 2015, Marrakech, Maroc

Taghbalout, I., Ataa Allah, F., El Marraki, M., (2017), "Pivot-based multilingual dictionary model for under- resourced languages". International Journal of Applied Engineering Research, Vol. 12, no. 20, pages 10342-10350

Talha, M., Boulaknadel, S. et Aboutajdine, D., (2014), "RENAM: Système de Reconnaissance des Entités Nommées Amazighes", 21ème édition du Traitement Automatique des Langues Naturelles, Marseille

Teixeira Martins, R., Maria das Graças V. , (2005), "On the aboutness of UNL".Universal Network Language: Advances in Theory and Applications.Instituto Politécnico Nacional Centro denvestigación en Computación, Mexico, pages 51–63

Thurmair, G., (2005), "Hybrid Architectures for Machine Translation Systems". Language Resources and Evaluation, Vol. 39, no. 1

Thuyen, P.T.L., (2016), "Multilingual Automatic Translation Based on UNL: A Case Study for the Vietnamese Language". IEIE Transactions on Smart Processing and Computing, Vol. 5, no. 2, pages 77-84

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., Sawaf, H., (1997). "Accelerated DP based search for statistical translation". 5th European Conference on Speech Communication and Technology, pages 2667–2670

Tinsley, J., Ma, Y., Ozdowska, S., Way, A., (2008), "Matrex : the DCU MT system for WMT 2008", 3rd Workshop on Statistical Machine Translation, pages 171–174,.

Uchida, H., (1989), "ATLAS II: A Machine Translation System Using Conceptual Structures as an Interlingua". Machine Translation Summit, pages 85-92.

Uchida, H., Zhu, M. and Senta, T.D., (1999), "The UNL, a gift for a millennium". Paper presented at the Institute of Advanced Studies, the United Nations University.

Vauquois B., (1968), "A Survey of Formal Grammars and Algorithms for Recognition and Translation", FIP Congress-68.Edinburg, pages 254-260.

Vilar, D., Xu J., Fernando D'Haro, L., Ney, H., (2006), "Error Analysis of Statistical Machine Translation Output". 5th International Conference on Language Resources and Evaluation

Levenshtein, V. I., (1966), "Binary Codes Capable of Correcting Deletions, Insertions and Reversals". Soviet Physics Doklady, vol. 10, no. 8, pages 707–710.

Wang, Y. et Waibel, A., (1998), "Modeling with structures in statistical machine translation".17th international conference on Computational linguistics – Vol. 2, COLING '98, pages 1357–1363

Wang, Y., et Waibel A., (1997), "Decoding algorithm statistical machine translation", 35th ACL, pages 366-372.

White J.S., et O'connell T.A., (1994), "The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches", AMTA Conference, Columbia, MD, USA

Witkam, T., (1988), "DLT an industrial R&D project for multilingual MT". 12th International Conference on Computational Linguistics, pages 756–759.

Yakoubi, N., Frain, J. et Ataa Allah, F. (2016). "Convertisseur numérique : Tifinaghe - Braille", 7th International conference TICAM, Rabat, Morocco.

Yamada, K. et Knight, K., (2001), "A Syntax-based Statistical Translation Model". 39th Annual Meeting on Association for Computational Linguistics, pages 523–530.

Zenkouar, L., (2004), "L'écriture amazighe tifinaghe et unicode ", Etudes et Documents Berbères, MSH Paris Nord, 22, pages 175-173.

Résumé

La mondialisation a influencé considérablement l'essor de l'industrie des langues, spécialement en traduction automatique où la demande ne cesse de croître. Ainsi, les besoins en matière de systèmes de traduction automatique fiables augmentent de plus en plus. L'objectif principal de cette thèse est de réaliser un système de traduction automatique au profit de la langue amazighe dans un contexte national visant à sa promotion et son développement. En fait, l'amazighe est une langue peu dotée qui manque de ressources linguistiques électroniques nécessaires pour toute application du Traitement Automatique des Langues Naturelles (TALN) en général et pour la traduction automatique en particulier. Face à cette limitation en ressources, le choix de l'approche de traduction automatique à adopter n'était pas évident. Certes l'approche à base des statistiques est celle la plus utilisée de nos jours vu ses avantages en termes de rapidité de développement et de facilité de maintenance mais elle reste tributaire de l'existence d'un corpus parallèle de taille suffisamment importante afin de bien entrainer des modèles probabilistes capables de donner de bons résultats de traduction. Partant de ce constat, nous avons proposé dans un premier temps un système de traduction automatique unidirectionnel de l'anglais vers l'amazighe à base de l'interlangue UNL (Universal Networking Language). L'utilisation de cette approche a produit plusieurs ressources linguistiques de valeur très importante à savoir un dictionnaire électronique bidirectionnel multilingue et un corpus parallèle anglais-amazighe. Ensuite, nous avons exploité ce corpus pour constituer, après avoir été enrichi par de nouveaux couples de phrases, le corpus d'apprentissage des modèles probabilistes nécessaires pour assurer la traduction automatique statistique bidirectionnelle de l'anglais vers l'amazighe et vice-versa

Abstract

Globalization has significantly influenced the language industry development, especially in machine translation domain where demands continue to grow. As a result, the need for reliable machine translation systems is growing more and more. The main objective of this thesis is to design an Amazigh multilingual machine translation system for the benefit of Amazigh language. In fact, the Amazigh is an under-resourced language that lacks electronic linguistic resources needed for any Natural Language Processing (NLP) tool development in general and for machine translation in particular. Given this limitation on resources, the adequate choice of machine translation approach is not obvious. Certainly, the statistical based approach is the most used nowadays given its advantages in terms of speed development and ease maintenance. However, it remains dependent on the existence of an important parallel corpus in order to well train probabilistic models able to give good translation results. Thus, we proceed at first to use a rule-based machine translation founded on the UNL interlanguage (Universal Networking Language). The proposed system is unidirectional, it allows translation toward Amazighe language. In this approach, many important linguistic resources have been produced namely a multilingual bi-directional electronic dictionary and an English-Amazigh parallel corpus. In a second time, we have exploited this corpus to constitute, after been enriched with new sentences' pairs, the training corpus for producing language and translation models necessary for English-Amazighe bi-directional statistical machine translation.