Faculté des **Langues**
Université de Strasbourg

Master
TdL

Master **Technologies des langues**
Option Traitement automatique des langues

CORPUS PARALLELES ET ANNOTATION DES COLLOCATIONS

Mémoire de deuxième année
Bastien Giordano

Co-dirigé par :

Amalia Todirascu
Professeur des universités
Université de Strasbourg

Frédéric Imbert
Professeur des universités
Aix-Marseille Université

TABLE DES MATIERES

LISTE DES TABLEAUX

LISTE DES FIGURES

REMERCIEMENTS

Ce mémoire vient mettre un point final à cinq années d'une reprise d'études qui aura tenu toutes ses promesses. Au-delà de ce projet de mémoire, je souhaite que ces remerciements englobent toute cette période de ma vie.

Je voudrais tout d'abord adresser mes premiers remerciements aux co-directeurs de ce mémoire, pour leurs enseignements au cours de mon cursus d'une part, mais aussi pour leurs conseils et leurs nombreuses et méticuleuses relectures du présent travail. Par la même occasion, je souhaite remercier tous les professeurs dont j'ai pu suivre les cours depuis cinq ans, pour qui j'ai un profond respect.

Je tiens également à remercier tous les camarades dont j'ai pu croiser la route au cours de ce parcours très hybride, ainsi que mes collègues de travail et responsables des cellules Pix d'Aix-Marseille Université et de l'Université de Strasbourg.

Je remercie également ma famille pour leurs encouragements infinis et pour tout le reste.

Enfin, je voudrais remercier mon épouse non seulement pour son soutien indéfectible malgré tous les sacrifices personnels qui étaient nécessaires, mais aussi pour tous ses conseils avisés qui m'ont amené où je suis aujourd'hui.

Wer fremde Sprachen nicht kennt, weiß nichts von seiner eigenen.

Johann Wolfgang von Goethe,
Werk Maximen und Reflexionen (1833)

I. ÉTAT DE L'ART

1. INTRODUCTION

1.1. Motivations

Les expressions polylexicales *(multiword expressions)* posent de nombreux problèmes en ce qui concerne le traitement automatique des langues (TAL), ce qui a mené à qualifier leur étude de *pain in the neck* (Sag et al., 2002). De nos jours, les expressions polylexicales constituent toujours une vive problématique pour le TAL, du fait sans doute de leur hétérogénéité, et demeurent au centre de nombreuses recherches. C'est le cas du projet international PARSEME[1], dont les acteurs s'efforcent de classer les expressions polylexicales sur le plan théorique, de développer des méthodes d'identification et d'annotation, et de créer des corpus annotés pour faire avancer l'ensemble de la recherche sur la question.

Au sein de la classe des expressions polylexicales, de nombreuses catégories peuvent être identifiées : les termes complexes *(canon à eau)*, les mots composés *(porte-drapeau)*, les expressions idiomatiques *(joindre les deux bouts)*, un nombre important d'entités nommées *(Ministre des Affaires Étrangères)*, les locutions *(peu à peu)*, ou encore les collocations *(conclure un contrat, sucre lent)* (Constant et al., 2017). C'est l'étude de ces dernières, caractérisées comme telles notamment de par l'affinité lexicale que présentent leurs composants internes (Seretan, 2008), qui feront l'objet de ce projet.

Malgré les connaissances linguistiques sur les collocations, leur traitement automatique demeure une tâche ardue du fait de certaines des propriétés qu'elles présentent, rendant les relations syntaxiques ou les mesures statistiques seules insuffisantes. De nombreux outils ont été développés pour leur détection automatique : c'est le cas du `mwetoolkit` (Ramisch, 2012) qui applique des patrons lexico-syntaxiques et des mesures statistiques, ou encore *Fips* (Wehrli, 2007), qui passe par une analyse syntaxique profonde. Cependant, le manque de ressources annotées en collocations fait cruellement défaut, notamment en ce qui concerne les corpus parallèles multilingues, sans compter que la plupart des ressources annotées ne prennent en compte qu'un type de collocations. Malgré de nombreuses tentatives pour y remédier (Tutin et al., 2015), tout cela contribue à rendre la recherche relativement éparse et difficile à homogénéiser, surtout dans le cas des langues peu ou moyennement dotées.

1.2. Objectifs

Les objectifs de ce projet sont multiples. Dans un premier temps, nous nous attèlerons à la constitution d'un corpus parallèle trilingue (français, anglais, arabe) multi-genre qui rassemblera des textes alignés issus de diverses sources et traitant de divers domaines. En nous inspirant de campagnes d'annotation déjà réalisées (Erjavec, 2012; Savary et al., 2015), nous tâcherons dans un deuxième temps d'établir un guide d'annotation qui nous permettra d'annoter notre corpus en collocations en réutilisant les bases établies par des projets comme SimpleApprenant (Todirascu & Cargill, 2019). Pour ce faire, nous appliquerons un outil d'identification d'EP verbales déjà existant pour le français, dont nous corrigerons et compléterons manuellement les résultats obtenus de manière automatique. Par la suite, nous projetterons les annotations du corpus français vers les corpus anglais et arabe, avant d'utiliser

[1] https://parsemefr.lis-lab.fr/doku.php et https://typo.uni-konstanz.de/parseme/

nos données annotées en langue arabe pour évaluer l'outil d'annotation automatique précédemment utilisé, car il n'a jamais été évalué pour cette langue à notre connaissance. Enfin, tout ce travail nous permettra de réaliser une étude des collocations à travers les genres et les langues.

2. LES NOTIONS D'EXPRESSIONS POLYLEXICALES ET DE COLLOCATIONS

2.1. Les expressions polylexicales *(multiword expressions)*

La première caractéristique des expressions polylexicales se retrouve directement dans leur nom : ce sont des expressions associant plusieurs mots (au moins deux), ou plus rigoureusement plusieurs lexèmes. Cette caractéristique peut sembler évidente pour les langues comme le français ou l'anglais, pour lesquelles la séparation des mots est marquée par des espaces *(tueur à gages)*, mais elle l'est moins pour les langues comme l'allemand, où les expressions polylexicales prennent la forme de mots composés obtenus par concaténation de lexèmes *(Auftragskiller).*

La deuxième caractéristique qui fait que les expressions polylexicales sont considérées comme telles tient à leur caractère idiomatique. En effet, c'est l'idiomaticité d'une langue qui la distingue de toutes les autres, qui la rend unique. C'est aussi cela qui rend le processus d'apprentissage d'une langue étrangère si complexe et si long, et qui distingue naturellement les locuteurs natifs des apprenants. À la suite de Baldwin et Kim (2010), nous estimons que cette idiomaticité peut se manifester, de manière non-exclusive, à plusieurs niveaux :

- Lexique : Lorsqu'aucun des constituants de l'expression n'appartient au lexique de la langue en question, l'idiomaticité de l'expression est lexicalement marquée. Par exemple, *ex nihilo* est composé des lexèmes latins *ex (hors de)* et *nihilo (néant, rien),* mais aucun des deux ne fait partie du lexique français.
- Syntaxe : Sag et al. (2002) définissent l'idiomaticité syntaxique comme une syntaxe déviante manifestée par les constituants de l'expression polylexicale. Par exemple, la locution adverbiale *tout à coup* est constituée de l'adverbe *tout*, de la préposition *à* et du substantif *coup* sans déterminant. Cette association dénote une syntaxe somme toute inhabituelle et peut être paraphrasée par l'unité lexicale simple *soudainement*. Il s'agit de ce qui a été défini comme un *extragrammatical idiom* (expression idiomatique extra-grammaticale) (Fillmore et al., 1988). En revanche, des expressions telles que *formuler une hypothèse* ne sont pas syntaxiquement marquées car tous les composants suivent un patron syntaxique canonique.
- Sémantique : L'idiomaticité sémantique caractérise les expressions polylexicales dont le sens est le plus opaque. L'interprétation de l'expression ne peut pas se faire grâce au sémantisme des constituants seuls. Comment interpréter, sans connaissances extralinguistiques, l'expression *passer au bleu* ? Impossible de déterminer que cette expression signifie littéralement *oublier volontairement*. La notion de sens figuré intervient très souvent dans les expressions idiomatiques sémantiquement marquées (Baldwin & Kim, 2010), notamment lorsque l'expression dénote d'un caractère métaphorique *(se serrer les coudes)*, hyperbolique *(quand les poules auront des dents)* ou métonymique *(prêter l'oreille)*.
- Pragmatique : L'idiomaticité pragmatique se définit par l'usage d'expressions dans un contexte restreint et fixe de situations d'énonciation. C'est le cas des pragmatèmes, qui font partie des clichés linguistiques (Mel'čuk, 2013). On n'envisagerait pas de dire *bon appétit* à quelqu'un qui ne serait pas sur le point ou en train de manger (hors cas d'ironie ou de sarcasme).

- Statistique : L'idiomaticité statistique se manifeste par la fréquence élevée de cooccurrence de certaines combinaisons de lexèmes. Certains lexèmes auront une préférence lexicale très marquée statistiquement : par exemple, l'adjectif *grandiose* apparaîtra statistiquement très souvent après le substantif *paysage*, moins souvent après *spectacle*, et extrêmement rarement après *falaise*. Les raisons peuvent être de nature phonologique, sémantique, etc. Baldwin et Kim (2010) mettent en avant un autre cas d'idiomaticité statistique avec les binômes du type *noir et blanc* : en français et en anglais, l'ordre reste toujours le même dans leur usage idiomatique *(?a white and black television, ?une télévision blanc et noir)* mais ce n'est pas le cas dans toutes les langues, et l'arabe, par exemple, dira le contraire : التلفاز الأبيض والأسود *(al-tilfāz al-ʾabyaḍ wa-l-ʾaswad,* « la télévision blanc et noir »*)*.

À travers la multiplicité d'exemples fournis plus haut, on pourra remarquer que la dimension arbitraire des associations de lexèmes dans les expressions polylexicales de tous types et leur relative opacité sont ce qui les rend si difficiles à cerner et à définir. Par ailleurs, certaines catégories d'expressions polylexicales peuvent se retrouver sous différentes formes et configurations en discours, c'est-à-dire qu'elles peuvent être discontinues, que ce soit à cause de l'insertion d'un adverbe *(prendre rapidement des mesures)*, d'un modifieur adjectival *(cacher un terrible secret)* ou les deux. De plus, les expressions à tête verbale se retrouvent très souvent conjuguées en discours et peuvent également apparaître à la diathèse passive *(son offre a été refusée)*, sauf en ce qui concerne les expressions idiomatiques possédant un sens figuré comme *jeter l'éponge (*l'éponge a été jetée)* (Todirascu et al., 2019).

Outre ces deux caractéristiques principales, d'autres particularités entrent en jeu dans la classification des expressions polylexicales. Une expression polylexicale peut souvent être paraphrasée par une unité lexicale simple *(marcher sur les pieds – maltraiter / dominer)* ou avoir un sens proverbial, de vérité générale *(un « tiens » vaut mieux que deux « tu l'auras »)*. Enfin, la variation des expressions polylexicales entre les langues est immense et beaucoup ne possèdent pas d'équivalents directs et / ou transparents d'une langue à l'autre (Villavicencio et al., 2004). C'est le cas de notre dernier exemple qui se traduira en anglais par *a bird in the hand is worth two in the bush* et en arabe par بيضة اليوم أفضل من دجاجة غد *(bayḍat al-yawm ʾafḍal min dağāğat ġad*, « l'œuf d'aujourd'hui est préférable à la poule de demain »*)*.

Ainsi, de toutes ces caractéristiques, nous pouvons dresser une liste des différents types d'expressions polylexicales. Il peut s'agir de termes complexes *(lentilles de contact)*, de mots composés *(tire-bouchon)*, d'expressions idiomatiques *(prendre le taureau par les cornes)*, d'entités nommées *(Président de la République)*, de locutions *(tout de suite)*, ou encore de collocations *(annoncer une nouvelle)*. Ce sont de ces dernières que nous discuterons dans la partie suivante et dont nous ébaucherons une typologie

2.2. Typologie des collocations

Malgré l'omniprésence des expressions polylexicales dans la langue, elles n'en sont pas moins difficiles à appréhender et à cerner. Les collocations, dont le nombre est « astronomique » (Mel'čuk, 2013), forment un sous-type d'expressions polylexicales et sont d'autant moins faciles à comprendre que les définitions divergent d'un chercheur à l'autre.

S'il devait y avoir une forme de consensus sur ce qu'est une collocation, ce serait vis-à-vis de l'affinité mutuelle qu'entretiennent les constituants de la collocation, affinité révélée par la cooccurrence de ses constituants trop fréquente pour être due au hasard (Seretan, 2008). Cependant, ce genre de définitions ne peut convenir pleinement au linguiste à cause de son caractère vague. Rien n'est dit sur les propriétés linguistiques intrinsèques et communes aux collocations. Dès lors, deux approches sont envisagées.

2.2.1. Approche statistique

L'approche statistique, compte tenu de la compréhension initiale des collocations comme étant une cooccurrence de *n* termes trop fréquente pour être le fruit du hasard, semble s'imposer comme l'évidence. C'est par ailleurs pour cette même raison que beaucoup des définitions des collocations adoptent cette approche.

Qui est le linguiste à avoir le premier utilisé le terme *collocation* pour désigner ce phénomène linguistique n'est pas avéré (Bartsch, 2004), mais l'un des premiers est sans conteste le contextualiste britannique Firth (Williams, 2001). Les contours qu'il en a dessinés demeuraient cependant encore flous et c'est seulement après sa mort que ses continuateurs, souvent qualifiés de néofirthiens, ont formalisé et affiné leur définition (Evert, 2008). Parmi eux figurait Sinclair, un des pères de la linguistique de corpus. Selon lui, une « collocation est la cooccurrence de deux mots ou plus à une distance restreinte l'un de l'autre. La mesure de proximité habituelle est au maximum de quatre mots intercalés[2] » (Sinclair, 1991). Le phénomène collocatif est cependant toujours considéré de manière relativement large.

Dans un contexte de TAL, de nombreux chercheurs définissent les collocations et basent leurs recherches sur des fondations purement statistiques (Seretan, 2008). Qu'il s'agisse de Cruse, qui décrit les collocations comme des items lexicaux qui cooccurrent habituellement (Cruse, 1986), de Benson qui les envisage comme des combinaisons de mots arbitraires et récurrentes (Benson, 1990), ou encore de Smadja, qui ajoute à cette dernière définition que ces cooccurrences ont lieu trop souvent pour être aléatoires (Smadja, 1993), le phénomène collocatif se trouve enfermé dans la mesure statistique. Malgré tout, on commence à distinguer certains critères qui feront consensus.

2.2.2. Approche linguistique

Même si le phénomène collocatif a été largement traité avec une approche purement statistique et en ignorant presque totalement la dimension linguistique, ce qui peut paraître paradoxal pour un fait de langue, il n'en demeure pas moins que, pour beaucoup, cette dimension doive se positionner comme une caractéristique qu'on ne peut ignorer.

Dans le domaine du TAL, ces approches, mettant davantage en valeur les particularités linguistiques des collocations plutôt que leur traitement statistique brut, ont été bien moins fréquemment utilisées (Seretan, 2008). Il apparaît cependant que des patrons syntaxiques sont régulièrement à l'œuvre dans leur formation, ce qui a mené de plus en plus de chercheurs à considérer le phénomène collocatif non pas comme une simple cooccurrence avec une fréquence élevée, mais une cooccurrence dont les constituants entretiennent une certaine

[2] La traduction est de nous : "Collocation is the cooccurrence of two or more words within a short space of each other in a text. The usual measure of proximity is a maximum of four words intervening."

relation syntaxique (Tutin & Grossmann, 2002). En effet, de nombreux linguistes ont ajouté à leur propre définition du phénomène collocatif cette dimension syntaxique. Bartsch (2004) les considère comme des « cooccurrences récurrentes lexicalement et / ou pragmatiquement contraintes d'au moins deux éléments lexicaux qui sont en relation syntaxique directe l'un avec l'autre[3] ». Hausmann (1989), lui, fournit une liste de patrons lexicaux que peuvent suivre les collocations : nom + adjectif (épithète), substantif + verbe, verbe + substantif (objet), verbe + adverbe, adjectif + adverbe, substantif + (préposition) + substantif.

La définition des collocations résultant d'une approche linguistique la plus poussée réside peut-être dans la théorie Sens-Texte initialement développée par Aleksandr Žolkovskij et Igor Mel'čuk. Cette théorie fournit un cadre conceptuel large pour la description linguistique formelle qui se prête bien aux technologies des langues, qu'il s'agisse d'outils TAL (bien que les ressources lexicales ne soient pas complètes pour les y intégrer), de traduction automatique (TA), ou encore de lexicographie (Mel'čuk & Zholkovsky, 1988). Le lexique est crucial dans la théorie Sens-Texte : il est composé d'unités lexicales (lexèmes, phrasèmes, collocations ou semi-phrasèmes (Mel'čuk, 1995), etc.) qui peuvent présenter un lien sémantique abstrait entre elles. Ces liens sont appelés *fonctions lexicales* et couvrent des fonctions paradigmatiques (lexies qui ne sont pas en cooccurrence immédiate) et des fonctions syntagmatiques (lexies en relation de cooccurrence). Un exemple de fonction lexicale est BON(*L*), qui représente les collocations à tête nominale faisant intervenir un modifieur adjectival ou adverbial pour véhiculer le sens *bon*. Un locuteur du français sait que pour une lexie L donnée telle que BON(*conseil*) = *précieux*, tandis que BON(*choix*) = *heureux*, etc. À l'inverse, la fonction lexicale contraire ANTIBON(*L*) se formalise comme suit : ANTIBON(*critique*) = *virulente* (Mel'čuk, 1995). Là où la théorie Sens-Texte s'avère précieuse, c'est que beaucoup de ces fonctions lexicales standards s'appliquent à toutes les langues.

Ces deux approches ont également une différence de point de vue notable. En effet, l'approche statistique considère que les constituants d'une collocation entretiennent une relation symétrique et ses sympathisants ne prêtent pas attention à l'importance relative des lexèmes en jeu (Seretan, 2008). Au contraire, les partisans de l'approche linguistique considèrent que la relation liant les constituants d'une collocation est fondamentalement dissymétrique, avec une partie ayant plus d'importance que l'autre. C'est le cas de Hausmann et Mel'čuk, qui séparent les collocations en une *base* d'un côté et un *collocatif* de l'autre : la base est choisie librement pour son sens, tandis que le collocatif est choisi en fonction du sens global à exprimer et du premier composant (Hausmann & Blumenthal, 2006; Mel'čuk, 2013). On appellera cette hiérarchie l'*orientation* de la collocation, orientation qui peut être considérée dans les deux sens. En effet, comme le font remarquer Hausmann et Blumenthal (2006), il est possible de dresser une liste de bases qui fonctionnent avec le collocatif *appeler* (*un taxi, un ascenseur)*, comme on peut en dresser une autre contenant les collocatifs de la base *taxi (appeler, héler)*.

[3] La traduction est de nous : "lexically and/or pragmatically constrained recurrent co-occurrences of at least two lexical items which are in a direct syntactic relation with each other"

2.2.3. Critères consensuels

Les différentes approches et définitions mentionnées plus haut n'empêchent pas que certains critères puissent être dégagés pour tenter d'élaborer une typologie consensuelle des collocations.

Grâce aux pionniers mentionnés ci-dessus et à la suite de Tutin et Grossmann (2002), les critères suivants peuvent être retenus :

- Aspect arbitraire de la combinaison lexicale : jusque dans une certaine mesure, les collocations peuvent apparaître comme étant une association arbitraire de mots. Dans le voisinage de *lumineux*, on trouvera plus facilement et fréquemment les noms *faisceau* ou *rayon* plutôt que *fil*. Cela est d'autant plus flagrant avec les collocations imagées comme *appétit d'ogre* et *faim de loup*, qu'on peut difficilement mixer (Tutin & Grossmann, 2002). Cependant, l'approche linguistique a pu démontrer que la formation de collocations se fait *via* des patrons syntaxiques relativement prévisibles. Par exemple, les collocations à tête verbale appelleront généralement un substantif (objet) ou un adverbe, mais difficilement un adjectif.
- Transparence et absence de figement sémantique : au contraire de certaines expressions idiomatiques totalement opaques pour les locuteurs non natifs *(filer un mauvais coton)*, les collocations ne souffrent généralement pas d'une telle opacité. Même pour celles qui ne sont pas totalement transparentes, leur sens peut être déduit dans la plupart des cas grâce à l'un des constituants *(buveur invétéré)*. Toutefois, toutes ne sont pas si transparentes : *peur bleue* et *colère noire* affichent toutes les deux des adjectifs dont le sens peut difficilement être déduit (Tutin & Grossmann, 2002).
- Binarité de la collocation : pour la plupart des linguistes et autres chercheurs, les collocations sont majoritairement composées de deux éléments. Il est cependant notable de préciser que ces associations binaires peuvent combiner des mots avec des syntagmes, pas uniquement des lexies (Tutin & Grossmann, 2002). Par exemple, *un vent à décorner les bœufs* résulte de l'association du substantif *vent* avec le syntagme *à décorner les bœufs*. Tutin et Grossmann font remarquer par ailleurs que certaines collocations se combinent entre elles et forment des expressions polylexicales à plus de deux termes. Par exemple, *poser une question existentielle* résulte de l'association des collocations *poser une question* et *question existentielle.* Néanmoins, toutes les collocations ne peuvent pas fusionner de la sorte *(noyer son chagrin + chagrin d'amour ≠ ?noyer son chagrin d'amour).*
- Dissymétrie des composants et sélection lexicale : à la suite de Mel'čuk, on considère que le rapport des constituants de la collocation est dissymétrique, c'est-à-dire qu'un des composants présente une plus grande importance que l'autre. D'un côté, la base est choisie librement pour son sens (et le conserve) ; de l'autre, le collocatif dépend de la base et est sélectionné pour véhiculer le sens de l'ensemble (et peut perdre son sémantisme intrinsèque dans le processus, p. ex. *colère noire*).

Malgré ces efforts de classification et de recherche de consensus, les contours du phénomène restent flous et chacun y va de sa classification. Mel'čuk (2013) classe les collocations en deux types : standard et non standard. Les premières présentent un lien sémantique systématique,

disposant d'une fonction lexicale standard qui s'applique à de nombreuses bases différentes et produisent un nombre important de collocatifs différents *(couvrir d'applaudissements)*. Quant aux collocations non standard, le lien sémantique entre la base et le collocatif n'est plus aussi systématique, et il s'applique à peu de bases (parfois une seule), n'impliquant que très peu de collocatifs, voire qu'un seul *(année bissextile)*. Tutin et Grossmann (2002), eux, les classent en trois types : opaques, transparentes, régulières. Les collocations opaques ont un collocatif dont le sens diffère de celui dont il dispose en dehors de cette association *(peur bleue, colère noire)*. Les collocations transparentes, elles, ont un collocatif sans statut lexical qui peut être interprétable en cooccurrence avec la base *(faim de loup)*. Enfin, les collocations régulières présentent une association motivée, dont le collocatif inclut le sens de la base ou a un sens générique *(nez aquilin, la chouette hulule)*. Ainsi, lorsqu'il s'agit de se lancer dans l'étude des collocations, choisir une définition et un cadre théorique précis s'avère obligatoire.

2.3. Choix de définition

Comme nous l'avons vu, qu'il s'agisse des expressions polylexicales ou plus spécifiquement des collocations, les définitions et les approches abondent. Afin d'inscrire notre travail dans un cadre théorique précis, il convient de faire un choix arrêté sur la définition que nous appliquerons. Nous baserons notre travail en considérant les collocations comme étant des unités lexicales qui peuvent être décomposées en plusieurs lexèmes et qui témoignent d'une idiomaticité importante, que ce soit sur le plan lexical, syntaxique, sémantique ou statistique (Baldwin & Kim, 2010). Nous estimons en effet que les collocations forment un sous-ensemble d'expressions polylexicales et que leur idiomaticité, qui peut être révélée sous chacun des angles cités avant, est centrale dans leur identification. De plus, nous nous concentrerons uniquement sur les collocations à tête verbale. De ce fait, à notre définition s'ajoute la contrainte que les constituants de la collocation doivent présenter une relation de dépendance syntaxique verbo-nominale dans lesquelles le substantif peut exercer la fonction de sujet ou d'objet.

Ce cadre nous guidera à travers toute notre stratégie d'annotation du phénomène collocatif, qu'il s'agisse de la rédaction du guide ou de l'activité même d'annotation. Dans la section suivante, nous traiterons *in extenso* de cette activité fondamentale de la linguistique de corpus et du TAL. Nous discuterons dans un premier temps du processus et de l'utilité de l'annotation linguistique, ainsi que de certaines campagnes d'annotation effectuées à une échelle internationale, avant de nous pencher sur les subtilités et les difficultés inhérentes à une telle activité dans une perspective trilingue.

3. L'ANNOTATION

3.1. L'annotation linguistique

L'annotation linguistique, qu'elle soit manuelle, semi-automatique ou automatique est une part essentielle de la linguistique de corpus et du TAL. À l'origine uniquement manuelle, cette activité requérait énormément de temps et d'efforts pour créer des corpus annotés, et ces ressources demeuraient rares. De nos jours, avec le développement de la puissance de calcul et de stockage des ordinateurs, des méthodes d'annotation automatique ont été développées, rendant la quantité de corpus annotés, qu'ils soient oraux ou écrits, de plus en plus importante (Ide & Pustejovsky, 2017). Ces ressources permettent non seulement de travailler sur des théories linguistiques, mais participent également à la création d'applications dans le domaine du TAL, qui nécessitent inévitablement d'être entraînées et testées grâce à ces corpus annotés.

L'annotation peut prendre plusieurs formes, incluant « l'analyse morphologique, l'étiquetage des parties du discours et la parenthétisation syntaxique ; la segmentation et l'étiquetage phonétique ; l'annotation des disfluences, du phrasé prosodique, de l'intonation, des signes, et de la structure du discours ; le marquage de la coréférence, l'étiquetage des "entités nommées" et l'étiquetage sémantique ; et les traductions aux niveaux de la phrase et du mot[4] » (Bird & Liberman, 2001), mais cette liste n'est pas exhaustive. Quelle que soit la forme de l'annotation, elle est évidemment choisie pour les informations que l'on souhaite mettre au jour et implique l'association de notes descriptives ou analytiques avec des données langagières (Ide & Pustejovsky, 2017).

Par ailleurs et outre le type d'informations que le processus d'annotation souhaite mettre en lumière, le format de l'annotation est lui aussi polymorphe. Gries et Berez (2017) en distinguent trois différents. Tout d'abord, le format le plus fréquent, qui est largement utilisé pour l'étiquetage des parties du discours et pour la lemmatisation, est l'annotation *inline* ou *embedded*. Les unités annotées sont alors simplement ajoutées aux données du corpus, directement sur la ligne même où se trouve l'unité. Un second format ajoute également les annotations directement aux données du corpus, mais elles sont reportées sur des lignes séparées. Enfin, une troisième possibilité est de stocker les annotations à un endroit différent que celui du corpus de données, par exemple dans des bases de données relationnelles.

Malgré quelques tentatives d'uniformisation et de création d'un cadre théorique homogène pour l'annotation linguistique, la plupart des initiatives se basent sur les besoins individuels de telle ou telle recherche. Quoi qu'il en soit, un élément est capital pour mener à bien tout projet d'annotation : l'élaboration d'un guide d'annotation.

3.1.1. Le guide d'annotation

Le guide d'annotation constitue l'ensemble des lignes directrices que les annotateurs devront suivre et respecter pour mener à bien un projet. Les caractéristiques des unités à annoter et les étiquettes qui doivent leur être associées doivent être décrites dans le guide aussi

[4] La traduction est de nous : "(…) morphological analysis, part-of-speech tagging and syntactic bracketing; phonetic segmentation and labeling; annotation of disfluencies, prosodic phrasing, intonation, gesture, and discourse structure; marking of co-reference, 'named entity' tagging, and sense tagging; and phrase-level or word-level translations."

rigoureusement et exhaustivement que possible, de sorte que plusieurs annotateurs humains face au même ensemble de données annotent de manière équivalente. En effet, l'uniformité et la cohérence sont des objectifs majeurs à tout projet d'annotation, et cela passe par des lignes directrices permettant d'obtenir un accord inter-annotateur (deux annotateurs ou plus annotent la même phrase de la même manière) élevé et un accord intra-annotateur (si un annotateur rencontre la même phrase ou certains segments plusieurs fois, il les annote de la même manière) régulier (Brants, 2000). Cette mesure permettra en outre de quantifier la fiabilité et la qualité du guide, et *de facto* sa reproductibilité. En effet, elle fait partie d'une méthodologie itérative qui consiste à a) écrire les lignes directrices et b) tester la fiabilité avec l'accord inter-annotateur obtenu sur une partie des données (Artstein, 2017). Tant que la fiabilité n'est pas assez élevée, les lignes directrices doivent être révisées et les étapes répétées, jusqu'à ce que la fiabilité soit satisfaisante. La campagne d'annotation peut dès lors démarrer.

Tout projet d'annotation se doit d'avoir un guide d'annotation pour pouvoir être entrepris, mais ce guide ne doit pas forcément être créé *ex nihilo*. En effet, de très nombreuses campagnes d'annotation, dont certaines d'ampleur internationale, en ont déjà créés et ces guides peuvent servir de base ou d'inspiration audit projet. À moins qu'il ne s'agisse d'étudier un phénomène linguistique qui n'a jamais ou très rarement été traité, cette alternative peut permettre au projet de se concentrer davantage sur la partie développement que sur l'annotation en elle-même (Ide & Pustejovsky, 2017), car un nouveau guide a de grandes chances de se voir amendé très souvent au fur et à mesure de l'avancement du projet et de la probable découverte d'exceptions, d'ambiguïtés ou de difficultés qui n'avaient pas été envisagées initialement.

3.1.2. Campagnes d'annotation

Baser le guide d'annotation d'un projet de recherche sur celui d'une campagne existante, qu'elle soit passée ou en cours, peut donc s'avérer judicieux. Nous prendrons dans cette section deux exemples de projets ayant lieu à une échelle internationale : MULTEXT et PARSEME.

3.1.2.1. MULTEXT (Multilingual Text Tools and Corpora)

MULTEXT *(Multilingual Text Tools and Corpora)* est un projet financé par le Programme de Recherche et d'Ingénierie Linguistique de la Commission des Communautés Européennes qui a contribué à la fois au développement d'outils pour manipuler et analyser des corpus textuels et à la création de corpus multilingues annotés (Ide & Véronis, 1994). Parmi les intentions du projet figuraient la volonté d'établir des conventions pour l'encodage des corpus et des lignes directrices pour le développement d'outils TAL afin qu'ils soient réutilisés pour d'autres projets. Avec cet objectif de créer un corpus de référence *(gold standard)* mais également des conventions d'encodage (ou de balisage) des corpus pour uniformiser les futures recherches, le projet avait choisi de distinguer 4 niveaux : un niveau 0 avec le balisage au niveau du document (description bibliographique, jeux de caractères et entités, etc.) ; un niveau 1 marquant la structure générale du document (volume, chapitres, titres, jusqu'aux unités de paragraphe) ; un niveau 2 marquant les structures à l'intérieur des paragraphes (phrases, mots, noms, dates, etc.) ; et enfin un niveau 3 pour l'annotation linguistique à proprement parler (informations morphologiques, syntaxiques, prosodiques, alignement des textes parallèles) (Ide & Véronis, 1994).

Les normes établies par ce projet ont été reprises, affinées et adaptées à d'autres langues. C'est le cas avec le projet MULTEXT-East, qui a été ravivé périodiquement depuis la fin des années 90 et qui en est à ce jour à sa sixième version. Au fil des années, le projet a développé des ressources pour plusieurs langues d'Europe de l'Est et pour l'anglais. Leurs principaux résultats furent la création de ressources lexicales et surtout de corpus annotés multilingues, dont le plus conséquent est le corpus parallèle du roman *1984* de George Orwell (Erjavec, 2004). Au niveau de l'annotation elle-même, les ressources morphosyntaxiques déployées se divisent en trois catégories de ressources différentes (Erjavec, 2012) :

- Un guide détaillant les spécifications morphosyntaxiques qui déterminent quelles sont les descriptions morphosyntaxiques valides et ce qu'elles signifient. L'annotation *Ncfp* peut être développée comme suit : *Noun Type=common Gender=feminine Number=plural*[5].
- Les lexiques morphosyntaxiques, dont chaque entrée fournit le token, son lemme et sa description morphosyntaxique associée. Par exemple : *widowers, widower, Ncmp.*
- Le corpus annoté *1984*, dans lequel chaque mot est balisé par son lemme et sa description morphosyntaxique en contexte. Par exemple : *<w lemma="she" ana="Pp3fsn">*She*</w>*.

3.1.2.2. PARSEME (Parsing and Multiword Expressions)

Plus récemment, c'est le projet international PARSEME qui occupe une place importante dans le paysage de l'annotation pour le TAL. Qui plus est, ce projet nous intéresse tout particulièrement car il est dédié à l'annotation et au traitement automatique des expressions polylexicales. Le projet PARSEME *(PARSing and Multiword Expressions)*, présenté dans (Savary et al., 2015), rassemble un réseau scientifique multidisciplinaire dont les membres sont issus de plus de 30 pays, couvrant au total 29 langues différentes. Grâce à cette impressionnante échelle, PARSEME a trois objectifs principaux : se concentrer sur le multilinguisme dans les études linguistiques et technologiques, créer un réseau de recherche en TAL à la croisée des langues, des théories et des méthodologies, et enfin de combler le fossé entre la précision linguistique et l'efficacité computationnelle des applications TAL.

Le groupe de travail 4, dédié à l'annotation des expressions polylexicales dans des corpus arborés *(treebanks)*, c'est-à-dire des corpus annotés syntaxiquement et / ou sémantiquement, a pour objectif l'amélioration des méthodologies de construction de corpus arborés prenant en compte les expressions polylexicales et vise notamment à établir des guides d'annotation pour leur représentation dans des corpus arborés en constituants et en dépendances. Ainsi, plusieurs catégories d'expressions polylexicales ont été identifiées (nominales, verbales, adjectivales, prépositives et autres types) et certaines de ces catégories présentent des sous-types. C'est le cas des expressions polylexicales verbales, abordées et décrites par Savary et al. (2017) qui, toujours dans un contexte multilingue et au terme d'un processus d'annotation en deux temps pour améliorer le guide initial, ont dégagé 3 types d'expressions polylexicales : les universelles, vraies pour toutes les langues incluses dans le projet, comme les constructions avec verbes faibles *(faire une promenade)* ou les expressions idiomatiques *(tourner la page)* ; les quasi-

[5] http://nl.ijs.si/ME/V6/msd/html/msd-en.html

universelles, vraies pour certains groupes de langues, comme les verbes intrinsèquement réflexifs *(se prélasser)* ou les constructions verbe-particule *(to do up*, « réparer ») ; et les autres, impossibles à classer dans les 2 premiers groupes, comme *court-circuiter*.

Le processus d'annotation se fait comme suit : tout d'abord, une séquence candidate (en d'autres termes, un verbe combiné avec au moins un autre mot pouvant potentiellement former une expression polylexicale verbale) est sélectionnée, puis une liste des composants lexicalisés et de ses formes canoniques est établie. Pour déterminer si la séquence sélectionnée est bien une expression polylexicale, une phase d'identification et puis une autre de classification sont menées (Savary et al., 2018). Le guide d'annotation a été travaillé de sorte à fournir aux annotateurs des arbres de décision contenant des exemples et des tests linguistiques, dont la plupart sont génériques et applicables à toutes les langues. En plus des défis de l'annotation linguistique elle-même (catégorisation, division en unités, sporadicité, chevauchement, imbrication) (Mathet et al., 2015), les expressions polylexicales verbales en présentent des supplémentaires. En effet, elles apparaissent très régulièrement de manière discontinue *([prendre] rapidement [des mesures])*. De plus, les constructions verbe-particule et les verbes réflexifs voient parfois leurs éléments être séparés *([to do] someone [in])* (Savary et al., 2017).

Pour effectuer une campagne d'annotation à une telle échelle, l'utilisation d'un outil d'annotation capable de gérer différents alphabets, de supporter les scripts allant de droite à gauche et qui permette de faire des annotations à la fois discontinues, imbriquées ou qui se chevauchent étaient nécessaire. Ainsi, le travail d'annotation du projet PARSEME a été réalisé sur FLAT, une plateforme web qui répond non seulement aux critères mentionnés précédemment, mais qui permet également d'administrer efficacement un projet composé de plusieurs groupes de travail (Savary et al., 2018). Malgré l'exhaustivité du guide d'annotation et outre les différentes phases d'annotation indépendantes effectuées par plusieurs annotateurs différents avant le calcul de l'accord inter-annotateur, l'obtention d'une annotation de qualité pour un projet d'une telle ampleur se fait par une phase de vérification et surtout d'homogénéisation. Les résultats du projet sont impressionnants : outre une terminologie, une méthodologie et un guide d'annotation uniformisés, un corpus contenant de plus de 5,4M mots et 62 000 annotations d'expressions polylexicales verbales en 18 langues a été construit.

Ces deux exemples de projets à très grande échelle démontrent qu'énormément d'efforts restent à être fournis, notamment pour ce qui est de l'homogénéisation des conventions d'annotation. Plus spécifiquement, le projet PARSEME a mis en lumière à quel point les expressions polylexicales étaient problématiques et l'établissement d'un guide pour leur annotation une opération complexe et de longue haleine.

3.2. L'annotation trilingue des collocations

L'annotation linguistique d'une manière générale, nous l'avons vu, n'est pas une mince affaire. Lorsque le projet d'annotation porte sur un phénomène aussi omniprésent que flou comme les expressions polylexicales, l'entreprise devient d'autant plus complexe. Pour Tutin et al. (2015), c'est une entreprise stimulante et complexe, qui s'apparente à une « réflexion profonde » sur les expressions polylexicales. Par-dessus le tout, si l'on s'attelle à étudier en particulier le phénomène collocatif dans une perspective multilingue, une couche supplémentaire de difficulté est ajoutée et les spécificités intrinsèques à chaque langue, d'autant

plus si elles sont issues de familles de langues différentes, doivent être prises en compte. Dans cette section, nous nous attacherons tout d'abord à décrire certaines des spécificités concernant l'annotation des collocations pour chacune des langues de notre projet, respectivement le français, l'anglais, et l'arabe, avant de faire une synthèse des difficultés liées à cet environnement trilingue.

3.2.1. Spécificités du français

Le français et l'anglais, si on les compare à l'arabe, sont des langues qui sont proches l'une de l'autre. Elles partagent l'alphabet latin (même si le français utilise un alphabet latin augmenté avec des diacritiques), se lisent de gauche à droite et emploient des auxiliaires. Dans notre trio de langues, on peut cependant noter deux spécificités propres au français, qui pourront affecter le processus d'annotation : les constructions pronominales et la formation de la négation *ne… pas*.

3.2.1.1. Les verbes pronominaux

Les collocations verbo-nominales en français, qu'il s'agisse d'une relation sujet-verbe ou verbe-objet, peuvent se manifester dans une configuration pronominale. Selon la transitivité desdits verbes pronominaux, on pourra évidemment ne pas en tenir compte. Les verbes pronominaux comme *s'évanouir* ou *se pavaner* sont intransitifs et *de facto* n'auront pas de collocatifs. En revanche, certains verbes pronominaux transitifs peuvent tout à fait apparaître dans des collocations : *s'arroger un droit, se poser une question*.

L'annotation de ces derniers peut s'avérer ardue car le pronom réfléchi est changeant selon le sujet *(il s'arroge un droit, vous vous arrogez un droit)* et selon l'ordre syntaxique de la phrase dans laquelle la collocation apparaît, l'annotation doit se faire de manière discontinue *(quels droits vous êtes-vous arrogés ?)*. Pour cette raison, les stratégies d'annotation des verbes pronominaux sont variables : ils ne sont parfois pas pris en compte ou les pronoms réfléchis sont ignorés dans l'annotation (Todirascu et al., 2019), ou l'annotation indique d'une manière différente que le verbe apparaît sous sa forme pronominale (par exemple par un « + » ou un « - ») (Tutin, 2004).

3.2.1.2. La négation *ne… pas*

De la même manière, la négation française *ne… pas* est régulièrement occultée de l'annotation. En effet, même si les collocations verbales ne se manifestent que rarement dans leurs configurations prototypiques, elles peuvent virtuellement toutes être niées, entraînant immanquablement une discontinuité dans l'annotation. Par exemple, la collocation *hisser le pavillon* en contexte pourrait très bien donner *ils ne hissèrent pas le pavillon* ou *le pavillon n'avait pas été hissé*. Outre le fait qu'il faille prendre en considération *avait… été* dans le deuxième cas, à ce moment-là, bien que l'annotation doive se faire de manière discontinue, les deux composants de la négation ne sont généralement pas annotés afin d'obtenir à la fois un meilleur accord inter-annotateur et de meilleures performances en vue d'une détection automatique.

Notons qu'au contraire, c'est une difficulté qu'évite l'arabe. En effet, les différentes particules de la négation de la phrase verbale (**لا** *lā* pour nier le présent, **لم** *lam* ou **ما** *mā* pour le passé et لن *lan* pour le futur) se placent avant le verbe qu'elles nient, et *de facto* également avant

le sujet, qui se place après le verbe dans la phrase verbale canonique (voir *infra*). Ce n'est toutefois pas le cas de l'anglais.

3.2.2. Spécificités de l'anglais

Nous l'avons dit, le français et l'anglais partagent un certain nombre de caractéristiques. Ce dernier cependant possède quelques particularités : les constructions verbales avec des particules, la négation *not* et son auxiliaire de prédilection *do* et les constructions modales à l'aide d'auxiliaires.

3.2.2.1. Les verbes à particules

Le projet PARSEME catégorise les constructions verbe-particule dans les expressions polylexicales verbales *quasi-universelles* (Savary et al., 2017). Ces constructions ne sont évidemment pas uniques à l'anglais et, comme cela est bien stipulé dans leur guide d'annotation[6], bien qu'absentes ou rares des langues slaves et romanes, elles sont omniprésentes non seulement en anglais, en allemand, en suédois et en hongrois (liste à laquelle nous ajoutons l'arabe).

Ces constructions peuvent être difficiles à identifier et posent plusieurs défis à l'humain et *a fortiori* à la machine. La particule adjointe au verbe vient modifier totalement le sens du verbe tête, prend généralement la forme d'une préposition et n'est pas toujours adjacente au verbe (Constant et al., 2017). À ceci, Side (1990) ajoute qu'il existe un nombre conséquent de ces constructions, que le choix de la particule apparaît presque aléatoire, que beaucoup d'entre elles sont polysémiques et sont tout à fait opaques. Notons enfin qu'elles peuvent apparaître dans un contexte qui les rend ambiguës et que leur lecture peut se faire de plusieurs manières.

3.2.2.2. L'auxiliaire *do* et la négation *don't / do not*

À l'instar du français, la négation en anglais se fait grâce à un adverbe rendant le processus d'annotation plus complexe dans le cas des collocations verbales mettant en jeu un sujet et son verbe. En effet, dans ces cas-là, l'annotation devient inévitablement discontinue car l'auxiliaire *do* et la particule de négation *not* viennent s'intercaler entre le sujet et le verbe, sous une forme contractée ou pas. Par exemple, *the plane took off* à la négation donnera *the plane didn't / did not take off*. Une fois de plus, la stratégie généralement adoptée est de ne pas inclure ces négations dans l'annotation.

3.2.2.3. Les auxiliaires modaux

Tout comme l'auxiliaire *do* dans les constructions négatives ou emphatiques, d'autres auxiliaires viennent rendre les annotations des collocations verbo-nominales sujet-verbe en anglais discontinues. En effet, les auxiliaires modaux en anglais et leurs équivalents périphrastiques, qu'ils véhiculent une modalité radicale ou épistémique, se placent avant la base verbale qu'ils modifient. En reprenant l'exemple précédent de l'avion qui décolle, qu'il ait la capacité de décoller *(the plane can take off)*, l'autorisation de décoller *(the plane may take off),* une probabilité faible de décoller *(the plane might take off)* ou l'obligation de décoller *(the plane must take off)*, l'auxiliaire vient irrémédiablement s'intercaler entre le sujet et le verbe.

[6] https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.1/?page=050_Cross-lingual_tests/050_Verb-particle_constructions__LB_VPC_RB_

Notons par ailleurs que l'auxiliaire *will*, qu'il soit utilisé pour véhiculer un sens futur ou une probabilité très élevée *(the plane will take off),* doit être traité de la même manière.

3.2.3. Spécificités de l'arabe

L'arabe est une langue bien différente du français et de l'anglais. Dans le cadre d'un projet d'annotation, il est important de considérer toutes les différences que présente cette langue, tant au niveau de sa forme que de son fond et de son histoire. Avant toute chose, il conviendra de préciser que l'arabe est une langue que Dichy (1994) qualifie de *pluriglossique*, c'est-à-dire que cette dernière présente un haut degré de variation linguistique, qu'il s'agisse de désigner des variétés « hautes » et « basses » (donc liées au *prestige*) ou des dialectes régionaux. Qui plus est, les variétés et registres observés en arabe ne distinguent pas seulement des variations linguistiques telles qu'elles peuvent être observées en français ou en anglais, mais bel et bien des systèmes linguistiques différents. Ainsi, Dichy fait la distinction entre les diverses *glosses* arabes, qu'il répartit en trois niveaux : le niveau littéraire, commun à l'ensemble du monde arabe, qui comprend l'arabe littéraire classique (textes arabes médiévaux, préislamiques et coraniques) et moderne (presse, pédagogie, littérature) ; le niveau moyen, caractérisé par l'utilisation de syntagmes de parlers régionaux dans des configurations syntaxiques appartenant à l'arabe littéraire (type 1) ou l'inverse (type 2) ; et le niveau du parler, qui englobe les dialectes régionaux / nationaux (le syro-libanais, l'égyptien, etc.) et les parlers locaux (spécifiques à un village, un quartier, etc.).

Aussi, nous ne considèrerons dans notre étude que le cas de l'arabe littéraire moderne, que nous désignerons par *arabe moderne standard*[7] (*Modern Standard Arabic* ou *MSA*), tel qu'il est plus communément appelé de nos jours, à l'exclusion de tous les autres registres.

3.2.3.1. Origines singulières et emprunts

Outre les différences d'alphabet et de sens de lecture vis-à-vis des deux premières langues, la langue arabe ne déroge pas à la règle et fait partie des langues particulièrement riches en collocations (Brashi, 2005). Ce phénomène n'avait déjà pas échappé aux lexicographes classiques comme al-Ṭaʿālibī (m. 1038) ou Ibn Sīdah (m. 1066) qui avaient compilé dans leurs « dictionnaires de sens » (*Fiqh al-luġa* et *al-Muḫaṣṣaṣ*, respectivement) un nombre important de collocations et autres expressions idiomatiques (Emery, 1991). Malheureusement, les travaux de ces compilateurs ne sont guère plus adaptés à la linguistique moderne du fait de leur organisation idiosyncratique et sont à certains égards complètement obsolètes, car listant des formulations n'ayant plus cours en arabe moderne standard.

Une particularité qui est unique à l'arabe est, comme le signale Brashi (2005), que de nombreuses collocations trouvent leur source directement dans le Coran, soit au début du VII[e] siècle selon la tradition islamique. On peut mesurer l'importance de ce texte par la multiplicité des expressions qui en sont issues et qui sont toujours en usage de nos jours. C'est le cas de certaines collocations verbe-objet : عاهد عهدًا (*ʿāhada ʿahdan*, « conclure un pacte »), ضرب مثلًا (*ḍaraba maṯalan*, « donner un exemple »), ou encore قصّ قصّةً (*qaṣṣa qiṣṣatan*, « raconter une

[7] Nous utilisons indifféremment « l'arabe » ou « la langue arabe » pour désigner l'arabe moderne standard.

histoire »). C'est le cas également de certaines collocations substantif-adjectif : صديق حميم (*ṣadīq ḥamīm*, « un ami proche ») ou نصر عزيز (*naṣr ʿazīz*, « une grande victoire »).

Une autre influence, temporellement plus proche de nous cette fois-ci, est celle de la presse occidentale. En effet, un grand nombre de collocations qui n'existaient pas en arabe ont vu le jour au fil du temps dans des textes journalistiques, à tel point que Blau (1981) écrivait que « le linguiste se demande même si oui ou non ces deux langues [l'arabe et l'hébreu] ne sont pas sur le point de devenir une partie du faisceau linguistique européen »[8]. Les journalistes, pressés par le temps, auraient alors traduit des expressions littéralement, et ces dernières sont restées et ont intégré la langue. Ainsi, on trouve régulièrement des collocations verbales « transparentes » du type قتل الوقت (*qatala al-waqt*, « tuer le temps ») ou encore أبدى اهتمام (*ʾabdā ihtimām*, « montrer de l'intérêt »).

3.2.3.2. Particularités syntaxiques et morphologiques

Par ailleurs, le processus d'annotation des collocations du type verbe-objet en arabe présente inévitablement un nombre très important de collocations discontinues, du fait de la nature même de l'organisation syntaxique de cette langue. En effet, il existe deux types de phrases en arabe. D'une part, il y a la phrase nominale *(ğumla ismiyya)* qui, dans sa forme la plus simple se compose de deux parties : le *mubtadaʾ* (le thème) et le *ḫabar* (l'information). La copule est élidée mais fait la jointure entre les deux parties de la phrase. D'autre part, il y a la phrase verbale *(ğumla fiʿliyya)*, qui présente deux cas de figure.

- Le sujet est un pronom personnel, auquel cas le sujet est incorporé au verbe sous sa forme flexionnelle (Ryding, 2005). On peut faire le parallèle avec une langue romane : l'italien (*parlo*, « je parle » ; *sapranno*, « ils sauront »). Dans ces cas-là, l'ordre canonique de la phrase est [VS]O, et l'annotation d'une éventuelle collocation verbe-objet se ferait de manière continue car les deux syntagmes sont adjacents : رفعت يدها (*rafaʿat yadahā*, « elle a levé la main »).
- Le sujet est un syntagme nominal, auquel cas l'ordre canonique de la phrase devient VSO et l'annotation d'une éventuelle collocation se ferait inévitablement de manière discontinue[9] : **اقترف** الرئيس السابق **جرائم** (*iqtarafa al-raʾīs al-sābiq ğarāʾim*, « l'ancien président a commis des crimes »).

Cependant, l'arabe littéral / standard étant une langue casuelle, l'ordre des mots est finalement relativement libre et l'organisation syntaxique canonique des phrases peut tout à fait être bouleversée (Imbert & Pinon, 2008). L'ordre SVO, par exemple, s'applique dans les titres d'articles de presse et dans les phrases commençant par كان *(kāna*, particule verbale qui exprime le passé), كاد (*kāda*, « faillir ») ou encore جعل (*ğaʿala*, « se mettre à »).

Nous pouvons ajouter à ces particularités syntaxiques des singularités morphologiques, qui rendent le traitement automatique de cette langue délicat (Habash, 2010). Nous reviendrons sur ces spécificités dérivationnelles et flexionnelles *in extenso* dans la section 4. En ce qui concerne l'annotation manuelle, le caractère agglutinant de l'arabe peut poser des problèmes, dans le sens

[8] La traduction est de nous : "(…) the linguist even wonders whether or not these two languages are about to become a part of the European language bundle."
[9] Les mots en gras sont les composants de la collocation discontinue.

où les clitiques, qu'il s'agisse de pronoms objets ou d'adjectifs possessifs, sont difficilement séparables des mots auxquels ils sont attachés. Si l'on considère l'exemple suivant : ضربكم مثلها (*ḍarabakum maṯalahā* – « il vous a donné son exemple (à elle) », les clitiques pronom objet كم (*-kum*, « vous ») et adjectif possessif ـها (*-hā*, « son ») sont agglutinés au mot qui les précèdent et doivent donc être inclus dans l'annotation.

Toujours dans le domaine morphologique, les formes verbales augmentées en arabe constituent un cas très intéressant. Ces dernières peuvent en effet régulièrement synthétiser en une seule forme orthographique, modifiée par des schèmes spécifiques agissant comme des patrons ou des modèles, des expressions complexes, du fait de leur valeur sémantique intrinsèque. Dans leur livre, Imbert et Pinon (2008) en dressent un inventaire complet. Nous les illustrons dans le tableau ci-après :

Forme verbale	**Valeur sémantique**	**Schème**		**Translittération**	
		accompli	**inaccompli**	**accompli**	**inaccompli**
I[10]	Forme nue ou simple. Valeurs diverses (actions principalement)	فَعَلَ	يَفْعَلُ	*faʿala*	*yafʿalu*
II	Valeur intensive / factitive / délocutive / dénominative	فَعَّلَ	يُفَعِّلُ	*faʿʿala*	*yufaʿʿilu*
III	Valeur de participation (avec réciprocité implicite) / action unilatérale / insistance ou exagération	فَاعَلَ	يُفَاعِلُ	*fāʿala*	*yufāʿilu*
IV	Valeur factitive / dénominative	أَفْعَلَ	يُفْعِلُ	*ʾafʿala*	*yufʿilu*
V	Valeur réfléchie passive (résultatif de la forme II) / de réfléchi intérieur / dénominative	تَفَعَّلَ	يَتَفَعَّلُ	*tafaʿʿala*	*yatafaʿʿalu*
VI	Valeur de réciprocité	تَفَاعَلَ	يَتَفَاعَلُ	*tafāʿala*	*yatafāʿalu*
VII	Valeur réfléchie passive (résultatif de la forme I)	اِنفَعَلَ	يَنْفَعِلُ	*infaʿala*	*yanfaʿilu*
VIII	Valeur réfléchie de la forme I / action réalisée à son propre profit / réciprocité	اِفْتَعَلَ	يَفْتَعِلُ	*iftaʿala*	*yaftaʿilu*
IX	Être ou devenir une couleur / une difformité	اِفْعَلَّ	يَفْعَلُّ	*ifʿalla*	*yafʿallu*
X	Valeur de « chercher / demander » / estimative / réfléchie passive de la forme IV / « nommer à une fonction »	اِسْتَفْعَلَ	يَسْتَفْعِلُ	*istafʿala*	*yastafʿilu*

Tableau 1 : Formes verbales augmentées en arabe

[10] Pour la forme I, d'autres vocalisations sont possibles. À l'accompli, فَعِلَ *(faʿila)* et فَعُلَ *(faʿula)*. À l'inaccompli, يَفْعِلُ *(yafʿilu)* et يَفْعُلُ *(yafʿulu)*.

Prenons un exemple : en nous basant sur la forme trilittère simple كَسَرَ (*kasara*, « casser »), si on applique le schème de la forme II (فَعَّلَ, *faʿʿala*), dont la valeur est intensive, la deuxième radicale est géminée et on obtient كَسَّرَ (*kassara*, « briser en mille morceaux »). En lui appliquant le schème de la forme V (تَفَعَّلَ, *tafaʿʿala*), dont la valeur est réfléchie passive (et résultat de la forme I), un *tāʾ* (تـ) à l'initiale est ajouté, la deuxième radicale de la racine est géminée et on obtient تَكَسَّرَ (*takassara*, « être fragmenté »). Evidemment, toutes les racines ne peuvent pas être déclinées dans toutes les formes augmentées. De même, certaines formes augmentées sont particulièrement rares et n'apparaissent pas dans le tableau ci-dessus (formes XI à XV).

3.2.3.3. Patrons de collocations uniques

À l'instar de ce qui se passe dans les autres langues, les contours des collocations et leur classification sont tout aussi flous en arabe, à commencer par les multiples appellations pour désigner le même phénomène : التلازم *(al-talāzum*, « indissociabilité »*)*, التضامّ *(al-taḍāmm*, « proximité »*)*, ou encore المتلازمات اللفظية / المصاحبات *(al-muṣāḥabāt / al-mutalāzimāt al-lafẓiyya*, « associations / corrélations verbales »*)* (Nofal, 2012). Outre cela, les classifications du phénomène collocatif sont également sujettes à de nombreuses variations d'un linguiste à l'autre. En effet, (Brashi, 2005) cite certains qui, à la manière des lexicographes classiques, ont tenté de classer les collocations suivant leur sens, quand d'autres ont proposé une typologie basée sur la dimension stylistique des collocations ; plus récemment, Grimm (2009) a soumis une nouvelle taxonomie en divisant les collocations en groupes et sous-groupes, en considérant les éléments constitutifs et leur profil sémantique. Cependant, comme le fait remarquer Izwaini (2015), chacune de ces classifications est imparfaite et souffre de faiblesses. Il leur reproche principalement que certaines catégories se superposent, que certaines typologies incluent des expressions figées et des expressions idiomatiques, que d'autres soient trop complexes et divisées, voire que des erreurs soient commises dans l'identification des patrons syntaxiques et lexicaux.

Ces erreurs possibles d'identification sont le fruit de confusions dues à certaines particularités de la grammaire arabe. En effet, certains patrons de collocations sont uniques à cette langue. En voici quelques-uns :

- Patron verbe + verbe, dont le premier verbe est un verbe factitif, inchoatif ou une « sœur » de *kāna* (auxiliaires). Le premier verbe est à l'accompli (الماضي) tandis que le second est à l'inaccompli (المضارع). Par exemple : جعله يقول (*ǧaʿalahu yaqūl*, « lui faire dire ») ou أخذ يبحث (*ʾaḫaḏa yabḥaṯ*, « se mettre à chercher »).

Les trois patrons verbaux suivants sont considérés comme des patrons verbe + adverbe car le collocatif, bien que sous une forme nominale ou adjectivale à l'accusatif, fonctionne comme un adverbe.

- Patron verbe + *mafʿūl muṭlaq* (مفعول مطلق, « complément absolu »). Ce dernier est utilisé en arabe comme une façon élégante de mettre l'emphase sur un verbe en faisant dériver un *maṣdar* (nom verbal) à l'accusatif (généralement indéfini) depuis le verbe principal ou la prédication principale, et peut inclure un adjectif (Ryding, 2005). Par exemple : فرح فرحًا شديدًا (*fariḥa faraḥan šadīdan*, « il s'est réjoui d'une forte joie *(litt.)* ».

- Patron verbe + *tamyīz* (تمييز, « complément spécificatif »). Ce dernier est utilisé en arabe pour spécifier la nature du référent. Il est encore une fois généralement sous une forme nominale à l'accusatif, mais fonctionne comme un adverbe complément de manière. Par exemple : نظر إليه شزرًا (*naẓara ʾilayhi šazran*, « regarder de travers »).
- Patron verbe + *ḥāl* (حال, « complément de manière »). Ce dernier est utilisé en arabe pour décrire les circonstances d'une action. Il est dans la plupart des cas construit à partir d'un participe actif (اسم الفاعل) à l'accusatif s'accordant en genre et en nombre avec l'agent. Par exemple : ولّى هاربًا (*wallā hāriban*, « tourner en fuyant *(litt.)* »).

D'autres patrons spécifiques à l'arabe (adjectif + complément absolu, substantif + adverbe de lieu, etc.) sont possibles (Izwaini, 2015), mais cela dépasse le cadre de ce projet. Cependant, il apparaît clairement que la langue arabe, du fait de ses particularités non seulement morphosyntaxiques mais également du fait de ses spécificités culturelles, présente un certain nombre de difficultés pour l'annotation de ses collocations.

3.2.4. Points communs et autres difficultés

Bien que les trois langues étudiées appartiennent à des familles de langues différentes (romanes, germaniques et sémitiques), les collocations dans les textes ont dans tous les cas un comportement dynamique et apparaissent rarement telles que décrites dans les dictionnaires (Tutin, 2004). Ainsi, les composants des collocations subissent un certain nombre de variations en contexte.

Ces variations peuvent être d'ordre lexical, auquel cas on pourra constater des variantes synonymiques sans changement de sens *(périr / mourir d'ennui)* ou l'apparition de diverses prépositions. La morphologie des composants des collocations subit énormément de variations également, qu'il s'agisse des verbes qui s'accordent en genre et en nombre avec leur sujet et réagissent au temps / aspect / mode requis, ou des substantifs qui peuvent être au singulier ou au pluriel dans les trois langues, au masculin ou au féminin en français et en arabe, et au duel dans le cas de cette dernière, sans mentionner les pluriels brisés qui sont omniprésents.

La plupart des collocations peuvent être étendues à cause de l'insertion d'adverbes et / ou d'adjectifs (Tutin, 2004) et peuvent ainsi apparaître de manière discontinue. C'est notamment vrai pour le français et l'anglais qui pourront avoir l'adjectif entre le verbe et l'objet (*to commit a horrendous crime*, *commettre un horrible crime*), mais pas pour l'arabe, pour qui l'adjectif est rejeté après (à gauche) du substantif qu'il qualifie (اقترف جريمةً مروعةً, *iqtarafa ǧarīmatan murawwiʿatan*[11], « commettre un crime horrible (litt.) »). Les variations peuvent finalement être d'ordre distributionnel (notamment dans le cas des collocations adjectif-nom) ou d'ordre syntaxique. En effet, les collocations n'apparaissent pas toujours à la diathèse active dans des phrases simples, mais peuvent figurer dans des constructions à la diathèse passive (*des mesures ont été prises*) ou dans des relatives (*les mesures qu'il a prises*).

[11] L'arabe moderne standard est une langue à désinences casuelles. Selon le cas grammatical d'un mot, ce dernier voit sa vocalisation finale modifiée : ـُ *-u* pour le cas sujet, ـَ *-a* pour le cas direct et ـِ *-i* pour le cas indirect. Cette vocalisation, lorsque le mot en question est indéfini, se manifeste par un تنوين (*tanwīn*, « nounation », soit la suffixation d'un *-n* non graphique) : ـٌ *-un* pour le cas sujet, ـًا *-an* pour le cas direct et ـٍ *-in* pour le cas indirect. Nous avons choisi de conserver ces désinences casuelles à la fois dans les segments en arabe mais également dans notre translittération.

Par ailleurs, il faut noter qu'une collocation verbo-nominale dans une langue n'aura pas forcément d'équivalent direct sous la même forme, c'est-à-dire que l'équivalent ne sera régulièrement pas une collocation. Brashi (2005) nous donne un exemple dans lequel l'équivalence arabe-anglais se fait de manière transparente, auquel on peut également ajouter la traduction française : *gagner la confiance*, *to win confidence*, كسب ثقةً (*kasaba ṯiqatan*, « gagner confiance (litt.) »).

Puis, il prouve qu'une telle équivalence n'est pas toujours le cas et que la dimension arbitraire de la combinaison de la collocation joue un rôle important : *donner une leçon*, *to teach a lesson*, لقّن درسًا (*laqqana darsan*, « dicter leçon (litt.) »). Si l'on devait donner des verbes français pour chacune des trois collocations ci-dessus, on aurait *donner*, *enseigner* et *dicter* respectivement.

Dans le même esprit, ces non-équivalences de traduction peuvent aussi se refléter dans des constructions grammaticales différentes. C'est dans ces cas-là que les spécificités de chaque langue sont les plus saillantes. L'auteur précédemment cité propose la paire anglais-arabe suivante, à laquelle nous ajoutons l'équivalent français : *se suicider*, *to commit suicide*, انتحر *(intaḥara,* « se suicider »*)*. Quand le français a recours à un verbe pronominal intransitif, l'anglais utilise plutôt une construction verbe-objet direct, et l'arabe, quant à lui, emploie un verbe de forme augmentée (en l'occurrence, forme VIII). Ces formes augmentées peuvent régulièrement synthétiser des expressions complexes en un seul mot du fait de leur sémantisme complexe, comme nous l'avons déjà présenté (voir section 3.2.3.2). Par exemple, la racine نحر (*naḥara*) peut se manifester sous trois formes. Nous les illustrons dans le tableau ci-après :

Forme verbale	**Valeur sémantique**	**Orthographe**		**Translittération**		**Traductions**
		accompli	**inaccompli**	**accompli**	**inaccompli**	
I	Forme simple	نَحَرَ	يَنْحَرُ	*naḥara*	*yanḥaru*	immoler, sacrifier rituellement
VI	Réciprocité	تَنَاحَرَ	يَتَنَاحَرَ	*tanāḥara*	*yatanāḥaru*	s'entretuer
VIII	Réflexivité	اِنْتَحَرَ	يَنْتَحِرُ	*intaḥara*	*yantaḥiru*	se suicider

Tableau 2 : Formes verbales pour la racine نحر (naḥara)

En passant de la forme simple نَحَرَ (*naḥara*, « immoler, sacrifier rituellement ») à la forme VI تَنَاحَرَ (*tanāḥara*, « s'entretuer ») par l'ajout d'un *tāʾ* (ـتـ) à l'initiale et d'un *alif* (ـا) entre la 1ère et 2e radicale, le sens devient alors réciproque. Si, à la place, c'est un *alif* (ا) qui est placé à l'initiale et un *tāʾ* (ـتـ) infixé entre la 1ère et la 2e radicale, on obtient la forme réflexive اِنْتَحَرَ (*intaḥara*, « se suicider »).

3.2.5. Projection des annotations

Un dernier point à prendre en considération dans une perspective d'annotation multilingue est la possibilité de transférer des annotations faites sur un premier corpus vers un ou plusieurs autres corpus. L'idée globale derrière la notion de projection d'annotation est de pouvoir obtenir automatiquement des annotations de qualité pour des corpus d'une langue-cible peu ou moyennement dotée grâce aux annotations d'un corpus d'une langue plus richement dotée.

Pour résoudre ce problème, Akbik et Vollgraf (2018) ont mis au point ZAP, un outil développé en Java permettant de faire le transfert des annotations depuis un texte anglais vers

un texte d'une autre langue sur plusieurs niveaux linguistiques : les catégories grammaticales, les dépendances syntaxiques, ainsi que les cadres et rôles sémantiques. À ce jour, la projection ne se fait que depuis l'anglais et les langues-cibles prises en charge sont le français, l'allemand, l'espagnol et dans une moindre mesure le chinois.

L'outil prend en entrée deux phrases, qui peuvent être fournies sous forme de texte (qui seront ensuite analysées syntaxiquement) ou au format CoNLL (ce qui fait gagner du temps sur le traitement) : une phrase source en anglais, et une phrase cible dans une des langues supportées. Ces deux phrases sont envoyées à un aligneur heuristique qui va trouver, dans une ressource de type table de traduction lexicale, l'équivalent le plus probable pour les tokens de la phrase source dans la phrase cible afin de retourner les alignements les plus précis possibles. Enfin, l'outil transfère les annotations correspondant aux alignements retrouvés sur la phrase cible, générant ainsi un fichier CoNLL enrichi. Qui plus est, l'outil est fourni avec une interface graphique appelée `TheProjector` qui permet de visualiser de quelle manière la projection a été effectuée, de sorte à pouvoir en évaluer la qualité. Dans notre objectif d'obtenir un corpus parallèle trilingue annoté en collocations, nous utiliserons ZAP. Nous en décrivons le fonctionnement plus en détail dans la section 10.2.

Un projet d'annotation des collocations dans un environnement trilingue, on l'a vu, pose un certain nombre de difficultés. Certaines sont liées aux originalités de chacune des langues étudiée : les constructions à base de verbes pronominaux en français peuvent être un obstacle ; les constructions avec des verbes à particule en anglais se présentent également comme une complication à dépasser ; et l'arabe, de par son sens de lecture et son script différents, ses particularités morphosyntaxiques et son système hautement flexionnel, ainsi que ses constructions grammaticales uniques, transforme le tout en un réel défi. Malgré toutes les spécificités inhérentes à chacune de ces trois langues et les disparités d'équivalences d'une langue à une autre, des universaux existent (Savary et al., 2017). En prenant en considération à la fois ces similarités et ces singularités, appliquer des techniques d'identification et d'extraction automatique est tout à fait possible.

4. DECOUVERTE ET IDENTIFICATION DES COLLOCATIONS

4.1. Définitions et portée

D'une manière générale, les processus pour repérer et annoter les collocations sont les mêmes qui sont appliqués pour les expressions polylexicales, les unes étant une sous-classe des autres. Nous emploierons donc ce dernier terme régulièrement en y incluant le phénomène qui nous intéresse pour définir les concepts relatifs à leur repérage.

Constant et al. (2017) ont défini un cadre conceptuel dans lequel ils distinguent deux sous-tâches dans le traitement automatique des expressions polylexicales, qu'il faut fondamentalement séparer. D'une part, il y a la découverte *(discovery)* des expressions, c'est-à-dire l'identification de nouvelles expressions dans des corpus, dans le but de les stocker dans des espaces de type lexique, afin de les réutiliser ultérieurement. D'autre part, il y a l'identification des expressions, qui consiste à annoter les expressions polylexicales détectées avec les différents types connus. Les résultats de ces deux tâches sont bien distincts : il y a d'un côté une liste de polylexèmes, et de l'autre une liste d'annotations.

Cette volonté de créer un cadre conceptuel est née pour pallier le manque de clarté quant aux différentes appellations utilisées dans la littérature pour définir ces tâches. Pour Baldwin et Kim (2010), les deux sous-tâches incluses dans le traitement automatique des expressions polylexicales sont d'une part l'identification et l'extraction des expressions depuis des corpus, tout en désambiguïsant leur syntaxe interne, et d'autre part l'interprétation de ces expressions. Ce que les premiers ont désigné par *découverte* est donc défini comme *identification* et *extraction* pour les seconds. Cependant, ces derniers font la distinction plus loin entre *identification* et *extraction*, alors qu'ils les incluaient dans la même sous-tâche précédemment.

Toujours est-il que ces deux tâches sont utilisées dans des *pipelines*, c'est-à-dire dans des chaînes de traitement, généralement avec des analyseurs syntaxiques et des applications de TA. Avant de nous pencher sur les différentes approches et les outils développés à ces fins, nous définirons dans un premier temps les deux tâches susmentionnées et nous discuterons des mesures d'association, mesures statistiques communément utilisées dans l'estimation du degré de dépendance des lexèmes.

4.1.1. Découverte des expressions polylexicales

Au sens de Constant et al. (2017), nous incluons ceux que d'autres ont appelé identification ou acquisition et extraction dans le processus de découverte. Tout d'abord, il s'agit de déterminer et de lister les occurrences de ce qui pourrait être des expressions polylexicales dans des corpus, on travaille donc au niveau du token. Un des défis principaux posés par la découverte d'expressions polylexicales est de déterminer si la combinaison de mots est faite dans un sens littéral ou pas. Baldwin et Kim (2010) donnent l'exemple de *make a face* qui peut avoir la double lecture selon le contexte : dans *Kim made a face at the policeman* (« Kim a fait une grimace au policier ») et *Kim made a face in pottery class* (« Kim a fait un visage en cours de poterie »), l'expression polylexicale a respectivement un sens idiomatique, puis un sens littéral. Le traitement automatique s'avère donc problématique pour distinguer les deux occurrences.

Les difficultés ne s'arrêtent pas là. Constant et al. (2017) en relèvent deux autres, que nous avons déjà évoquées dans la section précédente : les éventuelles discontinuité et variabilité des expressions polylexicales. En effet, nous l'avons vu, les constructions verbales, notamment, présentent très souvent ces deux caractéristiques. La discontinuité est généralement rendue moins problématique par les étapes de prétraitement comme l'analyse syntaxique, malgré l'introduction d'ambiguïtés dues à cette dernière. La variabilité a lieu régulièrement, notamment en ce qui concerne les langues morphologiquement riches comme l'arabe. Pour remédier à ce problème, des étapes de prétraitement supplémentaires, telles que l'étiquetage des parties du discours ou la lemmatisation pour se débarrasser de toutes les flexions qui ralentissent le processus, peuvent s'avérer nécessaires. Cependant, ces outils, bien que très performants et très utiles, demeurent imparfaits et les résultats obtenus nécessitent malgré tout une intervention humaine *ad hoc* pour vérification et correction. Savary et al. (2017) confirment toutes ces difficultés et ajoutent que les expressions polylexicales verbales se comportent différemment et doivent être modélisées en conséquence d'une langue à l'autre.

À la suite de ce repérage, c'est ce que de nombreux chercheurs comme Baldwin et Kim (2010) ont qualifié *d'extraction* qui a lieu. Selon eux, les deux tâches se distinguent par le fait que la première va générer une liste de candidats potentiels, tandis que la seconde va trancher et ne retenir que ceux qui ont bien les caractéristiques des expressions polylexicales. Cette sous-tâche est problématique pour les mêmes raisons que la précédente et est pertinente pour toutes les applications relatives au lexique, qu'il s'agisse de construction de dictionnaires, d'extraction d'information ou de construction de grammaires. Les différentes techniques et approches d'extraction des collocations sont abordées et décrites *infra*.

4.1.2. Identification des expressions polylexicales

Toutes les étapes et les sous-tâches exécutées pendant la découverte d'expressions polylexicales servent à étendre les ressources de type lexique qui sont ensuite utilisées dans le processus d'identification (Constant et al., 2017). Outre ces éventuels lexiques, l'ajout de règles ou de modèles d'apprentissage pour détecter les occurrences d'expressions polylexicales peut s'avérer nécessaire. L'identification automatique étant principalement intéressée par l'ajout d'annotations aux expressions polylexicales découvertes, l'usage d'un étiqueteur d'expressions polylexicales *(MWE tagger)* est de rigueur. On peut attendre de ce dernier qu'il mette en valeur les expressions polylexicales détectées par une annotation quelconque, qu'il arrive à distinguer les différentes expressions apparaissant dans une même phrase, voire qu'il identifie la classe des expressions détectées (dans le cadre d'une étude ne portant pas sur un type exclusif d'expressions polylexicales).

Cette identification est intéressante à la fois pour le développement et l'amélioration des performances des analyseurs syntaxiques et des systèmes de TA (Constant et al., 2017). Les premiers bénéficieraient d'une plus grande clarté avec la mise à l'écart d'un certain nombre d'ambiguïtés, tandis que les seconds pourraient générer de meilleures traductions, étant donné que de très nombreuses expressions polylexicales ne peuvent pas être traduites mot-à-mot. Ce ne sont pas les seules applications qui pourraient tirer avantage d'une identification efficace des expressions polylexicales. En effet, les systèmes de traitement et de désambiguïsation sémantiques pourraient gagner en précision dans leur étiquetage.

Les défis imposés à l'identification des expressions polylexicales sont similaires à ceux de la découverte : toutes deux rencontrent des difficultés à traiter ces dernières par leur variabilité syntaxique, leur ambiguïté sémantique et leur éventuelle discontinuité. À ces difficultés s'ajoute la possibilité d'avoir des expressions polylexicales qui se chevauchent ou qui sont imbriquées, comme dans une collocation verbale dont les deux objets d'un même verbe seraient coordonnés : *elle dirige le débat et les opérations.*

4.1.3. Mesures d'association

Dans un projet d'annotation des collocations, la notion de cooccurrence est d'une importance capitale. Seulement, il serait erroné de penser que tout mot apparaissant dans le contexte immédiat d'un autre assez fréquemment constituerait automatiquement une collocation. Certaines cooccurrences ne sont que le fait d'une association compositionnelle régulière *(manger une salade)* et d'autres peuvent être le fruit d'une juxtaposition de mots-outils n'ayant aucun rapport formel (وفي, *wa fī*, « et dans »).

La cooccurrence de certaines paires de mots peut n'être qu'une coïncidence. Ainsi, une mesure statistique est nécessaire pour déterminer le degré d'association réel entre les mots concernés (Evert, 2005). Plusieurs méthodes statistiques ont été développées au fil du temps, mais la plupart d'entre elles prennent en compte le nombre de cooccurrences observées d'un groupe de *n* mots (n_1, n_2, … n_n) et compare ce nombre avec celui qui était attendu (Constant et al., 2017).

Ces mesures d'association *(association measures)* calculent donc un score d'association entre deux mots pour déterminer s'il s'agit d'une simple cooccurrence ou d'une collocation (sur le plan statistique). Evert (2005) dresse un inventaire exhaustif des différentes mesures existantes et les classe en 4 groupes : *significance of association, degree of association, information theory* et *heuristics*. Les mesures d'association les plus fréquentes et les plus utilisées sont présentées par (Somers, 2001) :

- *Mutual information (MI)* : Introduit par Church et Hanks (1990), la mesure MI calcule la vraisemblance maximum pour la force logarithmique d'une association statistique pour une paire de mots donnée. Plus la valeur est haute, plus la vraisemblance d'avoir affaire à une collocation est élevée. À cause de cela, la force d'association est régulièrement surestimée pour les cooccurrences à fréquence basse. On considère f_a comme étant le nombre d'occurrences de la base dans le corpus, f_b le nombre d'occurrences du collocatif dans le corpus, f_{ab} le nombre de cooccurrences et n le nombre total de tokens.

$$\mathbf{MI} = \log_2 \frac{f_{ab} * n}{f_a * f_b}$$

- *Enhanced Mutual Information (EMI)* : Proposée par Zhang et al. (2009) pour pallier les problèmes de cooccurrence asymétrique du MI, cette mesure d'association est définie comme le rapport entre la probabilité d'occurrence d'une paire de mots et le produit des probabilités d'occurrence des mots individuels, à l'exclusion des occurrences de la paire de mots.

$$\mathbf{EMI} = \log_2 \frac{f_{ab}}{(f_a - f_{ab})(f_b - f_{ab})}$$

- *t*-score : Mis au point par Church et al. (1991) pour pallier les lacunes de MI, il est souvent combiné avec ce dernier pour s'assurer que l'association détectée est corroborée par assez de preuves.

$$\mathbf{t-score} = \frac{\left(\frac{f_{ab} - (f_a * f_b)}{n}\right)}{\sqrt{f_{ab}}}$$

- *chi-squared* : Prenant en compte la table de contingence de la paire de mots étudiée, le *chi-squared* (χ^2) offre une mesure plus précise de l'importance de l'association que le t-score. On considère *i* comme étant la base, *j* le collocatif, n_{ij} la fréquence de cooccurrence observée et e_{ij} la fréquence de cooccurrence attendue.

$$\mathbf{chi-squared} = \frac{\sum\sum n_{ij} - e_{ij}}{e_{ij}}$$

- *Log-likelihood ratio (LLR)* : Développé par Dunning (1993) en réponse aux performances relativement décevantes dans l'extraction de collocations avec le *chi-squared*, le LLR est devenu un standard en linguistique computationnelle.

$$\mathbf{LLR} = 2 * \sum_{j=1}^{2}\sum_{i=1}^{2} n_{ij} * \log\frac{n_{ij} * n_{**}}{n_i * n_j}$$

- *Dice* : Le coefficient Dice est une mesure utilisée en recherche d'informations et a été utilisé allègrement pour l'extraction de collocations. Il recherche également la vraisemblance maximum de la force d'une association mais ne fonctionne pas comme le MI pour autant.

$$\mathbf{DICE} = \frac{2 * f_{ab}}{(f_a + f_b)}$$

Malheureusement, ces mesures ont des limites (Constant et al., 2017). En effet, ces mesures d'association fonctionnent bien pour des paires de mots, mais beaucoup moins bien pour des associations de *n*-mots, et jusqu'à présent, aucune n'a pu être identifiée comme la meilleure solution.

Heureusement, ces mesures d'association peuvent être adaptées selon l'étude menée. Dans la section suivante, nous nous attacherons à décrire différentes approches d'extraction de collocations et nous montrerons que les différentes mesures d'association sont très largement combinées pour obtenir les meilleurs résultats possibles.

4.2. Extraction des collocations

Comme le précisent Todirascu et al. (2008) ou Zaidi et al. (2010), les outils d'extraction de collocations qui ont été développés sont de trois types : une première approche est purement statistique et s'appuie sur les résultats de fréquence de cooccurrences ; une deuxième est syntaxique et s'appuie sur des règles symboliques ; enfin, une troisième adopte une approche hybride en mixant les deux premières approches afin de pallier leurs faiblesses respectives, en y mêlant parfois d'autres techniques.

4.2.1. Approche statistique

D'une manière générale, l'extraction de collocations se fait en deux étapes : tout d'abord, un repérage des candidats est effectué (la plupart du temps après des étapes de prétraitement comme l'application d'un lemmatiseur et / ou d'un étiqueteur morphosyntaxique) ; dans un deuxième temps, ces candidats sont filtrés et classés à l'aide d'une ou plusieurs mesures d'association pour calculer leur vraisemblance (Ramisch, 2012). Historiquement, ce sont les méthodes principalement statistiques qui ont été utilisées en premier lieu.

Parmi les premiers outils développés sur des bases statistiques, on peut citer *Xtract*, développé par (Smadja, 1993) qui, grâce à la combinaison de *n*-grammes et de la mesure d'association MI, permettait d'extraire des collocations avec une précision supérieure à 80% (cité par Ramisch, Villavicencio et Boitet 2010).

Un peu plus tard, avec le développement de leur algorithme *Champollion* et l'utilisation de *Xtract*, les travaux de McKeown et al. (1996) s'inscrivirent comme pionniers pour ce qui est de l'extraction bilingue de collocations. En testant *Champollion* sur les données du corpus parallèle Hansard, ils ont pu atteindre une précision moyenne de 73% dans leur identification automatique d'équivalents de traduction de collocations en anglais et en français.

L'inconvénient principal de l'approche statistique est qu'elle identifie un trop grand nombre de candidats qui ne sont pas de réelles collocations. Afin de remédier à cela, même si les travaux sus-cités appliquaient déjà quelques informations syntaxiques de surface pour leur extraction, on s'est rapidement rendu compte qu'il était nécessaire de se focaliser un peu plus sur des systèmes symboliques à base de règles syntaxiques plus poussées.

4.2.2. Approche syntaxique

Cette approche symbolique a été adoptée notamment par les développeurs de l'outil *Fips* (Wehrli, 2007), un analyseur syntaxique profond basé sur les travaux de Chomsky sur la grammaire générative. Cette technique d'extraction basée sur les dépendances syntaxiques offre l'avantage de ne pas nécessiter que les composants de la collocation soient immédiatement adjacents, mais requière plutôt qu'ils soient liés syntaxiquement. Seretan et al. (2004) donnent un exemple avec une collocation verbale détectée alors que 39 mots séparent le sujet de son verbe. Ils avancent également un argument de poids en montrant qu'une approche basée entièrement sur la syntaxe en dépendances résout les difficultés liées à la diathèse passive, aux topicalisations, aux dislocations, ou encore aux structures clivées (Seretan, 2008).

Cette approche syntaxique a été à de nombreuses fois adoptée dans le cadre des collocations et autres expressions polylexicales arabes. En ce qui concerne les collocations, Zaidi et al. (2010) ont adapté GATE, un outil initialement développé pour l'extraction d'entités nommées, en créant de nouvelles règles JAPE *(Java Annotation Pattern Engine)*. Ils ont appliqué cet outil au *Crescent Quranic Corpus*, un corpus du texte coranique entièrement étiqueté en parties du discours et contenant des annotations morphologiques additionnelles. Attia (2006) a quant à lui présenté sa méthode semi-automatique à base d'expressions régulières pour extraire certains types d'expressions polylexicales.

Certaines méthodes adoptent des stratégies originales. Lü et Zhou (2004), dans le but d'extraire des collocations bilingues, utilisent des triplets syntaxiques depuis des corpus

monolingues. Ces triplets sont d'abord extraits de corpus chinois et anglais avec un analyseur syntaxique, puis un modèle de traduction du triplet de dépendances est estimé à l'aide de l'algorithme EM, basé sur une hypothèse de correspondance de dépendances. Ce triplet de traductions est ensuite appliqué pour extraire des équivalents de traduction depuis des corpus monolingues.

4.2.3. Approche hybride : apprentissage automatique, plongements lexicaux, réseaux de neurones

On peut se rendre compte que, finalement, les deux approches s'inspirent plus ou moins des stratégies de l'autre, au moins en surface. *A posteriori*, il serait facile de dire qu'une approche mêlant à la fois les statistiques et la syntaxe en dépendances semble le choix optimal pour obtenir des résultats concluants. C'est exactement ce vers quoi tendent la plupart des techniques d'extraction actuelles.

Ces systèmes hybrides diffèrent dans leur *orchestration* (Constant et al., 2017), c'est-à-dire à quel moment telle ou telle stratégie est utilisée dans la série de traitements imposés par l'outil. Par exemple, (Seretan et al., 2004) applique une méthode syntaxique pour filtrer les candidats, puis une méthode statistique pour classer les candidats retenus. C'est cet ordre-là que la plupart des approches hybrides intègrent.

Plusieurs travaux avec une approche similaire ont été réalisés pour l'arabe. Pour l'extraction de termes polylexicaux en langue spécialisée, Boulaknadel et al. (2008) ont appliqué des filtres linguistiques pour la détection de candidats termes, incluant différents types de variations (graphiques et orthographiques, flexionnelles, morphosyntaxiques), avant de comparer 4 mesures d'association (LLR, FLR, MI et *t-score*) pour classer les termes polylexicaux. Saif et Aziz (2011) ont appliqué une méthode quasi-identique pour l'extraction de collocations, mais ont préféré les mesures d'association EMI et $\chi 2$ à FLR et *t-score*.

Un des outils les plus complets car offrant un panel exhaustif d'applications pour à la fois la découverte et l'identification des expressions polylexicales est le `mwetoolkit` (Ramisch, 2012). Originalement conçu pour l'extraction de termes polylexicaux dans des corpus spécialisés, le `mwetoolkit` a ensuite été étendu pour le traitement automatique complet des expressions polylexicales dans les corpus spécialisés mais également généraux. Sa méthodologie a été implémentée en un ensemble de modules indépendants développés en Python, et chacun de ses modules joue un rôle dans la chaîne de traitement de l'extraction. Dans la pratique, les deux phases standards sont utilisées : repérage de candidats (basé sur des *n*-grammes et / ou des patrons morphosyntaxiques spécifiques, qu'il s'agisse des formes de surface, des lemmes, des étiquettes des parties du discours ou des dépendances syntaxiques) suivi du filtrage des candidats (avec combinaison de mesures d'association, des caractéristiques descriptives et contrastives et de l'apprentissage automatique). Là où le `mwetoolkit` est une *boîte à outils* à part entière, c'est qu'il offre un dispositif d'indexation de corpus avec l'intégration de moteur de recherche pour utiliser le web comme un corpus, des dispositifs de validation et d'annotation pour la phase d'évaluation, et il permet également l'intégration d'un outil d'apprentissage automatique pour la création de modèles supervisés d'extraction d'expressions polylexicales grâce à des données annotées.

Outre les modèles statistiques et syntaxiques standards, ainsi que les modèles d'apprentissage automatique, les changements paradigmatiques récents en TAL ont permis d'élargir le champ des possibles pour l'extraction de collocations et plus généralement d'expressions polylexicales. Un travail intéressant a été réalisé par Garcia et al. (2017) en utilisant des plongements lexicaux pour extraire des collocations trilingues depuis des corpus parallèles. Après les étapes de prétraitement (leurs corpus sont notamment entièrement lemmatisés), de découverte des candidats et de classement avec des mesures d'association (*t-score* et MI), les auteurs utilisent le modèle de plongements lexicaux *BiSkip* (Luong et al., 2015), lui-même basé sur le modèle *Skip-gram* de *word2vec* (Mikolov et al., 2013), pour l'apprentissage des représentations vectorielles bilingues depuis des corpus parallèles. Travaillant avec trois langues, après la construction de leurs trois modèles bilingues (espagnol-anglais, espagnol-portugais, portugais-anglais), ces derniers sont capables de calculer la distance cosinus entre leurs vecteurs, ce qui signifie, en ce qui concerne les collocations, qu'ils seraient à même d'identifier les équivalents bilingues d'une base mais également de stocker la relation sémantique la moins proche entre les collocatifs bilingues. Un alignement des équivalents bilingues est ensuite nécessaire. Pour ce faire, pour une collocation dans la langue source, la base est sélectionnée et les *n* lemmes les plus similaires dans la langue cible sont obtenus grâce au modèle bilingue. Ensuite, dans la liste cible, les collocations contenant les équivalents bilingues de la base sont recherchées et si une collocation répondant à ces critères est trouvée, la distance cosinus entre les deux collocations est calculée. Si leur similarité est supérieure à un seuil défini empiriquement, les collocations source et cible sont alignées et un score de fiabilité (la distance moyenne entre les bases et les collocatifs) leur est attribué.

Encore plus récemment, dans le cadre du PARSEME Shared Task 2018, Zampieri et al. (2018) ont soumis pour évaluation leur outil d'identification des expressions polylexicales verbales *Veyn*, basé sur un étiqueteur de séquences utilisant des réseaux de neurones récurrents (RNN). Cet outil est disponible gratuitement et couvre 19 des 20 langues du projet PARSEME (toutes sauf l'arabe). Pour représenter les expressions polylexicales, une variante du schéma d'encodage standard *begin-inside-outside* (BIO) proposé par Ramshaw et Marcus (1995) est utilisée : lorsqu'une expression polylexicale est identifiée, le token qui débute l'expression est annoté *B*, les tokens contenus dans l'expression sont annotés *I* et les tokens qui n'en font pas partie sont annotés *O*. Une étiquette spéciale *g* est utilisée pour annoter les tokens qui séparent les tokens d'une expression discontinue. Ce système d'annotation a été étendu davantage pour faire figurer la catégorie d'expression polylexicale en plus de l'encodage BIO. Basé sur des RNN, *Veyn* prend en entrée le lemme et le UPOS (la balise Universal Dependencies pour les parties du discours) de chaque token, représentés d'abord en vecteurs *one-hot* avant d'être transformés en vecteurs denses à 250 dimensions. Pour améliorer les performances, deux couches récurrentes sont utilisées : chacune présente deux réseaux récurrents à portes (*Gated Recurrent Units*, GRU) identiques, contient 512 neurones d'entrée et est bidirectionnelle. La première couche récurrente prend en entrée la concaténation des vecteurs précédemment obtenus, puis la seconde couche récurrente prend en entrée la concaténation des vecteurs *forward* et *backward* de la précédente concaténation. Ces vecteurs sont finalement utilisés par la couche d'activation pour donner les résultats du processus, qui sont ensuite transformés en

probabilités pour la vraisemblance de chaque étiquette, et la plus vraisemblable est finalement sélectionnée.

Toujours dans le cadre du PARSEME Shared Task 2018, Pasquer et al. (2018) ont développé VarIDE *(Variant IDEntification)* en formulant l'hypothèse que l'identification des expressions polylexicales verbales obtiendrait de meilleurs résultats grâce à l'apprentissage des différentes variantes morphologiques et / ou syntaxiques d'une expression polylexicale candidat. Ainsi, l'outil génère en premier lieu un nombre important de candidats basés sur les patrons syntaxiques les plus fréquents des expressions polylexicales verbales annotées dans un corpus d'entraînement, sous forme de tuples contenant les étiquettes grammaticales des composants de l'expression polylexicale. Des candidats sont ensuite générés avec les lemmes des composants de l'expression polylexicale. Ces candidats sont ensuite eux-mêmes déclinés avec les différentes formes flexionnelles des lemmes constituant l'expression polylexicale, puis filtrés pour ne conserver que les candidats respectant les patrons syntaxiques autorisés. En deuxième lieu, les caractéristiques morphosyntaxiques des candidats sont extraites pour finir la phase d'entraînement. Ces informations (morphologiques et dépendances syntaxiques) sont utilisées pour calculer la similarité entre un candidat et le reste des expressions polylexicales annotées. Enfin, la phase de prédiction comporte une étape d'extraction des candidats du corpus de test, à la manière de celle réalisée sur le corpus d'entraînement, avant qu'ils ne soient comparés aux tuples normalisés extraits du corpus d'entraînement. Pour finir, un classifieur bayésien naïve classe et annote finalement les expressions polylexicales. C'est cet outil que nous utiliserons pour l'annotation automatique du corpus français de notre projet. Nous en décrivons le fonctionnement plus en détail dans la section 7.3.

En 30 ans, l'extraction d'expressions polylexicales et plus spécifiquement des collocations a pris de multiples formes, en adoptant des approches qui ont évolué des mesures statistiques, aux règles symboliques, pour finalement s'affiner au rythme des découvertes et des changements de paradigme du TAL, avec l'apprentissage automatique, les plongements lexicaux et les RNN. Pour conclure cette partie, nous nous intéresserons aux problématiques intrinsèquement liées à la langue arabe et son traitement automatique.

4.3. L'arabe : vers des difficultés supplémentaires

Dans le trio de langues qui nous intéresse, le français et surtout l'arabe sont deux langues dotées d'une morphologie très riche quand on les compare à l'anglais (Green et al., 2012). Il faut noter qu'une approche statistique seule n'est pas suffisante pour la plupart des langues, et cela est d'autant plus vrai pour l'arabe. Notamment de par son caractère agglutinant, son traitement automatique doit être supplanté par des règles linguistiques (Boulaknadel et al., 2008).

Nous avons déjà parlé de l'ordre syntaxique VSO de la phrase verbale et l'absence de copule dans les phrases nominales. Nous avons également dit qu'au niveau morphologique, l'arabe est une langue agglutinante et les adjectifs possessifs et les pronoms objets se suffixent au mot précédent, nominal ou verbal respectivement. Outre cela, quelques mots grammaticaux sont monolitères et ne peuvent pas se trouver seuls sur la ligne. De fait, ils viennent se préfixer (et se ligaturer graphiquement dans la plupart des cas) au mot suivant. Le cumul de ces deux caractéristiques peut tout à fait donner une séquence comme ولأخواتكم *(wa li ʾaẖawātikum,* « et

pour vos sœurs »), qui ne forme qu'un seul mot graphique (aucune espace typographique) mais qui est pourtant composée de la conjonction و *(wa,* « et »*)*, de la préposition لـ *(li,* « pour »*)*, du substantif pluriel brisé (interne) أخوات *(ʾaḫawāt,* « sœurs »*)* et de l'adjectif possessif كم *(kum,* « vos »*)*. Par ailleurs, l'arabe utilise certains recours morphotactiques comme l'adjonction de la préposition لـ *(li)* et de l'article défini الـ *(al-)* qui donnera la forme للـ *(li-l),* qui elle-même sera préfixée au substantif ou à l'adjectif qu'elle définit. Ces amalgames sont problématiques pour les tokeniseurs et les erreurs de démarcation sont communes, ce qui entraîne inévitablement une réaction en chaîne d'erreurs (mauvais étiquetage morphosyntaxique, lemmatisation impossible, etc.) (Habash, 2010).

De plus, l'arabe est une langue sémitique qui fonctionne donc à base de racines consonantiques (trilitères ou quadrilitères) et de schèmes. La formation des verbes et des substantifs se fait par le choix d'une racine qui porte un sémantisme (unique ou pluriel), à laquelle on « plaque » des schèmes prenant la forme d'affixes et de diacritiques (Green & Manning, 2010). Ces diacritiques font office de vocalisation (ou de gémination) et sont parfois des marqueurs grammaticaux, notamment pour marquer le cas. Cependant, ce système graphique est hautement ambigu à plusieurs égards. D'une part, il faut noter que les vocalisations ne sont pas la norme des écrits en arabe : elles sont généralement présentes dans des cadres spécifiques (pédagogiques, coraniques, poétiques, ou dans de rares cas de désambiguïsation). Si l'on prend le mot كتب dont la racine est composée des lettres *kāf* ك, *tāʾ* ت et *bāʾ* ب, sans diacritiques et hors contexte, il peut se lire de plusieurs manières : كَتَبَ (*kataba,* « il a écrit »), كُتِبَ (*kutiba,* « il a été écrit / on a écrit »), كُتُب (*kutub,* « des livres »), كَتْب (*katb,* « le fait d'écrire », ou encore كَتَّبَ (*kattaba,* « faire écrire ». Ces ambiguïtés sont omniprésentes, rendant la lecture difficile pour les apprenants, et dans un cadre de traitement informatique, elles rendent le travail des étiqueteurs morphosyntaxiques complexe. Une étape de remise de diacritiques est souvent requise dans le prétraitement d'un corpus en arabe.

La liste des particularités ne s'arrête pas là (Habash, 2010). Nous l'avons mentionné, la plupart des racines peuvent avoir des formes dites « augmentées » avec des schèmes verbaux possédant un ou plusieurs sens et une réalisation graphique spécifique. Les verbes sont sujets à de nombreuses flexions, avec des réalisations différentes (avec affixes) selon l'aspect (accompli, inaccompli), la voix (passive / active), le temps, le mode (indicatif, subjonctif, impératif) et le sujet (accord en genre, nombre, personne). Notons que l'arabe connaît une forme duelle en plus du pluriel et du singulier. Cette forme duelle s'applique non seulement aux verbes, mais également aux substantifs et adjectifs, ainsi qu'aux pronoms démonstratifs et relatifs. Enfin, une orthographe malheureuse est régulièrement attestée dans les textes arabes. Parmi les cas les plus fréquents figurent la chute de la *hamza* (ء), l'écriture sans points diacritiques du *tāʾ marbūta* (ة, bien souvent la marque du féminin) ou du *yāʾ* final (ي), ou encore les cas de *taṭwīl / kašīda* (élongation graphique). Le tableau suivant illustre ces éléments avec quelques exemples :

Faute	Orthographe correcte	Orthographe déviante	Translittération	Traduction
Chute de la hamza (ء)	رأس	راس	*raʾs*	tête
***tāʾ marbūta* (ة) sans points diacritiques (ه)**	مكتبة	مكتبه	*maktaba*	bibliothèque
***yāʾ* final (ي) sans points diacritiques (ى)**	اقتصادي	اقتصادى	*iqtiṣādī*	économique
***taṭwīl / kašīda* (élongation graphique)**	تقطيع	تقطيـــــع	*taqṭīʿ*	découpage

Tableau 3 : Orthographes déviantes récurrentes en arabe

Finalement, dans le cadre d'un projet d'annotation et de traitement automatique des collocations dans un environnement trilingue, il est nécessaire de disposer d'un certain type de ressources. Dans la section suivante, nous tâcherons de définir la notion de corpus avec une attention particulière pour les corpus parallèles, avant de discuter des diverses applications de ces derniers et des différentes techniques d'alignement phrastique et lexical. Nous terminerons en présentant plus précisément les ressources que nous utiliserons pour notre projet.

5. CORPUS, LINGUISTIQUE DE CORPUS ET CORPUS PARALLELES

5.1. La notion de corpus

Comme toute notion en sciences humaines et sociales, apposer une définition définitive et universelle à un concept est une tâche ardue. La notion même de corpus peut s'avérer difficile à définir tant les formes se sont diversifiées et tant l'horizon des objectifs visés *via* leur utilisation s'est élargi au fil du temps. À l'instar du phénomène collocatif, l'abondance des définitions de *corpus* menait Pearson (1998) à écrire que le terme n'avait pas encore été totalement défini par la communauté linguistique.

Bien qu'il y ait de nombreuses manières de définir ce qu'est un corpus, certains points convergent vers un consensus : un corpus est une collection de textes à la fois lisibles par une machine, authentiques, et échantillonnés dans le but d'être aussi représentatifs que possible de la langue qui fait l'objet de l'étude (McEnery et al., 2006). Ces variables sont aussi importantes qu'elles sont difficiles à atteindre et à quantifier. En effet, il n'y a pas de méthode pour mesurer le degré de représentativité d'un corpus et / ou pour l'équilibrer, d'autant que ces critères sont fluides et fluctuent selon la question de recherche en vue.

Les définitions divergent mais dresser une typologie des différents corpus disponibles et leurs applications est possible.

5.1.1. Typologie et applications en linguistique

Pour toute étude de linguistique de corpus, il est nécessaire de connaître à la fois les types de corpus existants et disponibles et leurs différentes applications. En effet, parler de *corpora* au pluriel est plus exact que de parler de *corpus* au singulier car il n'existe pas un type de corpus unique qui serait adapté à toutes les situations de recherche (Tognini-Bonelli, 2001).

Pearson (1998) propose cinq catégories de corpus :

- Corpus généraux de référence *(general reference corpora)* : il s'agit de corpus largement homogènes créés pour représenter les variétés d'une langue et son vocabulaire caractéristique. Pour l'anglais, les corpus *Bank of English* (>200M mots) et *British National Corpus* (>100M mots) en sont des exemples. Les dictionnaires de langue générale créés à partir de corpus se basent sur des corpus généraux de référence.
- Corpus spécialisés *(specialized* ou *special corpora)* : il s'agit de corpus dans lesquels la langue utilisée est trop spécifique pour être considérée comme appartenant à la langue générale. Cela peut être dû par exemple au domaine spécifique traité (domaine médical, législatif) ou à la catégorie des locuteurs des textes constituant le corpus (enfants, apprenants d'une langue étrangère). La création de bases lexicales de termes d'une langue de spécialité peut être une activité issue de l'étude de ces corpus.
- Corpus-échantillon *(sample corpora) vs.* corpus de textes complets *(full-text corpora)*: les premiers ne contiennent que des extraits de textes tandis que les seconds comportent des textes dans leur intégralité.
- Corpus multilingues *(multilingual corpora)* : ces corpus sont notamment utilisés pour le développement de ressources bi- ou multilingues (nous y reviendrons plus largement dans la section suivante). Il en existe deux types :

    - Corpus comparable *(comparable corpora)* : il s'agit d'une collection de corpus le plus souvent monolingues qui peuvent être comparés car élaborés dans les mêmes conditions et traitant d'un même sujet.
    - Corpus parallèle *(parallel corpora)* : autre type de corpus multilingue, les corpus parallèles sont des collections de textes bi- ou multilingues contenant les textes originaux dans une ou plusieurs langues sources traduits vers une ou plusieurs langues cibles. Contrairement aux corpus comparables, les textes sont identiques car il s'agit d'originaux et de leur(s) traduction(s).
- Corpus pour un but spécifique *(special purpose corpora)* : cette catégorie (dont l'appellation ne fait pas l'unanimité) est proposée pour y classer les corpus créés avec un objectif spécifique en tête et qui ne peuvent pas être rangés dans les catégories mentionnées précédemment. Pearson donne en exemples la recherche de définitions ou l'analyse des problèmes liés au genre.

Cependant, à l'instar de la définition, cette typologie n'est pas fixée et, encore une fois, les différents types de corpus peuvent varier selon leur application en linguistique. Les études peuvent porter sur la grammaire (en synchronie ou en diachronie), sur la dialectologie et la variation linguistique, sur l'apprentissage et l'enseignement des langues, ou encore sur les différents champs telles que la sémantique, la pragmatique, la sociolinguistique, l'analyse de discours ou la stylistique et les études littéraires (McEnery et al., 2006).

La section suivante s'attèle à décrire en particulier un des objets principaux de ce projet, à savoir les corpus multilingues, et plus précisément les corpus parallèles. Nous tâcherons de les définir, de fournir des exemples et de parler de leurs applications en linguistique contrastive et en traduction. Nous aborderons ensuite les techniques d'alignement phrastique lexical utilisées pour créer ces ressources.

5.2. Corpus bi- et multilingues

Bien que la langue anglaise soit la langue dominante et de très loin, en atteste le nombre incalculable de ressources textuelles disponibles, les corpus dans d'autres langues se répandent de plus en plus. Ces corpus sont d'une très grande importance pour tout ce qui est relatif au développement de systèmes de traduction assistée par ordinateur (TAO), de bases de données lexicales et terminologiques multilingues, ou encore d'outils pour aider les traducteurs, les lexicographes et les chercheurs (Altenberg & Aijmer, 2000). L'on distingue cependant deux types de corpus multilingues : les corpus parallèles et les corpus comparables.

5.2.1. Corpus parallèles et comparables

Une fois de plus, la terminologie pour ces corpus peut apparaître floue et un certain nombre d'appellations sont attestées. Dans la littérature, ce que certains appellent *corpus parallèles* sont en fait des *corpus comparables* pour d'autres (McEnery & Xiao, 2007), tandis que d'autres encore vont parler de *corpus concurrents*, *corpus alignés* ou *corpus comparatifs* (Kenning, 2010).

Bien que ces deux types de corpus puissent sembler de prime abord relativement similaires, ce qui distingue principalement les corpus parallèles et les corpus comparables, c'est évidemment les données qui les composent (Mitkov, 2004), mais surtout le lien

qu'entretiennent les textes à l'intérieur du corpus. D'un côté, le corpus parallèle bilingue prototypique sera constitué d'un ou plusieurs textes dans une langue A et de sa ou leurs traductions vers une langue B. Les textes sont finalement équivalents et liés par le sens. De l'autre côté, les textes d'un corpus comparable sont liés parce qu'ils ont été collectés et classés ensemble sur la base de critères communs (leur taille, leur sujet, un auteur, une période, etc.), mais ils restent indépendants les uns des autres. Kenning (2010) postule que la différence majeure entre ces deux types ne réside pas dans le fait que le premier consiste en un ensemble de traductions et pas le second, mais plutôt dans le fait que les corpus parallèles impliquent un texte source commun.

Il est à noter que parfois les corpus parallèles peuvent avoir le texte original absent et ne présenter que des traductions dudit texte original, sauf dans un contexte de traduction automatique, dans lequel le texte original est nécessairement présent. Une autre caractéristique de ces corpus est le sens dans lequel va le processus de traduction. Ce dernier peut être unidirectionnel (de l'arabe au français ou du français vers l'arabe uniquement), bidirectionnel (des textes sources de l'arabe au français mais également des textes sources du français vers l'arabe) ou multidirectionnel (McEnery & Xiao, 2007). Kenning (2010) fait cependant remarquer que les corpus parallèles ne sont pas forcément des corpus bi- ou multilingues, car ils peuvent faire figurer plusieurs traductions dans une seule et même langue d'un texte source commun absent. On ne peut alors parler d'autre chose que de corpus parallèle monolingue.

La frontière est donc relativement fine entre un corpus parallèle et un corpus comparable. D'aucuns pourraient dire qu'un corpus parallèle, selon la manière dont il est divisé en sous-corpus, peut alors devenir un corpus comparable. Dans l'acception qu'un corpus parallèle contient un texte-source et sa ou ses traduction(s), certains ont montré la voie pour le développement de ce type de ressources.

5.2.2. Exemples de corpus parallèles

Malgré le gain en popularité des études comparatives depuis les années 1990 et malgré le très grand nombre de corpus qui sont construits chaque année, le manque de ressources parallèles et comparables se fait toujours ressentir. Qui plus est, la majorité des corpus parallèles existants conséquents couvrent relativement peu de genres (principalement de la fiction, des travaux parlementaires ou des manuels techniques) et les paires ou ensembles de langues sont souvent limités (Kenning, 2010). En effet, à moins qu'il ne s'agisse de textes bibliques ou de documents produits par des multinationales ou des institutions internationales, la plupart des ressources parallèles sont bilingues plutôt que multilingues, et certaines paires de langues sont évidemment bien plus représentées que d'autres.

Parmi les corpus multidirectionnels les plus impressionnants, on peut citer le corpus biblique de Christodouloupoulos et Steedman (2015). Les auteurs se sont efforcés de construire un corpus rassemblant des traductions complètes ou fragmentaires de la Bible dans 100 langues différentes, faisant passer le total de paires de langues possibles à 4950. Il s'agit de la plus grande quantité de textes parallèles disponible.

D'autres projets à grande échelle ont fourni des corpus très impressionnants. C'est le cas des corpus parallèles bilingues suivants : le corpus *Hansard French/English* (travaux du Parlement

canadien en anglais et en français canadien), le corpus *Arabic English parallel news* (articles journalistiques en arabe traduits en anglais), ou le corpus *Hong Kong Laws* (du chinois vers l'anglais) (Kenning, 2010). Dans une perspective multilingue, on pourra citer le corpus *JRC-Acquis* (textes législatifs européens avec plus de 20 langues), le projet *MULTEXT-East* dont nous avons parlé plus haut (de l'anglais vers 9 langues de l'Europe de l'Est), ou encore le *Oslo Multilingual Corpus* (textes sources en allemand, français et finnois et leurs traductions dans plusieurs combinaisons) (Lee, 2010).

Comme précisé, ces corpus parallèles ont plusieurs applications, que ce soit en linguistique contrastive, en traduction (humaine et automatique) et traductologie, dans l'enseignement des langues et en TAL. Nous discuterons de ces applications dans la section suivante.

5.2.3. Applications

5.2.3.1. Traduction humaine et automatique et traductologie

Les applications premières des corpus parallèles se trouvent dans le champ de la traduction (humaine et automatique) et de la linguistique contrastive (McEnery & Xiao, 2007). Dans le cas de la traduction, les études basées sur corpus se placent sur deux plans. Sur le plan théorique, c'est le processus même de la traduction qui entre en considération ; le transfert d'un concept d'une langue source vers une langue cible est au centre de la comparaison linguistique en jeu. Sur le plan pratique, ces corpus sont de bons outils pour les traducteurs en herbe, mais ils sont surtout de bons supports pour développer des systèmes de TA et de TAO.

Les corpus parallèles offrent de multiples possibilités aux traductologues de tester leurs hypothèses. D'abord fondées sur des intuitions ou sur un nombre limité d'exemples, ces grandes quantités de textes et leur contexte leur permettent de les confirmer ou de les infirmer. Cela est particulièrement vrai avec les textes bidirectionnels dans le cas où l'on cherche à montrer le seuil de correspondance mutuelle de deux éléments d'une langue à l'autre et vice versa (Kenning, 2010).

Pour les traducteurs, les corpus parallèles offrent la possibilité de chercher en contexte comment un terme ou une expression ont été traduits par leurs pairs et peuvent directement s'en inspirer. De ces corpus peuvent également être extraite une terminologie spécialisée (McEnery & Xiao, 2007), qui a son tour s'avèrera une aide précieuse pour le traducteur, notamment pour les domaines qui se développent très vite et pour lesquels la terminologie est en évolution permanente.

Cette technique d'extraction fait notamment partie des possibilités offertes par le TAL grâce aux corpus parallèles. Qu'il s'agisse de la lexicographie, de l'ingénierie des connaissances, de la construction de bases de données terminologiques et d'outils de références bilingues, c'est autant d'instruments mis à disposition de la traduction (Kenning, 2010). Ces corpus ont également joué un rôle majeur dans le développement de la TA, notamment dans le paradigme de la TA statistique et de la TA basée sur des exemples, qui a supplanté le paradigme précédent, qui adoptait une approche symbolique à base de règles.

Cependant, l'utilisation de ces corpus pour la traduction, bien que de plus en plus importante, reste controversée. En effet, une traduction reste une *interprétation* vers une langue cible d'un élément d'une langue source ; elle n'est jamais une production spontanée, et les données

obtenues dans une traduction d'une langue A seront inévitablement différentes des données d'un texte source dans cette même langue (McEnery & Xiao, 2007).

5.2.3.2. Linguistique contrastive

Les corpus parallèles, comme l'indique (Kenning, 2010), peuvent être utilisés à de très nombreuses fins dans le champ de la linguistique contrastive. Les comparaisons peuvent être faites à tous les niveaux : celui du lexique, de la syntaxe, du discours, des marqueurs de modalité, etc. Non seulement les (a)symétries d'une langue à une autre sont mises en exergue par les corpus parallèles, mais ces derniers peuvent également améliorer la précision des descriptions individuelles des langues.

À la suite des travaux de Johansson, Altenberg et Granger (2002) proposent de résumer l'utilité des corpus multilingues en linguistique appliquée comme suit :

- Ils offrent une base empirique solide qui fournit des informations fiables sur le degré de correspondance d'éléments lexicaux,
- Ils offrent de nouvelles perspectives sur les langues comparées qui auraient pu être occultées dans des études monolingues,
- Ils peuvent fournir des informations spécifiques à une langue, qu'elles soient typologiques ou culturelles, ainsi que des caractéristiques universelles,
- Ils mettent en lumière les différences entre les textes sources et leurs traductions, ainsi que les différences entre des textes produits par des locuteurs natifs et ceux produits par des non-natifs ou apprenants,

Ils peuvent être utilisés pour approfondir des théories mais également dans de nombreux champs pratiques, comme la lexicographie, l'enseignement des langues ou le TAL.

5.2.3.3. Enseignement des langues

Bien que l'usage des corpus parallèles soit très majoritairement ciblé pour la traduction et la linguistique contrastive, ces derniers ne s'y cantonnent pas pour autant et peuvent servir de base pour soutenir des buts pédagogiques et d'apprentissage des langues.

À l'instar des champs d'application cités précédemment, l'usage de corpus pour l'enseignement des langues est controversé. En effet, bien que la linguistique de corpus ait fait avancer la recherche à de multiples niveaux, et malgré le fait que bien souvent la langue contenue dans les ouvrages pédagogiques, donc destinée à être enseignée, soit basée sur des intuitions fautives sur la manière dont on utilise ladite langue (O'Keeffe & Farr, 2003), cette pratique reste contestée.

Pourtant, l'utilisation de corpus parallèles peut s'avérer bénéfique à de multiples égards. Du côté de l'enseignant, ils peuvent adapter les recherches dans ces corpus pour leurs propres buts pédagogiques, qu'ils pourront ensuite jauger et évaluer de sorte à les comparer avec ce qui peut être présenté dans les manuels scolaires classiques. D'autre part, les données langagières « réelles » trouvées en corpus leur permettraient de mieux gérer la recontextualisation et la médiation socioculturelle de la langue-culture étudiée (O'Keeffe & Farr, 2003). Enfin, les applications et / ou outils pédagogiques développés à partir desdits corpus pourraient aider les apprenants à développer leurs capacités linguistiques à tous les niveaux.

Pour toutes ces applications, les textes des corpus parallèles nécessitent au préalable d'être alignés efficacement et convenablement. Dans la section suivante, nous tâcherons d'en décrire quelques techniques.

5.3. Techniques d'alignement automatique

Dans le but de maximiser l'utilité des corpus multilingues, il est nécessaire d'appliquer des méthodes d'alignement de texte. En effet, une fois qu'un fragment d'un texte original se retrouve lié avec sa ou ses traductions directes, ces fragments peuvent être affichés les uns à côté des autres afin d'être comparés, notamment grâce à des concordanciers ou d'autres outils de recherche multilingue (Altenberg & Granger, 2002). Cette opération peut être menée au niveau du paragraphe, au niveau de la phrase, ou au niveau du mot. Bien qu'elle puisse paraître relativement aisée, cette opération est loin d'être triviale dans le cas de la construction de corpus parallèles avec des paires de langues qui ne partagent que très peu de caractéristiques communes (McEnery & Xiao, 2007).

5.3.1. Alignement phrastique

Plusieurs approches sont possibles pour aligner plusieurs textes. Certaines sont basées sur des analyses traditionnelles du texte, qu'il s'agisse d'analyse syntaxique, d'étiquetage ou du recours à des dictionnaires bi- ou multilingues ; d'autres sont entièrement automatiques (Somers, 2001). Les premiers programmes développés dans le but d'aligner des textes multilingues partaient du postulat relativement simple qu'un segment source et un segment cible sont d'une longueur plus ou moins équivalente (Gale & Church, 1993), tant sur le nombre de mots que sur le nombre de caractères. Cependant, il semble évident que ce postulat n'est vrai que pour des paires de langues qui sont assez similaires syntaxiquement mais également sur le plan lexical.

Malgré cela, « force est de constater que depuis les travaux pionniers du début des années 1990, peu de choses ont bougé » (Kraif, 2015). En effet, ces travaux étaient assez robustes pour demeurer utiles jusqu'à aujourd'hui et servir de base pour la plupart des systèmes actuels. L'alignement phrastique repose sur des corrélations diverses entre les textes étudiés, la plus triviale étant, nous l'avons dit, la longueur moyenne des phrases. Cependant, la traduction d'une phrase dans le texte source peut régulièrement se réaliser en plusieurs phrases dans le texte cible (ou le contraire), à telle enseigne que la longueur d'une phrase n'a plus vraiment d'importance, et la segmentation au niveau de la ponctuation ne peut pas non plus fonctionner (Kraif, 2014). De fait, d'autres éléments devront servir à déterminer des « points d'ancrage », afin de repérer quelles parties du texte cible correspondent à celles du texte source. Il y a tout d'abord les *cognats* (ou mots apparentés), des items lexicaux partageant une forte ressemblance d'une langue à l'autre car issus d'un fonds lexical commun (p. ex. *kangourou*, *kangaroo*, *Känguru*). Ces derniers sont facilement exploitables par la machine grâce à une comparaison de chaînes de caractères. Il y a ensuite les *transfuges*, c'est-à-dire des éléments qui passent d'une langue à l'autre sans altération, même si elles ne partagent pas le même script. C'est le cas par exemple de certaines entités nommées (noms propres, dates, etc.) et dans une moindre mesure la ponctuation.

Réaliser un alignement automatique revient à extraire automatiquement un « chemin d'alignement » (Kraif, 2014) représenté par une succession de points dans un espace bidimensionnel. Pour calculer l'alignement, deux types d'algorithmes sont utilisés. Les algorithmes matriciels quadrillent cet espace bidimensionnel pour identifier des points pour les couples de forme potentiellement alignables, qu'il s'agisse de ressemblances graphiques ou d'une distribution similaire, avant de les filtrer pour établir un quadrillage plus fin et réitérer le processus. Les algorithmes linéaires, eux, cherchent à effectuer des regroupements de phrases et calculer récursivement le meilleur chemin. Gale et Church (1991) utilisent ce type d'algorithmes en se basant sur la probabilité du rapport des longueurs des phrases pour déterminer si les aligner est statistiquement pertinent. En vue d'améliorer ces techniques, de nouvelles ont ensuite été développées combinant les divers indices et en les hiérarchisant dans le processus d'alignement. Ainsi, pour son outil Alinéa, Kraif (2001) a mis au point une architecture en trois temps : il procède tout d'abord à l'extraction de points d'ancrage à partir des transfuges, puis effectue la même chose à partir de la densité des cognats, avant de calculer l'alignement des phrases en se basant sur leur longueur. Ce traitement permet d'aller du plus similaire au plus arbitraire, le dernier processus évoluant ainsi dans un espace de recherche réduit grâce aux précédents.

D'autres systèmes performants existent pour l'alignement phrastique : *cwb-align* (Evert & Hardie, 2011), intégré au IMS CWB Open Corpus Workbench, un ensemble d'outils *open source* pour gérer et interroger des corpus avec des annotations linguistiques ; *Hunalign* (Varga et al., 2007), écrit en C++ et initialement conçu pour aligner des textes en langues moyennement dotées avec l'aide potentielle de ressources externes telles que des dictionnaires ou lexiques bilingues ; ou encore *NATools* (Simões & Almeida, 2003) et *YASA* (Lamraoui & Langlais, 2013). Il sera intéressant de noter que l'alignement phrastique *multitexte* (plutôt que *bitexte*) fait l'objet de développements relativement récents (Chiao et al., 2006), notamment avec les travaux de Kraif (2015) et son outil *JAM*.

5.3.2. Alignement lexical

L'alignement lexical, c'est-à-dire au niveau du mot ou des unités lexicales, suit généralement l'alignement phrastique et révèle un nouveau niveau de difficulté. En effet, les divergences linguistiques d'une langue à l'autre sont récurrentes et aligner des unités lexicales dans des phrases régies par des configurations syntaxiques différentes, ayant fait l'objet d'une transposition (autrement dit ayant changé de catégorie grammaticale) ou encore faisant partie d'expressions polylexicales, s'avère un problème de taille. Ce problème est d'autant plus complexe que définir ce qu'est un token au niveau du mot est une autre difficulté qu'il s'agit de surmonter, notamment dans le cas des langues comme le chinois où les frontières de mots ne sont pas symbolisées par une espace (Wu, 2000), comme c'est le cas avec le *wāw* (و, « et ») de coordination en arabe. Contrairement à l'alignement phrastique, l'alignement lexical est donc beaucoup moins *monotone* (c'est-à-dire « linéaire »). De fait, à moins que les langues alignées soient extrêmement similaires, les techniques basées sur la longueur ne sont que peu précises.

Deux approches sont généralement employées (Tiedemann, 2003). La première de ces approches est une approche d'estimation, notamment utilisée dans le cadre de la TA statistique.

Le principe est de modéliser les paramètres d'alignement comme des paramètres cachés (comme les modèles de Markov cachés) dans un modèle de TA statistique, comme pour l'outil précurseur développé par IBM, *GIZA++* (Och & Ney, 2003). La deuxième approche concerne plus généralement l'extraction de lexiques bilingues et se base sur les mesures d'association que nous avons déjà présentées, à l'instar de *word_align* (Dagan et al., 1999).

5.3.3. Mesure des performances : précision, rappel et F-mesure

Quelle que soit l'approche adoptée, comme pour beaucoup d'autres systèmes développés en TAL, les performances des outils d'alignement, qu'ils agissent au niveau phrastique ou lexical, se quantifient grâces à trois mesures : la précision, le rappel et la F-mesure (la moyenne harmonique de la précision et du rappel). Nous reprenons (Santos, 2011) et considérons que *A* est l'aligneur utilisé, *D* est l'ensemble de textes alignés et C_{total} le nombre de correspondances correctes entre eux. Après alignement, $T_{A(D)}$ représente le nombre total de correspondances trouvées par *A* dans *D*, et $C_{A(D)}$ le nombre de correspondances correctes identifiées par *A* dans *D*. La précision *P* et le rappel *R* sont alors calculés comme suit :

$$\mathbf{P} = \frac{C_{A(D)}}{T_{A(D)}}$$

$$\mathbf{R} = \frac{C_{A(D)}}{C_{total}}$$

La moyenne harmonique (*F-mesure*) est représentée par l'équation suivante :

$$\mathbf{F-mesure} = \frac{2 * P * R}{P + R}$$

Enfin, comme le font remarquer Kübler et Aston (2010), les mémoires de traduction, dans le cas où les textes source et cible ont été sauvegardés dans un format dans lequel les textes sont alignés, peuvent elles-mêmes être considérées comme des corpus parallèles. Elles offrent cependant moins d'informations sur le contexte car elles contiennent généralement uniquement des paires de segments et non pas des paires de textes. Malgré cela, dans le contexte de la traduction automatique, tous les corpus parallèles sont utilisés. Depuis, la plupart des outils de TAO mêmes proposent des fonctionnalités d'alignement lexical que l'on peut paramétrer. De plus, selon les besoins et ressources spécifiques à chaque projet de construction de corpus alignés, de nombreux outils et de nouvelles techniques d'alignement ont été développées. C'est le cas de certains des corpus que nous utiliserons pour notre travail. Nous tâcherons de les présenter dans la section suivante.

6. CONCLUSION ET DIRECTION DU PROJET

6.1. Récapitulatif

Au travers de cet état de l'art, nous avons couvert les sujets que nous serons amené à traiter au cours de notre projet. Après avoir introduit ce dernier en présentant nos motivations et nos objectifs, nous avons présenté les notions d'expressions polylexicales et de collocations, tout en exposant les différentes approches concernant ces dernières.

Nous avons ensuite couvert les différents pans de l'activité d'annotation, en nous attachant particulièrement aux spécificités de l'annotation des collocations pour chacune des trois langues de notre projet, leurs points communs et les difficultés anticipées.

Dans la section suivante, nous avons défini ce que l'on entendait par découverte et identification des collocations avant de présenter les différentes approches historiques utilisées pour l'extraction des collocations, jusqu'aux techniques les plus récentes. Nous avons fait un arrêt sur la langue arabe, langue pour laquelle le traitement automatique amène des difficultés supplémentaires.

Dans la section finale, nous avons fait un bref historique de la notion de corpus, avant d'en dresser une typologie, pour nous arrêter plus largement sur les corpus bi- et multilingues et leurs applications, qui sont au centre de notre projet. Nous avons finalement discuté des différentes techniques d'alignement, phrastique et lexical, pour terminer sur les métriques pour la mesure des performances.

II. METHODOLOGIE : ANNOTATION DES COLLOCATIONS DANS LE CORPUS FRANÇAIS

7. PRESENTATION DES RESSOURCES ET OUTILS UTILISES

7.1. Plan

La direction que prendra ce projet se fera en deux temps. En premier lieu, c'est la phase d'annotation du corpus français qui nous intéressera. Nous rédigerons un guide d'annotation basé sur les travaux de PARSEME (Savary et al., 2015) et de SimpleApprenant (Todirascu & Cargill, 2019). Nous nous attellerons ensuite à la construction de notre corpus trilingue multi-genre en échantillonnant 4 corpus parallèles que nous présentons *infra,* de sorte à obtenir environ 100 000 triplets de phrases alignées. Puis, nous adapterons et utiliserons VarIDE (Pasquer et al., 2018) pour annoter automatiquement le corpus en collocations verbales en français, avant de corriger et d'augmenter les résultats obtenus.

Dans un deuxième temps, nous nous attacherons à projeter les annotations du corpus français vers les corpus anglais et arabe avec ZAP (Akbik & Vollgraf, 2018). Au préalable, nous devrons utiliser GIZA++ (Och & Ney, 2003) pour la création de tables de traduction bilingues, notamment en ce qui concerne l'arabe, car l'outil de projection n'a pas été développé pour cette langue. Une fois l'adaptation du code source de ZAP faite, nous projetterons les annotations vers les corpus anglais et arabe, ce qui nous permettra, grâce à l'intégralité de nos données annotées, de mener une étude contrastive de l'usage des collocations entre les trois langues et entre les genres. Enfin, l'objectif final est d'évaluer l'outil d'annotation automatique VarIDE avec nos données annotées en langue arabe car, à notre connaissance, cela n'a jamais été fait pour cette langue jusqu'à présent.

7.2. Corpus parallèles exploités

Pour composer notre corpus parallèle trilingue et multi-genre, notre principale source pour obtenir les sous-corpus qui seraient susceptibles de nous intéresser a été OPUS (Tiedemann, 2012). En effet, collecter des données textuelles en vue de les exploiter et de les étudier nécessite la permission des instances intéressées et peut engendrer des coûts lorsqu'il s'agit d'œuvres sous copyright (O'Keeffe & McCarthy, 2010). C'est pour cela que nous avons opté pour des corpus librement exploitables. Outre les corpus collectés sur la plateforme OPUS, l'impressionnant corpus parallèle des Nations Unies, libre de droit lui aussi, a été notre choix de prédilection pour ce qui est relatif au domaine juridique. Chacun d'entre eux appartient à un genre différent, ce qui nous permettra de mener une étude comparative de l'usage des collocations entre les genres. Dans cette section, nous nous attacherons à les décrire.

7.2.1. Corpus parallèle des Nations Unies v1.0 (textes juridiques)

L'Organisation des Nations Unies (ONU) est une organisation internationale qui remplace la Société des Nations depuis 1945. À travers sa promotion du respect des droits de l'homme et du développement durable, des aides humanitaires qu'elle fournit, l'ONU vise principalement à maintenir la paix et la sécurité au niveau international, notamment pour ses 193 états-membres. Elle compte six langues officielles : l'arabe, le chinois, l'anglais, le français, le russe et l'espagnol.

Une institution d'une telle ampleur qui reconnaît et utilise un panel linguistique relativement large était susceptible de produire des données textuelles multilingues importantes. Dans leur papier, Ziemski et al. (2016) décrivent le processus de création et présentent les statistiques du

corpus parallèle officiel de l'organisation (que nous désignerons par le sigle UN), est en fait une collection de documents traduits manuellement sur 25 ans (de 1990 à 2014), de l'anglais vers les cinq autres langues officielles. La plupart des documents sont alignés par paires (par exemple anglais-français ou français-arabe), mais un sous-corpus contenant des textes alignés pour les six langues est également disponible.

Lors de la création du corpus parallèle, les auteurs ont différencié deux types de paires de langues : les paires primaires (l'anglais et une des cinq autres langues) et les paires secondaires (deux langues hors anglais). Le processus d'alignement quant à lui s'est fait en deux temps, afin d'assurer un alignement de haute qualité (Ziemski et al., 2016). Tout d'abord, les textes dans la langue qui n'est pas l'anglais sont d'abord traduits vers celle-ci et sont alignés au niveau phrastique avec l'outil *Hunalign* (Varga et al., 2007). Le bruit résultant des erreurs d'alignement est ensuite filtré, avant d'appliquer leur propre outil d'alignement monolingue BLEU-Champ qui applique l'algorithme *Champollion* (Ma, 2006). Enfin, avec le texte anglais tokenisé et le texte de l'autre langue traduit, BLEU-Champ crée un fichier échelle qui combine les deux fichiers TEI (anglais et l'autre langue) dans un fichier TEI unique avec un alignement au niveau de la phrase. Dans le cas des paires de langues secondaires, les deux textes sont traduits en anglais dans la première étape et sont alignés ensuite grâce à ces traductions.

Les statistiques du corpus sont impressionnantes. Pour ne pas surcharger le présent travail avec des chiffres, nous nous contenterons de ceux des paires de langues qui nous intéressent ici :

	EN / AR		**EN / FR**		**AR / FR**	
Documents	111 241		149 741		112 605	
Lignes	18,5M		25,8M		18,2M	
Tokens	512M	456,5M	668,5M	782,9M	452,8M	597,6M

Tableau 4 : Statistiques du corpus parallèle des Nations Unies pour notre trio de langues

7.2.2. Plateforme OPUS

La plateforme OPUS *(Open Parallel corpUS)* est une ressource accessible gratuitement en constante expansion. Son objectif principal est d'être une source importante pour obtenir des corpus parallèles, et ce dans un maximum de domaines possibles avec autant de combinaisons de langues possibles. Cette plateforme ne se cantonne pas pour autant de proposer des dizaines de corpus parallèles plus ou moins conséquents. En effet, comme le précise son fondateur dans son papier (Tiedemann, 2012), la plateforme vise également à proposer des outils pour traiter automatiquement les données présentes dans les corpus parallèles (mais également monolingues), notamment à travers d'outils d'annotation, ainsi que plusieurs interfaces pour explorer les données.

Nous l'avons vu, les corpus parallèles avaient tendance à être focalisés sur certains domaines et concernaient uniquement certaines paires de langues. Un des autres objectifs d'OPUS est d'élargir ces horizons-là, en incluant autant que possible les langues faiblement dotées et en fournissant un nombre toujours plus grand de corpus parallèles de qualité. À l'époque de la publication de son article, la plateforme de Tiedemann couvrait plus de 90 langues, soit plus de 3 800 paires de langues avec des données alignées au niveau de la phrase (2,7+ milliards de segments alignés pour plus de 40 milliards de tokens).

Les corpus proposés, disponibles sous divers formats (XML / XCES, TMX, texte brut), couvrent un panel relativement large de genres également, fondamental pour le présent travail. Les sous-sections suivantes décrivent ces corpus.

7.2.2.1. Corpus parallèle Global Voices v2018q4 (textes journalistiques)

Le corpus parallèle Global Voices (maintenant GV) comporte des articles journalistiques issus du site web[12] éponyme. Il est compilé et fourni par CASMACAT[13] et la version que nous utiliserons (v2018q4) a été ajustée pour la plateforme OPUS pour être réellement multilingue. Les statistiques totales sont correctes, bien que bien moindres si on les compare à celle du corpus précédemment présenté (46 langues, 5,4M segments alignés pour quelques 100M tokens).

Pour les paires de langues qui nous intéressent, le corpus GV dispose de :

	EN / AR	**EN / FR**	**AR / FR**
Paires de phrases	63 071	195 387	42 229
Mots	2,32M	7,38M	1,63M

Tableau 5 : Statistiques du corpus parallèle Global Voices v2018q4 pour notre trio de langues

7.2.2.2. Corpus parallèle TED 2020 (transcriptions de conférences orales)

Le corpus parallèle TED 2020 (maintenant TED) (Reimers & Gurevych, 2020) rassemble environ 4000 transcriptions de conférences orales données à l'occasion de différents TED Talks, traduites par une communauté de volontaires vers plus de 100 langues. Ces conférences couvrent une multitude de sujets différents et sont données à travers le monde entier par des conférenciers de tous horizons. Les statistiques des bitextes pour chacune de nos trois paires de langues sont données dans le tableau suivant :

	EN / AR	**EN / FR**	**AR / FR**
Paires de phrases	407 595	410 443	399 617
Mots	12.54M	14.08M	12.52M

Tableau 6 : Statistiques du corpus parallèle TED 2020 pour notre trio de langues

7.2.2.3. Corpus parallèle WikiMatrix (textes encyclopédiques)

WikiMatrix (maintenant WM) est un corpus parallèle rassemblant 135 millions de phrases alignées issues d'articles du site Wikipedia pour 1620 paires de langues. Ce corpus a été compilé par des chercheurs de Facebook IA (Schwenk et al., 2019) qui ont automatiquement extrait les phrases alignées en utilisant une approche basée sur des plongements phrastiques multilingues. Le corpus WM est intéressant du fait que plus de 100M de ses phrases alignées le sont avec deux langues autres que l'anglais, rendant les recherches possibles pour des paires de langues éloignées sans pour autant passer par l'anglais en tant que langue pivot. En voici les statistiques pour les trois paires de langue de notre projet :

[12] https://globalvoices.org/
[13] http://casmacat.eu/corpus/global-voices.html

	EN / AR	EN / FR	AR / FR
Paires de phrases	999 763	2.75M	163 550
Mots	41.98M	120.89M	5.90M

Tableau 7 : Statistiques du corpus parallèle WikiMatrix pour notre trio de langues

7.3. VarIDE

7.3.1. Fonctionnement de l'outil

L'hypothèse principale qui a mené au développement de VarIDE *(Variant IDEntification)* est la suivante : l'identification des EP verbales devrait être plus efficace si l'apprentissage des différents motifs de variabilité morphologique et / ou syntaxique est effectué (Pasquer, 2017). Son fonctionnement est décrit dans Pasquer et al. (2018) et nous le reprenons ci-après.

Le programme applique une chaîne de traitement en 3 étapes successives. Dans un premier temps, l'outil extrait un très large ensemble de candidats potentiels (ensemble accompagné d'énormément de bruit). Cette extraction est basée sur les patrons syntaxiques les plus fréquents pour les EP verbales annotées dans le corpus d'entraînement. Ces patrons syntaxiques sont générés après une étape de normalisation, dans laquelle les étiquettes grammaticales des composants de l'EP sont extraits, classés lexicographiquement pour obtenir des tuples du type (NOUN,VERB). Une deuxième étape de normalisation a lieu pour les lemmes, afin de neutraliser l'ordre et les différentes flexions des mots. On obtient alors des tuples, toujours classés lexicographiquement, du type (hommage,rendre). Etant donné qu'une EP verbale n'apparaît généralement pas sous toutes ses formes flexionnelles dans le corpus d'entraînement, les variantes d'un tuple de lemmes sont générées. Chaque candidat doit avoir un patron syntaxique autorisé par la première normalisation en vue d'être extrait. Un filtre de longueur 20 est également appliqué, limitant le nombre de mots entre le premier et le dernier élément d'une EP à 20. Selon le résultat de cette extraction, le candidat est considéré comme positif ou négatif.

Dans un deuxième temps, ce sont les caractéristiques morphosyntaxiques des candidats qui sont extraites en vue de la classification finale. Ces informations (morphologiques et relations de dépendances syntaxiques) sont présentes dans les corpus au format `conllu` (`cupt` pour PARSEME). Pour VarIDE, elles sont à classer dans deux groupes :

- Les caractéristiques absolues `(ABS)` : elles sont obtenues localement, c'est-à-dire sur les composants du candidat eux-mêmes. Par exemple, pour un tuple normalisé de lemmes <*poser*,*question*> issu de la phrase *Il a posé une question*, une des caractéristiques absolues obtenues serait `ABS_morph_NOUN_Number=singular`, car le nom est au singulier.

- Les caractéristiques relatives `(REL)` : elles sont obtenues par comparaison du candidat avec tous les autres candidats partageant le même tuple normalisé de lemmes, à l'exception de lui-même. Ces dernières informations servent à calculer la similarité d'un candidat avec le reste des EP annotées. Les caractéristiques relatives peuvent prendre trois valeurs : *false* quand il n'y a aucune équivalence avec une EP verbale du corpus d'entraînement, *true* s'il en existe au moins une, et *-1* quand la comparaison n'est pas faisable (une seule occurrence). Par exemple, tout en admettant que ce sont les deux seuls candidats partageant le tuple de lemmes

<*poser,question*>, si l'on devait comparer le tuple introduit plus haut et le comparer à un tuple équivalent mais issu d'une phrase différente (p. ex. *Posez votre question !*), une des caractéristiques relatives obtenues serait `REL_morph_NOUN_Number=true`, car le nombre du substantif est singulier dans les deux cas. En revanche, une autre caractéristique absolue serait `REL_morph_VERB_Mood=false`, car le mode du verbe est différent entre le premier et le second tuple (indicatif et impératif respectivement).

Ces deux premières étapes constituent la phase d'entraînement. Une fois complétée, c'est au tour de la phase de prédiction d'avoir lieu. Tout d'abord, tous les candidats du corpus de test sont extraits, suivant la même méthodologie que pour le corpus d'entraînement, à la différence qu'ils ne sont considérés ni positifs ni négatifs. L'obtention des caractéristiques absolues sont obtenues de la même manière que décrite précédemment, tandis que les caractéristiques relatives sont obtenues après comparaison avec les EP verbales du corpus d'entraînement avec le même tuple normalisé de lemmes, dont la valeur booléenne est fixée selon s'il y a un résultat positif dans le corpus d'entraînement ou pas. Enfin, le classifieur bayésien naïf de la librairie Python `nltk` prend le relais et, en prenant en compte leurs caractéristiques, classent les candidats en positifs ou négatifs.

Un élément important est à prendre en considération pour cet outil : il a été développé dans l'optique d'identifier uniquement les variantes d'EP verbales annotées dans un corpus d'entraînement. Il est donc entièrement dépendant d'annotations manuelles effectuées au préalable et ne peut pas annoter d'EP qu'il n'aurait pas apprises. Ainsi, toute annotation automatique réalisée par VarIDE recevra inévitablement une augmentation manuelle humaine après elle. Dans la sous-section suivante, nous nous attachons à présenter les résultats obtenus par cet outil au cours de la Shared Task v1.1 de PARSEME (Ramisch et al., 2018).

7.3.2. PARSEME Shared Task v1.1

VarIDE a pris part à la Shared Task v1.1 de PARSEME, qui a rassemblé 12 équipes de 9 pays différents pour le développement de systèmes d'identification d'EP. VarIDE a obtenu des résultats honorables, se classant en 5e position du classement général sur les 13 systèmes présentés[14]. Avec une précision de 61,49, un rappel de 36,71 et une F-mesure de 45,97, l'outil peut être considéré comme performant compte tenu de la difficulté que représente l'identification des différents types d'EP verbales reconnues par PARSEME, qui plus est dans 19 langues différentes. Malgré les résultats moyens en anglais et l'absence de résultats pour l'arabe pour des raisons de licence, il sera intéressant de noter que le système se place en 2e position pour ce qui est de l'identification d'EP verbales en français. Avec une précision de 55,24, un rappel de 46,59 et une F-mesure de 50,54, ces résultats nous permettent d'anticiper des résultats qui devraient s'avérer très corrects pour notre propre projet, d'autant plus que nous nous concentrons uniquement sur un type d'EP verbale, les collocations, qui sont sensiblement moins complexes à identifier automatiquement que les expressions idiomatiques.

[14] Source : multiword.sourceforge.net

7.4. Guide d'annotation

Notre guide d'annotation[15] a été le premier élément à rédiger *in extenso* pour pouvoir mener à bien notre projet d'annotation, notamment en ce qui concerne l'annotation du corpus d'entraînement. Nous avons choisi de le baser sur deux projets d'envergure existants : SimpleApprenant (Todirascu & Cargill, 2019) et PARSEME (Savary et al., 2015).

7.4.1. Délimitation

En ce qui concerne la délimitation des annotations, nous avons fait plusieurs choix obligatoires vis-à-vis du format des corpus traités par VarIDE (format `conllu` étendu). L'annotation se fait dans la 11^{e} colonne du fichier : le premier token de l'annotation est signalé par un chiffre, deux-points et l'annotation `COLL` (p. ex. `1:COLL`), les tokens suivants de l'expression sont signalés par le même chiffre uniquement. Tous les tokens ne faisant partie d'aucune collocation sont signalés par un astérisque (`*`). Des exemples sont donnés dans le guide d'annotation.

L'annotation prend en considération uniquement les termes faisant directement partie de la collocation, à savoir le substantif et le verbe (et sa préposition le cas échéant). Tous les modifieurs adjectivaux et adverbiaux, les déterminants, ou encore les relatives ne font pas partie de l'annotation. Lorsque le verbe est conjugué à un temps composé, l'auxiliaire accompagnant le participe passé est également ignoré. Notons par ailleurs que si le verbe de la collocation est sous une forme participiale (p. ex. *prenant les mesures nécessaires à…*), il doit être annoté. Quant aux pronoms réfléchis, ils sont à annoter s'ils font partie intégrante du régime du verbe.

De fait, bien que les expressions puissent être continues, elles sont principalement discontinues (voir les résultats de l'annotation *infra*).

7.4.2. Tests linguistiques : SimpleApprenant

Les tests linguistiques pour identifier si un candidat-collocation est réellement une collocation sont issus du projet SimpleApprenant (Todirascu & Cargill, 2019). Ce dernier est un projet de création de ressources linguistiques autour des EP verbales pour les apprenants du français langue étrangère (FLE). Trois catégories d'expressions polylexicales sont considérées dans ce projet : les expressions figées, les expressions idiomatiques et les collocations.

Pour distinguer ces dernières des autres, l'équipe qui a travaillé sur le projet a formulé les tests linguistiques suivants :

- (T1) Passage à la diathèse passive :

  ils ont abrogé cette loi → cette loi a été abrogée [par eux]

- (T2) Changement de déterminant :

  ils ont abrogé cette / une / la loi

- (T3) Ajout de modifieurs :

  ils ont rapidement abrogé cette loi injuste

[15] Voir Annexe A.

Si le candidat ne récuse pas ces tests, l'expression peut être une collocation. Outre ces tests, les collocations se distinguent des expressions idiomatiques car leur sens est plutôt compositionnel, tandis que celui des expressions idiomatiques est bien souvent métaphorique ou figuré *(jeter l'éponge, passer au bleu)*. De plus, elles présentent une plus grande variabilité syntaxique.

En ce qui concerne les expressions figées, elles se distinguent des collocations en ce que leur objet est totalement fixe *(être d'accord, faire confiance)* et leur sens peut être pragmatique *(il fait beau)*. De plus, la tête de ces expressions est souvent un verbe faible *(être, faire, avoir)* et elles n'acceptent que peu de modifications syntaxiques.

7.4.3. Arbre de décision : PARSEME

Le modèle de prise de décision à l'aide d'un arbre qui décline les différents tests linguistiques à appliquer s'inspire directement des travaux de PARSEME (Savary et al., 2018). Bien que les catégories considérées dans le projet PARSEME soient plus nombreuses et malgré le fait que les collocations ne constituent pas une catégorie d'expressions polylexicales à part entière dans leur cadre théorique[16], ce modèle nous a paru à la fois approprié et efficace.

Reprenant les tests linguistiques présentés précédemment, l'annotateur humain peut décider au fur et à mesure qu'il descend dans l'arbre d'éliminer ou de conserver un candidat-collocation. Il peut commencer par identifier un candidat-collocation et se poser la question préliminaire de savoir si ce dernier dispose d'un élément verbal (qui ne soit pas un verbe faible) et d'un élément nominal en relation de dépendance syntaxique. S'il répond par l'affirmative, il peut passer au test T1 et vérifier que le passage à la diathèse passive est faisable. Si le test n'est pas récusé, il peut passer au test T2 et vérifier que le changement de déterminant est faisable. Dans le cas où ça fonctionne, il peut passer au test T3 et vérifier que l'ajout ou la suppression d'un modifieur adjectival ou adverbial est faisable. Si ce dernier test est accepté et que l'information mutuelle (IM) de la collocation est suffisamment élevée (en la vérifiant sur une ressource telle que Voisins de Wikipédia), la collocation peut être annotée.

[16] La traduction est de nous : « Une collocation est une cooccurrence de mots dont l'idiosyncrasie est uniquement de nature statistique. Les collocations ne sont pas considérées comme des expressions polylexicales verbales dans cette tâche. » Source : https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/?page=080_Glossary

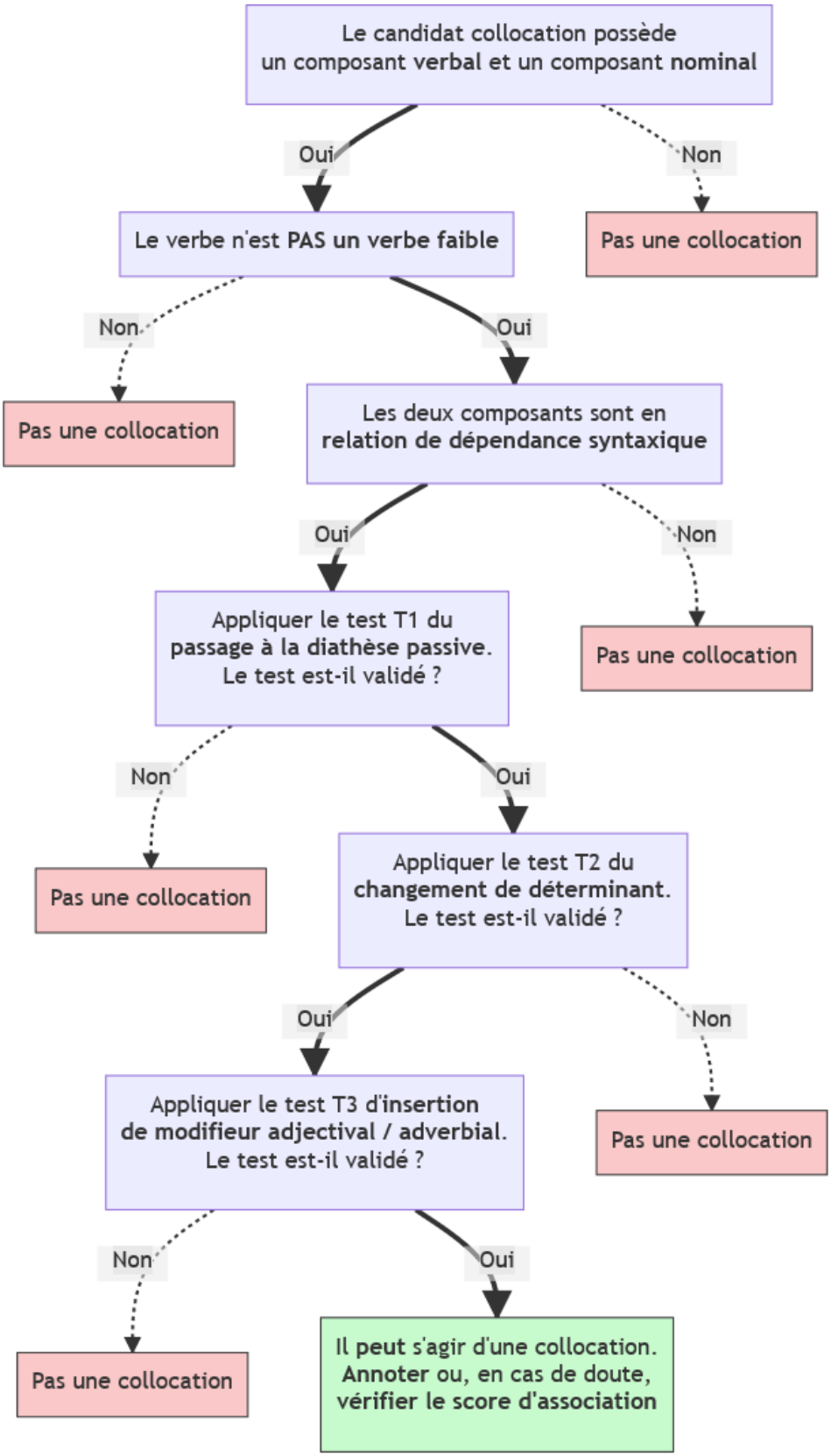

Figure 1 : Arbre de décision pour l'annotation d'un candidat-collocation

Après avoir sélectionné les corpus, choisi et pris en main l'outil d'annotation automatique et rédigé le guide d'annotation, la phase d'annotation, d'abord manuelle puis automatique, pouvait démarrer.

8. ANNOTATION AUTOMATIQUE

8.1. Préparation à l'annotation automatique

Avant de pouvoir utiliser l'outil d'annotation automatique, quelques étapes préalables étaient nécessaires : le corpus d'entraînement devait être adapté, les corpus parallèles devaient être échantillonnés, et ces mêmes échantillons devaient être convertis au format adéquat. Dans cette section, nous nous attacherons à décrire les différentes étapes nécessaires à la préparation des données.

8.1.1. Corpus d'entraînement

Les résultats d'une annotation automatique dépendent en grande partie de la qualité des données sur lesquelles le système a été préalablement entraîné. Pour obtenir une annotation de qualité, nous avons adapté le corpus d'entraînement français annoté manuellement en EP verbales qui avait été fourni PARSEME lors de la Shared Task 1.1 (Ramisch et al., 2018),. Il est constitué de 17 225 phrases issues de divers corpus et a servi à l'évaluation d'outils comme Veyn ou VarIDE.

Pour adapter ce corpus à notre projet, de multiples changements ont dû être apportés. En effet, toutes les EP verbales considérées dans le projet PARSEME ne sont pas exploitables pour nous. Trois d'entre elles pouvaient d'ores et déjà être ignorées et remplacées par des astérisques, car aucune ne respecte notre contrainte morphosyntaxique pour les collocations verbo-nominales :

- Les constructions multi-verbes (`MVC`) : ces constructions, comme *faire savoir* ou *laisser tomber*, sont une juxtaposition de deux verbes.
- Les constructions verbe-particule (`VPC`) : ces constructions n'existent pas en français et ne concernent donc pas notre corpus d'entraînement. En anglais, il s'agit de constructions comme *give up* (« abandonner ») ou *sleep in* (« faire la grasse matinée »), soit la juxtaposition d'un verbe et d'une particule adverbiale.
- Les verbes intrinsèquement réflexifs (`IRV`) : il s'agit de verbes pronominaux comme *s'incliner* ou *se rapprocher*, soit la juxtaposition d'un pronom et d'un verbe.

Deux autres types de constructions qui, contrairement aux précédentes, incorporent des éléments nominaux, ont été modifiées en `COLL` dans un premier temps :

- Les constructions à verbe faible (`LVC`) : ces constructions englobent des constructions avec de « vrais » verbes faibless (comme *faire l'historique* ou *avoir conscience*), mais aussi avec des verbes pleins (comme *apporter un témoignage* ou *dresser un bilan*).
- Les idiomes verbaux (`VID`) : ces constructions englobent les expressions idiomatiques (comme *jeter l'éponge* ou *couper l'herbe sous le pied*), des tournures de phrase idiomatiques (comme *il y a, il est question de* ou encore *il faut*), mais aussi des expressions avec un sens plus compositionnel (comme *poser une question* ou *attirer l'attention*).

Ces expressions ont en effet de bonnes chances d'être considérées comme d'éventuelles collocations. Ensuite, nous avons vérifié l'exactitude de ces annotations, en supprimant les expressions récurrentes qui ne correspondaient pas à notre définition à l'aide d'expressions régulières et en conservant les cas (minoritaires) où l'annotation était correcte. La définition que nous avons adoptée diverge en effet des choix faits dans le cadre du projet PARSEME. Nous ne considérons pas que les `LVC` avec les verbes *être*, *avoir* ou encore *faire* puissent être considérées comme des collocations. En outre, les `VID` sur le modèle de *jeter l'éponge* ou *il y a* ne répondent pas non plus positivement aux tests d'identification des collocations verbales. En revanche, celles ayant un sens plus compositionnel ont été conservées. C'est le cas, par exemple, des `LVC` *apporter un témoignage* et *dresser un bilan* et des `VID` *poser un problème* et *attirer l'attention*, qui répondent toutes les quatre positivement à la fois à notre définition et à nos tests.

Une fois toutes ces modifications effectuées, nous avons repassé chacune des phrases du corpus et ajouté manuellement les expressions répondant au phénomène collocatif. Enfin, à l'aide d'un script d'extraction, nous avons corrigé les quelques erreurs commises lors de l'annotation manuelle, puis vérifié la validité du format `cupt` grâce à un script de validation fourni par PARSEME. En voici les statistiques finales après extraction des annotations et classement lexicographique des patrons de collocation et de leurs composants :

Nombre de phrases	17 225
Nombre de tokens	432 389
Nombre total d'annotations	958
Nombre de patrons uniques	469
Pourcentage de phrases avec collocation	5,56%
Pourcentage de tokens annotés	0,45%

Tableau 8 : Statistiques du corpus d'entraînement français

Parmi les patrons uniques de collocations, on compte au moins deux occurrences pour 149 d'entre eux, dont 15 seulement apparaissent au moins 10 fois. Les patrons de collocation les plus fréquents, leur nombre d'occurrences et leur fréquence relative sont illustrés dans le diagramme suivant :

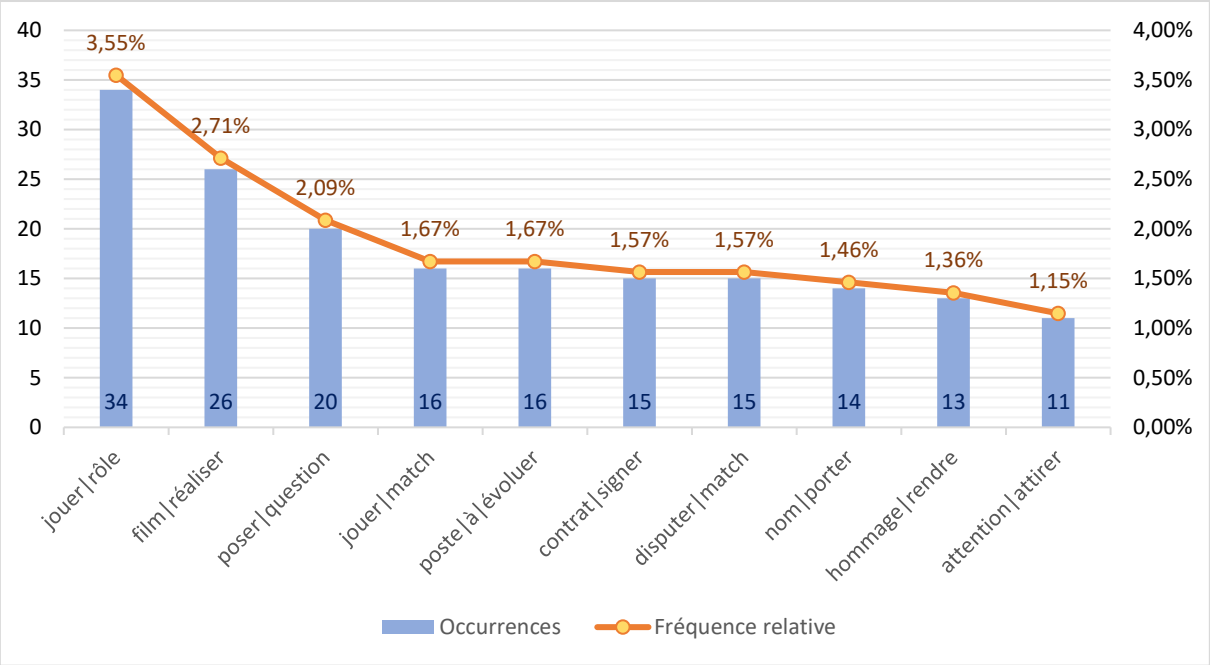

Figure 2 : Collocations les plus fréquentes (corpus d'entraînement)

Compte tenu du nombre relativement important de phrases traitant les domaines du cinéma et du football dans le corpus, il n'est pas étonnant de voir de nombreuses occurrences de collocations comme *disputer|match*, *jouer|match* ou *film|réaliser*.

8.1.2. Echantillonnage et création de « tritextes »

La préparation du corpus d'entraînement a été une tâche chronophage et complexe, mais une fois terminée, nous pouvions procéder à l'annotation automatique. Étant donné la portée modeste d'un travail de mémoire de master, nous ne pouvions imaginer travailler sur les corpus parallèles complets. Ainsi, il a fallu les échantillonner. Pour affiner un peu plus notre travail, nous avons décidé de créer des « tritextes », c'est-à-dire des triplets de phrases alignées dans les trois langues.

Pour construire ces tritextes, nous avons écrit un script qui, étant donnés deux bitextes B_1 et B_2 (p. ex. EN-FR et EN-AR), recherchait les phrases anglaises communes à B_1 et B_2 et en extrayait la phrase française alignée dans B_1 et la phrase arabe dans B_2. Selon le corpus concerné, des filtres supplémentaires étaient appliqués. Par exemple, pour le corpus parallèle UN, nous avons appliqué des filtres concernant la longueur minimum et maximum des segments, pour ne pas nous retrouver avec des noms de pays ou des noms de participants aux débats seuls, ou encore avec le nom de certains articles, chapitres et autres décrets. Nous avons également dû appliquer un filtre pour ne pas prendre en compte les doublons qui étaient, malheureusement, très nombreux dans le corpus parallèle GV.

Après correction manuelle des quelques erreurs qui subsistaient dans la création de ces tritextes, voici les statistiques finales du corpus complet et de ses sous-corpus :

Corpus	Nombre de phrases	Nombre de tokens		
		FR	EN	AR
GV	25 544	863 332	747 450	714 510
TED	29 296	617 678	571 975	538 550
UN	20 899	900 238	748 491	771 881
WM	29 992	976 769	852 323	879 849
Tous	105 731	3 358 017	2 920 239	2 904 790

Tableau 9 : Statistiques du corpus parallèle trilingue (phrases et tokens)

On pourra constater que, malgré l'apparent déséquilibre des sous-corpus quant au nombre de phrases, ils sont finalement relativement équilibrés en termes de tokens.

8.1.3. Conversion des « tritextes » au format `cupt`

VarIDE, comme tous les outils développés dans le cadre de PARSEME, traite des corpus au format `cupt`, c'est-à-dire dans un format `conllu` étendu. En plus des 10 colonnes « classiques » dédiées au numéro du token, à sa forme, à son lemme, etc., le format `cupt` introduit une 11e colonne recevant les annotations d'EP, format d'annotation que nous avons décrit plus haut.

Pour convertir les tritextes toujours au format texte brut en `cupt`[17], nous avons utilisé deux scripts. Le premier (et le plus important) nous a servi à passer du format texte brut au format `conllu` classique avec la bibliothèque Python `stanza` (Qi et al., 2020). Pour chaque phrase d'un fichier texte, le programme transforme la phrase au format `conllu` et l'écrit dans un fichier. Tous les fichiers contenant une phrase au format `conllu` sont finalement fusionnés en un seul. Le second script quant à lui nous a servi à transformer le fichier `conllu` dans un format `cupt` valide, c'est-à-dire avec l'insertion d'une 11e colonne contenant des astérisques, une ligne d'en-tête contenant la désignation de chaque colonne, ainsi que deux lignes avant chaque phrase, la première contenant la source et l'identifiant de la phrase (`# source_sent_id`) et la seconde la phrase elle-même (`# text`).

Notons que nous avons mené quelques tests pour les textes en arabe. Partant de l'hypothèse que la conversion au format `conllu` pourrait être plus performante et précise avec du texte avec les diacritiques complétés, nous avons effectué une comparaison des résultats de `stanza` avec et sans diacritiques sur une dizaine de phrases. Pour ce faire, nous avons utilisé la bibliothèque Python `CAMeL tools` (Obeid et al., 2020). Les résultats sont sans appel : paradoxalement, les résultats de `stanza` sont substantiellement meilleurs sans aucun signe diacritique ajouté (malgré la grande qualité de la diacritisation automatique de l'outil). La majorité des tokens ne sont alors pas reconnus. Nous avons donc décidé, pour les textes arabes, de faire exactement le contraire de ce que nous anticipions initialement, c'est-à-dire de supprimer les diacritiques complètement, afin d'améliorer la qualité du fichier `conllu`.

8.2. Evaluation

Le corpus d'entraînement finalisé, les tritextes construits et convertis en `cupt`, nous pouvions dès lors procéder à l'annotation automatique. Dans cette section, nous présentons tout d'abord l'évaluation standard de l'outil d'annotation automatique VarIDE. Ensuite, nous détaillons la

[17] Voir Annexe C pour un exemple de phrase au format `cupt`.

procédure que nous avons suivie pour effectuer l'annotation automatique, la corriger et l'augmenter, puis nous dressons une typologie des erreurs d'annotation, avant de présenter nos résultats et d'en proposer des interprétations.

8.2.1. Evaluation standard

Après avoir modifié légèrement le code source du programme afin qu'il fonctionne pour nos annotations et pour ne considérer que le modèle français (dans le fichier `code/config.cfg`), lancer le processus d'annotation automatique de VarIDE se fait simplement à l'aide de la commande `python3 varIDE.py` dans le terminal. Le programme s'exécute comme décrit précédemment (voir section 7.3.1) et génère les fichiers tableurs correspondants.

Pour effectuer une évaluation standard, nous avons utilisé notre corpus d'entraînement entièrement annoté manuellement. De ce dernier, nous avons extrait aléatoirement 500 phrases afin de constituer un jeu de test. Ces phrases ont été soustraites au corpus d'entraînement dans le cadre de l'évaluation. Le tableau suivant[18] résume le contenu de chacun de ces corpus :

	Corpus d'entraînement	**Corpus de test**
Nombre de phrases	17 175	500
Nombre d'annotations	912	46

Tableau 10 : Statistiques des corpus d'entraînement et de test pour l'évaluation standard

Les 500 phrases annotées extraites du corpus d'entraînement ont servi de corpus de référence, tandis que les annotations ont été supprimées du corpus de test. Après annotation automatique avec VarIDE, nous obtenons les résultats suivants :

Précision	**Rappel**	**F-mesure**
34 / 38 = **89,47**	34 / 46 = **73,91**	**80,95**

Tableau 11 : Résultats de l'évaluation standard de VarIDE (précision, rappel, F-mesure)

On constate que l'annotation automatique des collocations verbo-nominales en français avec VarIDE obtient des résultats corrects. Malgré un score de rappel de 73,91, son score de précision proche de 90% mène à une moyenne harmonique de 80,95. Ces résultats sont très encourageants pour l'annotation automatique massive de la partie française du corpus parallèle de notre projet. Les sous-sections suivantes s'attachent à en décrire les résultats.

8.2.2. Evaluation de notre corpus parallèle

Avant toute correction des résultats, les statistiques obtenues pour chaque sous-corpus ont été les suivantes :

[18] Dans la colonne Précision : le premier nombre correspond aux vrais positifs (annotations correctes), le second à toutes les annotations automatiques (soit la somme des vrais positifs et des faux positifs).
Dans la colonne Rappel : le premier nombre correspond aux vrais positifs (annotations correctes), le second à toutes les annotations après correction et augmentation (soit la somme des vrais positifs et des faux négatifs).
Le chiffre décimal suivant le signe « égale » correspond au résultat.

Corpus	Nombre de collocations	Pourcentage de tokens annotés	Pourcentage de phrases avec collocations
GV	1 577 (22,63%)	0,36%	6,17%
TED	792 (11,37%)	0,25%	2,70%
UN	2 829 (40,60%)	0,62%	13,53%
WM	1 770 (25,40%)	0,36%	5,90%
Tous	6 968	0,41%	6,59%

Tableau 12 : Statistiques de l'annotation automatique avant correction

On peut constater d'emblée un certain déséquilibre entre les corpus. Si GV et WM sont très proches, TED et UN sont très différents. En effet, une collocation apparaît dans 13,53% des phrases du corpus onusien, tandis qu'elles apparaissent dans moins de 3% des phrases des conférences orales de TED. Ces données nous donnent d'ores et déjà une idée de la disparité qu'il peut exister entre le registre oral et le registre écrit. En revanche, GV et WM obtiennent des résultats extrêmement proches, avec chacun 0,36% des tokens totaux annotés.

Après avoir obtenu les résultats de l'annotation automatique, la suite s'est faite en deux temps. Tout d'abord, nous avons corrigé les faux positifs. Autrement dit, nous avons supprimé les annotations qui n'auraient pas dû être faites, ce qui nous a permis ensuite de calculer la précision de l'outil. Pour ce faire, nous avons passé en revue manuellement toutes les phrases contenant des annotations pour vérifier leur exactitude.

Ensuite, nous avons augmenté le corpus avec les annotations que l'outil n'a pas faites car les collocations en question n'avaient pas été apprises pendant la phase d'entraînement, ne faisant pas partie du corpus d'entraînement utilisé. Autrement dit, nous avons ajouté les faux négatifs, ce qui nous a permis de calculer le score de rappel de l'outil. Pour ce faire, plutôt que de passer en revue toutes les phrases du corpus une à une, nous avons procédé différemment : nous avons tout d'abord extrait tous les lemmes des tokens dont l'étiquette grammaticale était `VERB`, supprimé les doublons et les erreurs de lemmatisation. Avec cette liste de plus de 3 500 verbes, nous avons consulté les Voisins de Wikipédia pour examiner les résultats pour chacun d'entre eux. Quand un candidat-collocation semblait possible et ne faisait pas partie des patrons du corpus d'entraînement, nous interrogions le contenu du corpus annoté pour voir si ledit candidat y était présent. Pour ce faire, nous avons rédigé un programme court permettant de saisir les deux termes du candidat (un verbe racinisé et un substantif sous sa forme lemmatisée) pour n'afficher que les phrases et leur identifiant contenant les deux saisies. Nous pouvions ainsi rapidement identifier quelles phrases contenaient les deux termes du candidat et si, en l'occurrence, il s'agissait bien d'une collocation, nous l'annotions.

Avant de présenter les résultats finaux de l'annotation semi-automatique et d'en discuter, nous présentons dans la sous-section suivante une typologie des erreurs commises par l'outil.

8.2.3. Typologie des erreurs commises par l'outil

Les annotations erronées faites par l'outil sont de natures diverses. Nous en dressons ici la typologie identifiée.

8.2.3.1. Erreur de lemmatisation préalable

Tout d'abord, lorsque le programme transformant les textes bruts au format `conllu` comportait des erreurs au niveau des lemmes des verbes ou des substantifs, VarIDE n'a pas pu éviter d'annoter incorrectement. Par exemple, dans la phrase « 5. *Prie* le Directeur exécutif de rendre compte de l'exécution de la présente *décision* au Conseil à sa dix-huitième session ordinaire. » (UN:722), l'impératif *Prie* en début de segment a été lemmatisé en *prendre*, ce qui a mené à l'annotation fautive *décision*|*prendre*.

8.2.3.2. Différence de traitement `conllu`

De la même manière, le traitement au format `conllu` de certains tokens est différent entre le corpus d'entraînement et notre corpus. Par exemple, les pronoms relatifs amalgamés du type *au(x)quel(s)* ne sont pas entièrement développés dans le corpus d'entraînement, alors que notre programme décline bien le token *auxquelles* en la préposition *à* et le pronom *lesquelles*. Ainsi, ce qui est annoté par défaut *confronter*|*difficulté* dans le corpus d'entraînement n'est pas annoté automatiquement comme nous le voudrions dans notre corpus, à savoir *confronter*|*difficulté*|*à*.

8.2.3.3. Cas de polysémie

Un autre type d'erreur récurrente était dû à la présence de deux acceptions d'un terme polysémique faisant partie d'une collocation dans un même segment. Par exemple, dans la phrase « (…) l'Etat partie importateur peut *prendre* dans toute la *mesure* du possible des *mesures* bilatérales appropriées (…) » (UN:2054), le verbe *prendre* a été annoté deux fois avec le substantif *mesure*, alors que seul le second est correct, le premier faisant partie du groupe prépositionnel *dans la mesure du possible*.

8.2.3.4. Longs segments et absence de dépendance syntaxique

D'autres erreurs étaient commises lorsque plusieurs combinaisons de verbes et de substantifs faisant partie de différentes collocations identifiées dans le corpus d'entraînement apparaissaient dans la même phrase, menant VarIDE à « se mélanger les pinceaux ». Par exemple, dans la phrase « (…) des mesures seront nécessaires pour assurer une évaluation adéquate des progrès réalisés et des obstacles *rencontrés* avant et pendant l'Année, afin d'en *assurer* le *succès* et de *prendre* les *mesures* de *suivi* voulues » (UN:3455), l'outil a fautivement annoté les collocations *rencontrer*|*succès* (tandis que *rencontrer* se rapporte à *obstacles* et *succès* à *assurer*) et *assurer*|*suivi* (tandis que *assurer* se rapporte à *succès* et *suivi* complète le nom *mesures* le précédant). La dernière collocation *mesure*|*prendre* était la seule bien identifiée dans ce segment. Ce type d'erreurs est de loin le plus fréquent.

8.2.3.5. Dépendance syntaxique différente de celle attendue

Parfois, les composants d'une collocation ne se trouvent pas dans la configuration dans laquelle elle devrait être trouvée pour bel et bien former une collocation. Par exemple, la collocation *mesure*|*prendre* ne fonctionne pas si *mesure* est sujet de *prendre*. Dans la phrase « (…) les *mesures* actuelles *prennent* une direction beaucoup plus dérangeante. » (GV:13045), l'outil a annoté fautivement *mesure*|*prendre*.

8.2.4. Résultats et interprétations

Malgré les erreurs inévitables décrites ci-dessus, le fait que des erreurs dans la transformation au format `conllu` aient un peu réduit les performances de VarIDE, et le fait que l'outil ne puisse pas identifier les expressions qu'il n'a pas apprises pendant la phase d'entraînement, les résultats de l'annotation automatique, présentés dans le tableau suivant, sont satisfaisants.

Corpus	Précision	Rappel	F-mesure
GV	1342 / 1577 = **85,10**	1342 / 1721 = **77,98**	**81,38**
TED	684 / 792 = **86,36**	684 / 768 = **89,06**	**87,69**
UN	2417 / 2829 = **85,44**	2417 / 2875 = **84,07**	**84,75**
WM	1504 / 1770 = **84,67**	1504 / 1847 = **81,43**	**83,16**
Tous	5947 / 6968 = **85,35**	5947 / 7211 = **82,47**	**83,88**

Tableau 13 : Evaluation de l'annotation automatique du corpus parallèle après correction

On constate que les résultats sont très équilibrés, notamment en ce qui concerne la précision qui ne diffère quasiment pas entre les sous-corpus. Le rappel diverge plus grandement entre les sous-corpus, mais pas de manière drastique. Avec une F-mesure globale de 83,88, l'annotation automatique semble être un succès encourageant. Le script d'évaluation de VarIDE offre quelques informations plus détaillées intéressantes : sur les 7211 annotations finales (585 patrons uniques) du corpus annoté, seules 749 d'entre elles (environ 10%) apparaissent de manière continue. De plus, 81% des collocations finales sont des variantes de celles du corpus d'entraînement.

De prime abord, on constate également que même après correction, l'écart du nombre d'annotations par sous-corpus est toujours aussi important. L'hypothèse que nous pouvons formuler, compte tenu des sources de chacun de ces corpus, c'est qu'au plus le langage qui est utilisé doit répondre à une norme stricte, au plus l'usage de collocations verbales est important. En effet, les rapports parlementaires du corpus UN sont plus surveillés et normés que ne sont les conférences TED. En outre, la dimension diamésique semble également importante : la transmission orale et « grand public » d'une conférence TED est bien plus relâchée que ne le sont les codes de rédaction d'articles de journaux, d'encyclopédies ou une fois de plus, de ceux des rapports parlementaires. Rappelons également que le corpus TED est le fruit de traductions de volontaires et n'est peut-être pas aussi qualitative que celle du corpus UN.

Par ailleurs, même si quelques collocations verbales reviennent très fréquemment indifféremment d'un genre à l'autre *(poser|question* ou *jouer|rôle)*, d'autres sont plus spécifiques à un genre donné et témoignent des thématiques abordés. Regardons tout d'abord les 10 collocations les plus fréquentes sur l'ensemble du corpus :

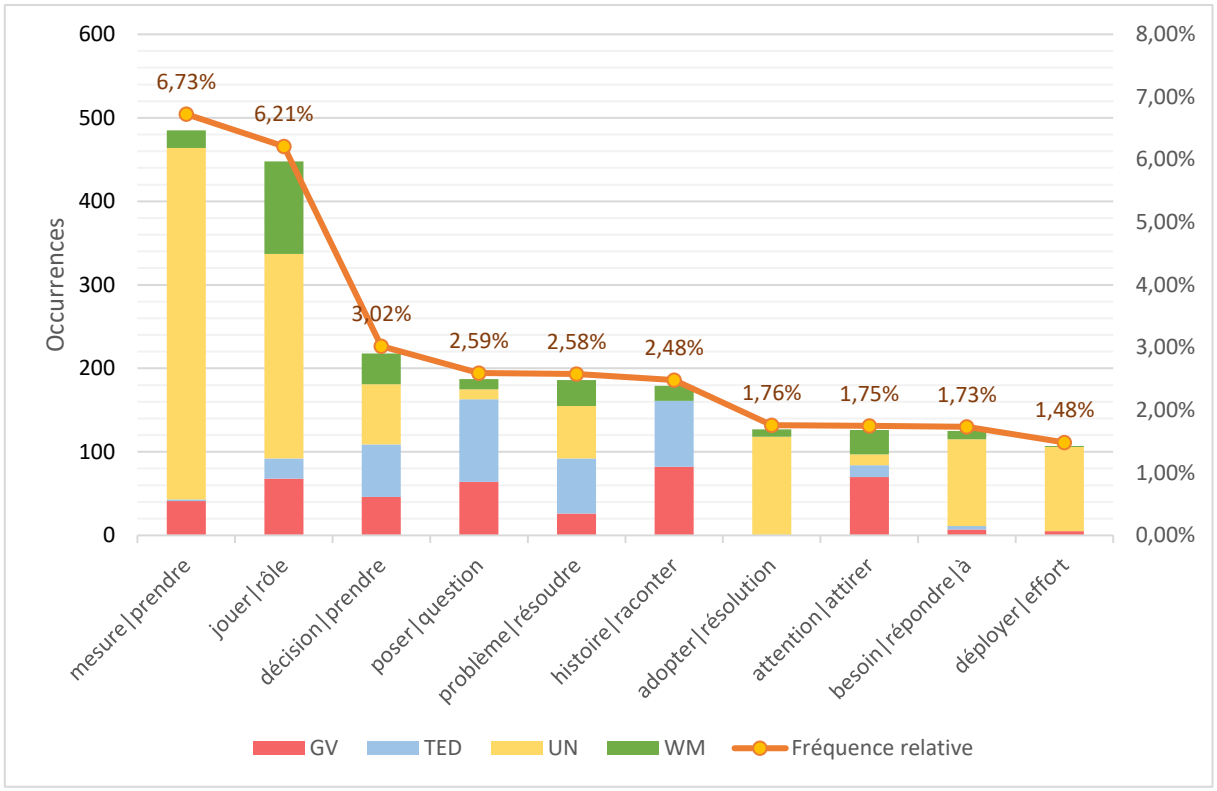

Figure 3 : Collocations les plus fréquentes et proportion par sous-corpus

On constate que deux collocations se détachent très largement du reste *(mesure|prendre* et *jouer|rôle)* avec respectivement 485 (soit une fréquence relative de 6,73%) et 448 occurrences (fréquence relative de 6,21%). En comparant ces résultats à ceux de chaque sous-corpus, nous serons plus à même de comprendre ceux-ci. Les 10 collocations les plus fréquentes pour le corpus GV sont les suivantes :

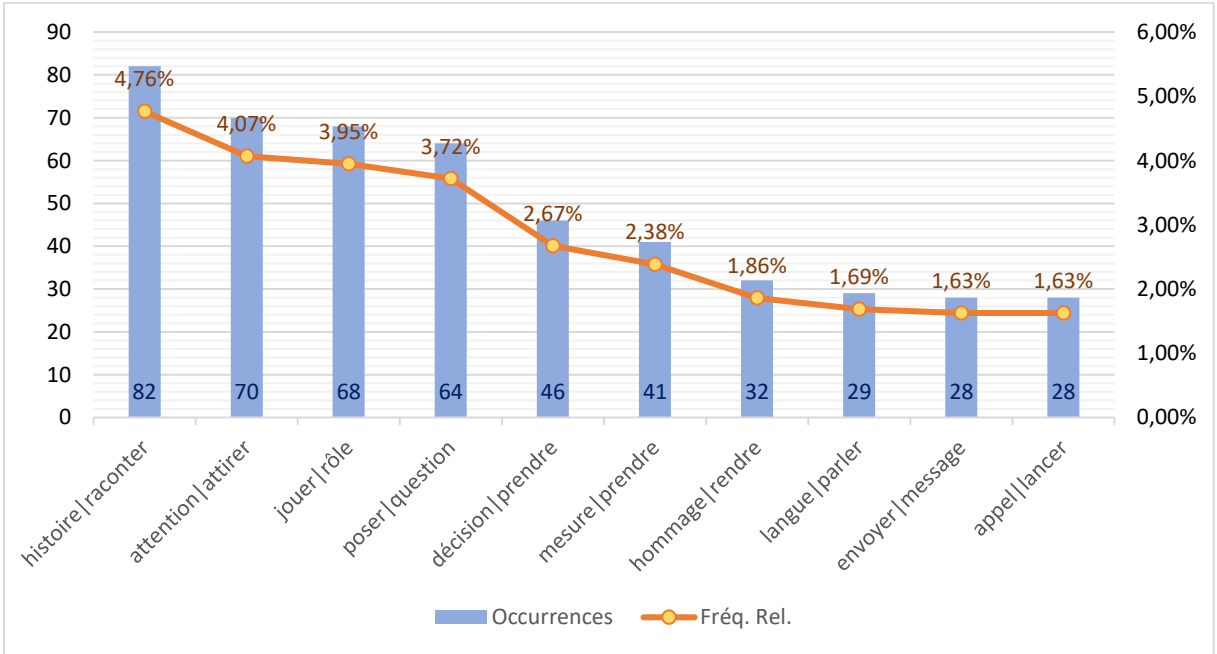

Figure 4 : Collocations les plus fréquentes (sous-corpus Global Voices)

Outre les collocations *jouer|rôle* (68 occurrences) et *poser|question* (64 occurrences) très présentes, les 2 expressions les plus répétées sont *histoire|raconter* (82 occurrences) et *attentation|attirer* (70 occurrences). Si l'on examine les contextes dans lesquels elles apparaissent, on en trouve rapidement la cause : nos extraits du corpus GV reprennent largement les messages publiés dans des vidéos ou les billets de blogs postés par des activistes. À travers ces médias, ce sont leur histoire ou l'histoire d'autres personnes qui y sont racontées et les messages qui y sont relayés sont envoyés dans le but d'attirer l'attention des internautes, souvent du monde occidental.

Plus que les collocations les plus fréquentes, ce sont peut-être les collocations spécifiques à ce sous-corpus ayant un certain nombre d'occurrences qui sont les plus parlantes. En effet, 6 collocations ont plus de 5 occurrences[19] et sont spécifiques à nos extraits GV : *scander|slogan* (10 occurrences), *liberté|restreindre* (8 occurrences), *encourir|peine* (7 occurrences), *peine|purger* (6 occurrences), *censure|contourner* (6 occurrences) et *jeûne|rompre* (5 occurrences). En se contentant de ces expressions, on peut confirmer l'hypothèse que GV est un journal relativement engagé, traitant régulièrement de sujets relatifs aux manifestations populaires et au manque de liberté(s), bien souvent ayant lieu dans le monde arabo-musulman. Après examen des contextes dans lesquels apparaissent ces expressions, cette hypothèse est confirmée. Penchons-nous maintenant sur les résultats du corpus TED :

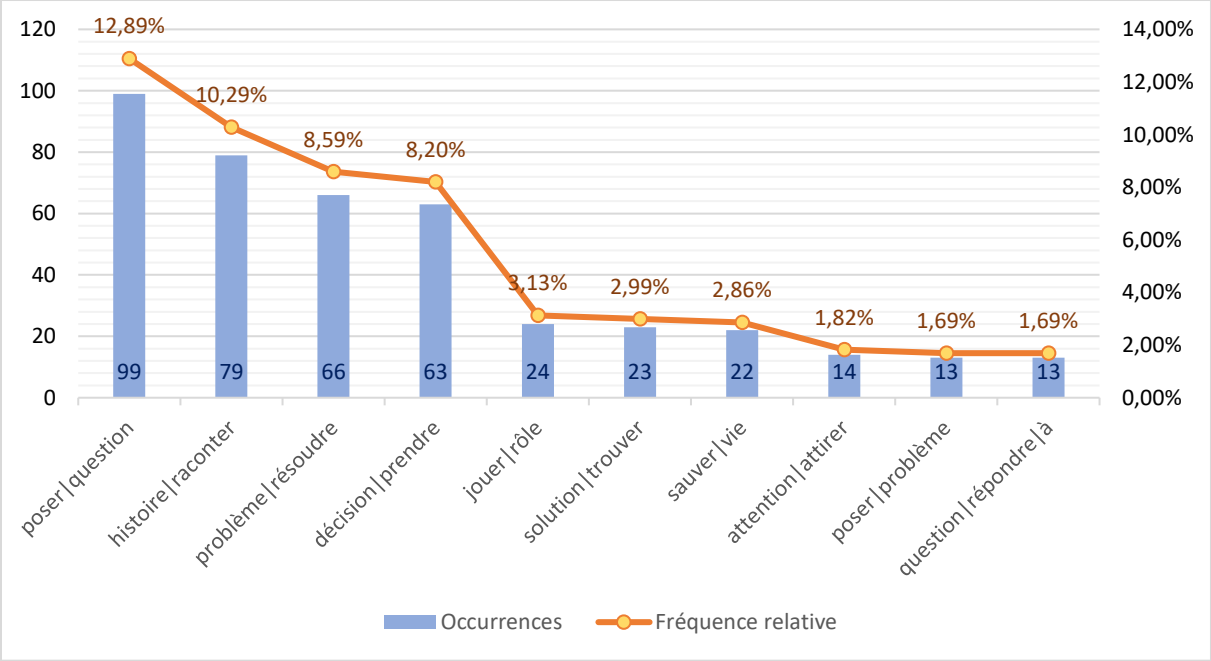

Figure 5 : Collocations les plus fréquentes (sous-corpus TED 2020)

On constate que deux collocations se détachent relativement largement : *poser|question* et *histoire|raconter*. Les conférenciers des TED Talks ont souvent recours à des stratégies pour capter l'attention des spectateurs, comme des amorces du type « La dernière question que nous aimerions poser est (…) » (TED:4641) ou encore « Je voudrais aujourd'hui vous raconter une histoire sur (…) » (TED:4910). Contrairement au premier sous-corpus, les collocations

[19] Nous avons choisi ce seuil de 5 occurrences plutôt que 10 car très peu de collocations spécifiques à un corpus apparaissent à cette fréquence. Selon le corpus, ce seuil de 10 n'est pas du tout atteint.

spécifiques à TED 2020 sont rares, et exceptée l'expression *blague|raconter* (3 occurrences), rien ne nous donne d'indices particuliers quant aux sujets prédominants ou à la dimension orale des conférences. Le sous-corpus UN a beaucoup plus de choses à nous dire :

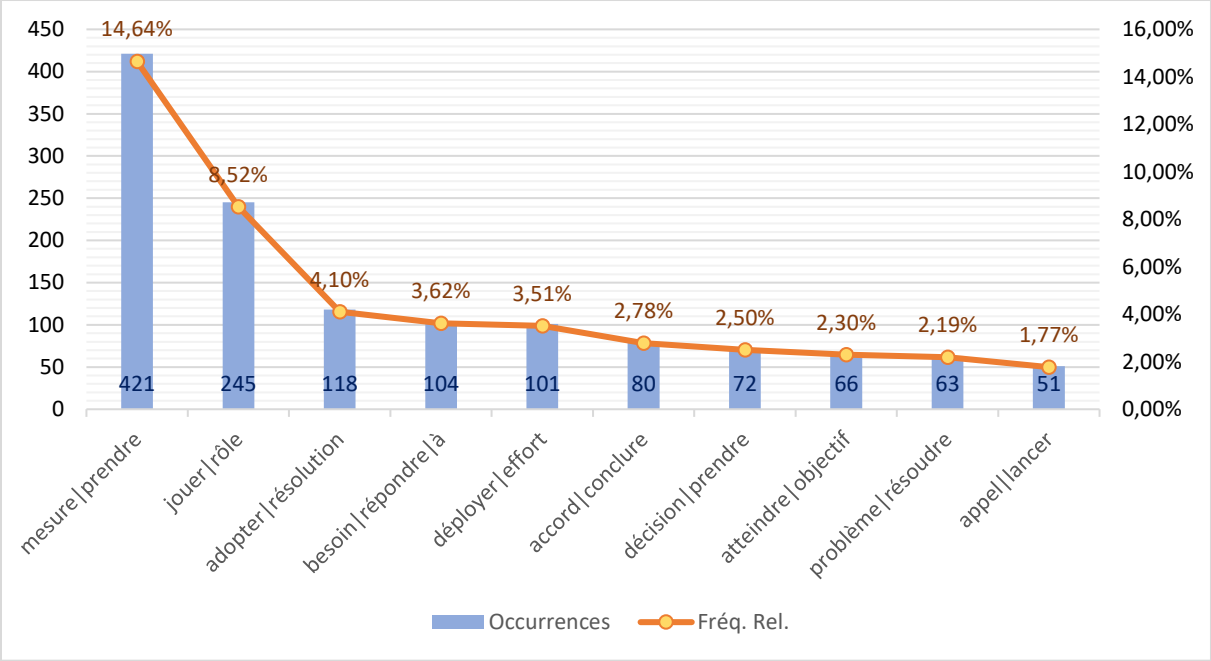

Figure 6 : Collocations les plus fréquentes (sous-corpus United Nations)

Après *jouer|rôle* approchant les 250 occurrences, c'est, de très loin, la collocation *mesure|prendre* qui domine toutes les autres dans ce sous-corpus. C'est peu surprenant quand on songe au contenu des rapports parlementaires : une partie très importante de ces documents traitent des mesures que tel pays ou telle organisation a mis, met ou mettra en œuvre. La langue parlementaire étant plutôt rigoureuse, il y a peu de variation lexicale pour exprimer une idée similaire. Ainsi, on constate que les 5 collocations les plus fréquentes apparaissent plus de 100 fois chacune dans notre sous-corpus, qui n'est pas pour autant très volumineux.

À l'instar des collocations les plus fréquentes, celles qui sont spécifiques à ce sous-corpus nous renseignent tout aussi bien sur les thématiques de prédilection abordées lors des sessions parlementaires. En effet, autant *adopter|résolution* (118 occurrences) que *fixer|modalité* (5 occurrences) nous donnent un indice sur la dimension juridico-législative des documents ; de la même manière, autant *accord|conclure* (80 occurrences) que *accorder|crédit* (7 occurrences) nous donnent un indice sur la dimension économique des documents ; enfin, autant *déployer|effort* (101 occurrences) que *causer|préjudice* (8 occurrences) nous donnent un indice sur la dimension politico-collaborative des documents. Penchons-nous finalement sur le sous-corpus WM :

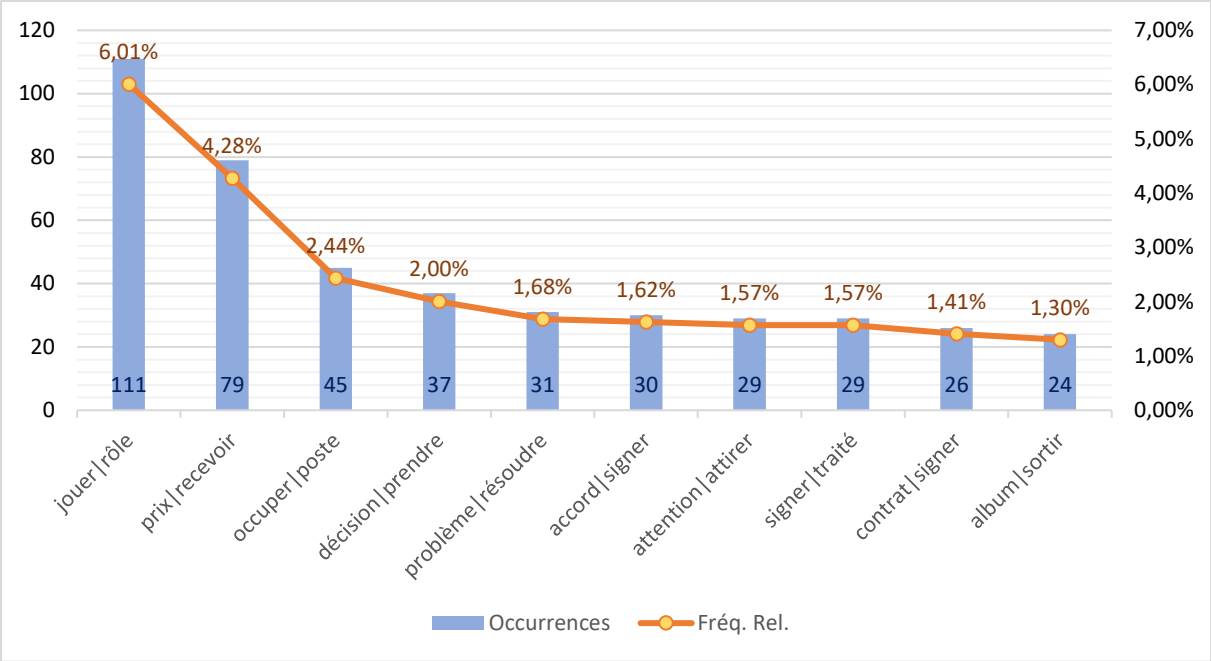

Figure 7 : Collocations les plus fréquentes (sous-corpus WikiMatrix)

Exceptées les deux premières collocations, le relatif équilibre du reste des expressions témoigne de la diversité des sujets abordés par notre échantillon du corpus WM. On constate qu'un nombre non négligeable des collocations les plus fréquentes ont sans doute un rapport avec des évènements historiques *(accord|signer* 30 occurrences, *signer|traité* 29 occurrences, *contrat|signer* 26 occurrences*)*. Cependant, ce sont surtout les collocations spécifiques (dont le nombre s'élève à 88, contre 15 pour le sous-corpus TED) à ce sous-corpus qui montrent que des grands thèmes comme le sport (*jouer|match* 13 occurrences, *but|inscrire 7* occurrences, *record|établir* 6 occurrences) ou encore la culture (*album|sortir* 24 occurrences, *animer|émission* 5 occurrences) sont également représentés de manière importante.

8.2.5. Observations finales

De ce que nous avons abordé dans cette section concernant l'évaluation de l'annotation automatique, nous pouvons retenir les conclusions suivantes. Tout d'abord, il semblerait que le nombre de collocations utilisées soit plus important selon qu'il s'agisse d'un canal de communication écrit ou oral. En effet, la dimension diamésique semble être un facteur important quand on compare les chiffres des conférences TED (768 annotations, 176 patrons uniques et seulement 15 patrons spécifiques) à ceux des rapports parlementaires du corpus UN (2875 annotations, 252 patrons uniques et 32 patrons spécifiques) ou des articles de journaux de GV (1721 annotations, 384 patrons uniques et 82 patrons spécifiques).

De plus, il semblerait également que le degré de normalisation de la langue utilisée influe énormément sur l'usage des collocations verbales. En effet, le corpus onusien équivaut à 40,60% des annotations totales du corpus parallèle, soit quasiment 4 fois plus que TED (11,37%) et un peu moins que 2 fois plus que GV (22,63%) et WM (25,40%). En d'autres termes, seulement 2,70% des phrases de TED contiennent une collocation, tandis que ce pourcentage monte à 13,53% pour UN (soit environ le double que pour les deux autres sous-corpus). Pour autant, le nombre de patrons uniques et surtout de patrons spécifiques n'est pas

particulièrement élevé. En revanche, le nombre d'occurrences des collocations fréquentes est bien plus élevé. Sur ce dernier point, un parallèle peut être fait entre la langue juridico-parlementaire très normée de l'ONU et la langue journalistique de Global Voices, dont le degré de normalisation est un peu moindre. En outre, nous pourrons retenir que c'est la combinaison des collocations les plus fréquentes et les collocations spécifiques à un genre qui nous éclairent sur les différents thèmes abordés.

Enfin, en ce qui concerne les performances de VarIDE spécifiquement, avec une précision de 85,35, un rappel de 82,47 et une F-mesure de près de 83,88, nous pouvons conclure que l'outil fonctionne bien pour l'identification des collocations verbales en français. Nous avons établi une typologie des sources des erreurs dans l'annotation automatique. Elles peuvent être dues à des erreurs préalables (p. ex. mauvaise lemmatisation du texte brut), à des différences de traitement de la tokenisation entre le corpus d'entraînement et de test (p. ex. *auquel* qui n'est pas détaillé en *à + lequel*), à des considérations d'ordre sémantique (cas de polysémie), à la présence concomitante de plusieurs composants de collocations mais n'ayant pas relation de dépendance syntaxique dans la phrase donnée (notamment dans les longs segments), ou encore à des relations de dépendance syntaxique différentes et incompatibles avec celles attendues. Malgré ça, nous considérons que l'annotation automatique est un succès et pourrait être encore améliorée avec un corpus d'entraînement plus conséquent ou avec une source supplémentaire de patrons de collocations (p. ex. un dictionnaire).

9. CONCLUSION

Dans cette section, nous avons balayé toutes les étapes nécessaires de la méthodologie pour mener à bien le processus d'annotation automatique des collocations verbales pour la partie française de notre corpus. Dans un premier temps, nous avons présenté les ressources exploitées, à savoir les 4 sous-corpus de notre corpus parallèle multi-genre, ainsi que l'outil d'annotation automatique que nous avons utilisé, VarIDE, développé pour l'identification automatique des EP verbales dans le cadre de la Shared Task 1.1 du projet international PARSEME. Nous avons également détaillé la manière dont nous avons rédigé notre guide d'annotation en nous inspirant largement des projets SimpleApprenant pour les tests linguistiques utiles à la reconnaissance des collocations verbales (passage à la diathèse passive, changement de déterminant, ajout de modifieurs, mesure d'association) et PARSEME pour leur modèle d'arbre de décision, aiguillant l'annotateur humain quant à la possible annotation d'un candidat-collocation.

Dans un second temps, nous avons présenté le processus d'annotation semi-automatique de notre corpus français, de la préparation des données à leur évaluation. Pour la phase de préparation, nous avons détaillé comment nous avons adapté le corpus d'entraînement fourni lors de la Shared Task 1.1 de PARSEME à notre projet, avant de discuter de la façon dont nous avons échantillonné les corpus parallèles complets de sorte à obtenir ce que nous avons qualifié de « tritextes », c'est-à-dire des triplets de phrases alignées. Nous avons ensuite convertis ces derniers au format `cupt`, un format `conllu` étendu contenant les annotations des EP verbales, obligatoire pour utiliser VarIDE. Pour la phase d'évaluation, nous avons détaillé la procédure que nous avons exécutée, de la correction de l'annotation automatique à l'augmentation manuelle de cette dernière, tout en détaillant les chiffres obtenus. Nous avons dressé une typologie des erreurs commises par l'outil et leurs possibles explications. Enfin, nous avons présenté nos résultats en tâchant de les interpréter au mieux, en comparant chacun des sous-corpus tant sur le plan quantitatif que qualitatif. De notre point de vue, avec une F-mesure de 83,88, nous estimons que VarIDE fonctionne bien pour identifier des collocations verbales en français.

Dans la prochaine section principale, nous traiterons de la dernière phase de notre projet : la projection des annotations françaises vers les corpus anglais et arabe. Nous concluons avec l'analyse linguistique contrastive de l'usage des collocations verbo-nominales entre les langues et entre les genres.

III. METHODOLOGIE : PROJECTION DES ANNOTATIONS

10. PRESENTATION DES RESSOURCES ET OUTILS UTILISES

Projeter des annotations nécessite d'avoir d'une part des alignements lexicaux du type table de traduction d'une langue-source vers une langue-cible et d'autre part un outil permettant d'exploiter ces alignements pour transférer les annotations d'un document vers un autre. Pour notre projet, à ce stade d'avancement, nous disposons d'un corpus français au format `cupt` entièrement annoté semi-automatiquement en collocations verbo-nominales. L'objectif étant d'enrichir les corpus parallèles anglais et arabe, nous avions besoin dans un premier temps de générer des tables de traduction bilingues à partir de grands volumes de textes alignés. Pour ce faire, nous avons utilisé GIZA++, un outil d'alignement lexical développé par Och et Ney (2003). Une fois les tables de traduction générées, nous avons tiré profit de ZAP (Akbik & Vollgraf, 2018), un outil développé en Java permettant de transférer des annotations à plusieurs niveaux linguistiques d'un document source vers un document cible. Dans cette section, nous nous attacherons à les décrire.

10.1. GIZA++

GIZA++ est un outil d'alignement lexical en C++ qui demeure une référence en ce qui concerne ce type d'outils, malgré la vingtaine d'années qui s'est écoulée depuis son développement par les équipes d'IBM. Très robuste, il permet de créer des tables de traduction bilingues à partir de corpus parallèles.

GIZA++ prend en entrée deux textes : un pour la langue-source et un autre pour la langue-cible. Les deux textes doivent être alignés au niveau de la phrase pour pouvoir calculer les alignements au niveau lexical. Dans un premier temps, ce sont les fichiers liés au vocabulaire (`.vcb`) et aux phrases (`.snt`) des textes qui sont générés.

```
./plain2snt.out langue-source.tok langue-cible.tok
```

Les fichiers de vocabulaire contiennent une liste (identifiant unique, chaîne de caractères, nombre d'occurrences) pour chaque mot. Les fichiers de phrases contiennent une liste de trois lignes : le nombre de fois que cette paire de phrases apparaît, la phrase source (avec chaque token remplacé par son identifiant unique), et la phrase cible dans le même format.

Dans un deuxième temps, ce sont des fichiers correspondant aux cooccurrences qui sont générés, à partir des fichiers générés précédemment.

```
./snt2cooc.out langue-source.vcb langue-cible.vcb langue-source_langue-
cible.snt > langue-source_langue-cible.cooc

./snt2cooc.out langue-cible.vcb langue-source.vcb langue-cible_langue-
source.snt > langue-cible_langue-source.cooc
```

Dans un troisième temps, ce sont des fichiers classes qui sont générés à partir des fichiers de vocabulaire et des textes tokenisés.

```
./mkcls -plangue-source.tok -Vlangue-source.vcb.classes

./mkcls -plangue-cible.tok -Vlangue-cible.vcb.classes
```

Ces fichiers contiennent une liste de tous les mots du corpus classés par ordre alphabétique, ponctuation incluse, ainsi que la fréquence correspondant à chaque mot. D'autres fichiers sont également générés et contiennent une liste de toutes les fréquences et un ensemble de mots correspondant à chacune de ces fréquences. Finalement, GIZA++ effectue les alignements.

```
./GIZA++ -S langue-source.vcb -T langue-cible.vcb -C langue-source_langue-
cible.snt -CoocurrenceFile langue-source_langue-cible.cooc -p0 0.98 -o
giza-langue-source-langue-cible

./GIZA++ -S langue-cible.vcb -T langue-source.vcb -C langue-cible_langue-
source.snt -CoocurrenceFile langue-cible_langue-source.cooc -p0 0.98 -o
giza-langue-cible-langue-source
```

Plusieurs types de fichiers sont générés. D'une part, il y a les fichiers `*.actual.ti.final` qui contiennent les tables de traduction inverses avec la probabilité de traduction au niveau lexical entraînées par le modèle. La probabilité de traduction lexicale *t(e|f)* correspond à la probabilité qu'un mot *f* dans la langue-source soit traduit par le mot *e* dans la langue-cible. Etant donné qu'il s'agit de tables inverses, le fichier contient la probabilité de traduction lexicale *t(f|e)*. Voici un extrait :

```
exceptions différents 0.0889517
exceptions donneur 2.9108e-05
exceptions exception 0.0876911
exceptions exceptions 0.809032
exceptions idiosyncratiques 0.0142956
```

La somme des probabilités correspond bien à 1.0. D'autre part, il y a les fichiers `*.AA3.final`. Ces derniers contiennent les alignements Viterbi, qui correspondent aux alignements les plus probables (ceux qui ont la probabilité d'alignement maximum). Voici un extrait de ce fichier pour une paire de phrases :

```
# Sentence pair (22834) source length 7 target length 8 alignment score :
1.25846e-05
This is a photograph of the object .
NULL ({ }) Voici ({ 1 2 }) une ({ 3 }) photographie ({ 4 }) de ({ 5 }) l'
({ 6 }) objet ({ 7 }) . ({ 8 })
```

La première ligne renseigne sur la longueur en nombre de mots de la phrase en langue-source (ici le français) et en langue-cible (ici l'anglais), accompagnée du score d'alignement Viterbi déjà mentionné. La deuxième ligne correspond à la phrase en langue-cible, tandis que la troisième est la phrase en langue-source annotée avec les informations d'alignement. Chaque mot source est annoté avec l'ensemble des indices des mots cibles qui sont alignés avec ledit mot source. Le premier `NULL ({ })` signifie que tous les mots cibles ont été alignés avec un mot source, tandis que `Voici ({ 1 2 })` signifie que les mots cibles *This* et *is* sont alignés avec *Voici*.

Grâce à cet outil toujours aussi robuste, il est possible de créer des tables de traduction bilingues. Ces dernières sont utiles pour de nombreuses tâches. Pour nous, elles seront utiles

pour projeter les annotations françaises vers le corpus anglais, puis les annotations anglaises vers le corpus arabe. Un autre outil nous permettra d'exploiter ces tables de traduction pour la projection : ZAP (Akbik & Vollgraf, 2018). Dans la section suivante, nous nous attachons à en décrire son fonctionnement.

10.2. ZAP

ZAP est un outil développé relativement récemment en Java. Le problème auquel ses développeurs souhaitent trouver une solution est le suivant : obtenir des annotations linguistiques de qualité pour des langues moyennement ou peu dotées, malgré la rareté des données dont elles jouissent. La proposition faite est de tirer parti d'alignements lexicaux obtenus entre une langue bien dotée et une langue moins bien dotée, de sorte à pouvoir étiqueter en sortie le document en langue-cible. En pratique, l'outil propose actuellement comme langue-source uniquement l'anglais, et comme langues-cibles le français, l'allemand, l'espagnol et dans une moindre mesure le chinois.

Le principe est simple. Il faut tout d'abord disposer de tables de traduction (ZAP en dispose par défaut pour l'anglais vers le français, l'allemand et l'espagnol). Ensuite, il faut fournir en entrée un document source (un objet `Sentence`) soit au format texte (qu'il est alors nécessaire d'analyser avec un objet `PipelineWrapper`), soit au format CoNLL. Il faut également créer un objet `Sentence` pour le document en langue-cible (tokenisé ou au format CoNLL).

```
PipelineWrapper pipeline = new PipelineWrapper(Language.ENGLISH);

Sentence sourceSentence = pipeline.parse("The cat eats cheese.");

Sentence targetSentence = Sentence.fromTokenized("Le chat mange du fromage
.");
```

Il est ensuite nécessaire d'utiliser l'aligneur heuristique de ZAP en initialisant un objet `HeuristicAligner` avec la langue-cible choisie. C'est grâce à l'aligneur heuristique que l'alignement entre les tokens du document source et du document cible sera effectué. Pour ce faire, il faut initialiser un objet `BiSentence`, qui va effectivement se charger d'aligner les deux phrases fournies en entrée. Il est également possible de récupérer les alignements réalisés.

```
HeuristicAligner aligner = HeuristicAligner.getInstance(Language.FRENCH);

BiSentence biSentence = new BiSentence(sourceSentence, targetSentence);

biSentence.align(aligner);

System.out.println(biSentence);

System.out.println(biSentence.aligments);
```

Pour le présent exemple, la sortie serait la suivante :

```
        The  cat  eats  cheese  .
Le       X
chat           X
mange               X
du
fromage                    X
.

{4 cheese={5 fromage=1.0}, 2 cat={2 chat=1.0}, 3 eats={3 mange=1.0}, 1
The={1 Le=1.0}}
```

Enfin, la projection des annotations linguistiques de la `Sentence` source (obtenues avec `pipeline.parse` dans cet exemple) peut se faire en appelant la classe `AnnotationTransfer`. Il est alors possible de récupérer les projections au format CoNLL-U.

```
new AnnotationTransfer().transfer(biSentence);

System.out.println(biSentence.getSentenceTL().toConllU());
```

Dont la sortie est la suivante :

```
1    Le         _    DET    DT     _    2    det      _    _    _    _         _
2    chat       _    NOUN   NN     _    3    nsubj    _    _    _    _         A0
3    mange      _    VERB   VBZ    _    0    _        _    _    Y    eat.01    _
4    du         _    _      _      _    0    _        _    _    _    _         _
5    fromage    _    NOUN   NN     _    3    dobj     _    _    _    _         A1
6    .          _    _      _      _    0    _        _    _    _    _         _
```

Les étiquettes grammaticales, les dépendances syntaxiques et les cadres sémantiques ont bien été projetés.

ZAP fournit également une interface web, `TheProjectorUI`, pour permettre de visualiser les différentes projections réalisées. La figure suivante est la visualisation obtenue pour la projection des annotations de l'objet `BiSentence` précédent :

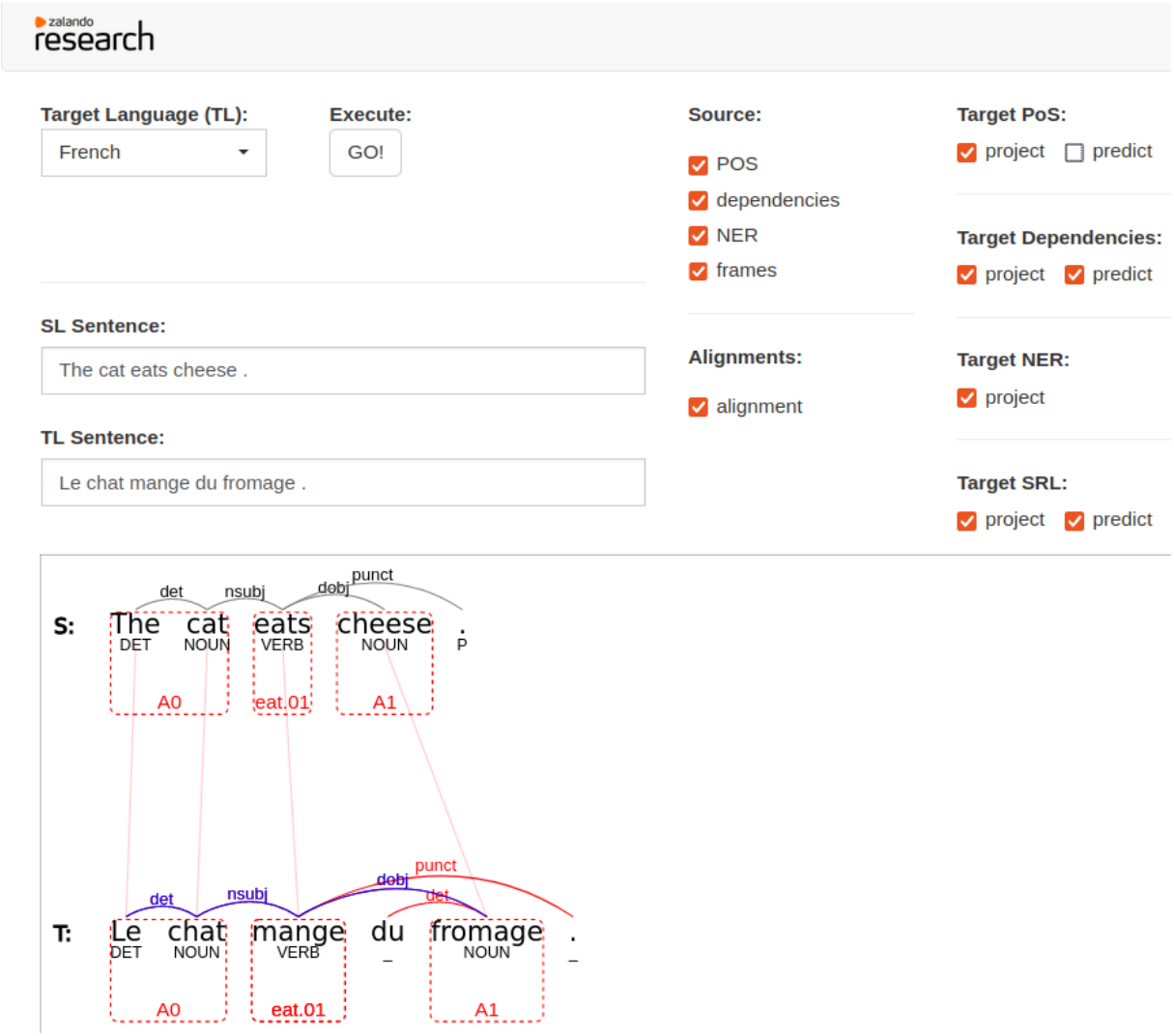

Figure 8 : Exemple de visualisation de projection avec TheProjectorUI

De la phrase-source à la phrase-cible, les termes alignés sont reliés par un trait plein rouge. Les différents constituants, notés `A0` et `A1` ici, sont encadrés en pointillés rouges. Le verbe l'est également, et le cadre sémantique `eat.01` est bien projeté. Les relations de dépendance de la phrase-cible notées en bleu sont les relations projetées, tandis que celles en rouge sont des relations prédites par le modèle. On constate par exemple que le déterminant amalgamé *du*, dont l'équivalent dans la phrase-source serait le déterminant zéro entre *eats* et *cheese*, a reçu une relation de dépendance `det` avec le terme *fromage*.

Pour notre projet, la projection des informations morphosyntaxiques importe peu car nos « tritextes » sont déjà au format CoNLL-U étendu et les cadres sémantiques ne nous sont d'aucune utilité. En outre, nous n'utiliserons pas l'interface graphique. De fait, ce qui nous a intéressé plus particulièrement avec ZAP est son aligneur heuristique, que nous avons exploité pour procéder à notre propre projection. Cette étape se décline en deux temps, avec la projection du français vers l'anglais tout d'abord, puis de l'anglais vers l'arabe ensuite. Les sections suivantes s'attachent à en décrire la méthodologie.

11. PROJECTION DES ANNOTATIONS (FRANÇAIS-ANGLAIS)

11.1. Projection automatique

La projection automatique, au départ, devait suivre la méthodologie suivante : 1) obtenir une table de traduction français-anglais avec GIZA++, 2) adapter le code-source de ZAP pour pouvoir projeter une colonne supplémentaire (celle contenant nos annotations des collocations), et 3) écrire et exécuter un script pour projeter nos résultats. Cependant, nous nous sommes heurté à un problème que nous n'avions pas anticipé et avons dû modifier la méthodologie à employer. Dans cette section, nous décrivons dans un premier temps les problèmes rencontrés et la solution envisagée pour y remédier. Ensuite, nous discutons de la projection automatique en elle-même et en évaluons les résultats, avant de dresser une typologie des différences observées. Enfin, nous analysons et interprétons les résultats obtenus après correction et augmentation de notre corpus anglais.

11.1.1. ZAP : problèmes rencontrés et solutions envisagées

ZAP est un outil très complet, mais nous n'avions pas réalisé que la version distribuée ne pourrait pas satisfaire entièrement notre projet. En effet, il est possible de projeter des annotations uniquement depuis l'anglais vers une autre des langues supportées. La projection du français vers l'anglais était alors impossible en l'état sans développer une tout autre version de l'outil. Les contraintes temporelles liées à notre projet ne nous permettaient donc évidemment pas de faire cela, d'autant plus que notre projection devait non seulement se faire vers l'anglais, mais également vers l'arabe, langue qui n'est actuellement pas gérée par ZAP.

Ainsi, nous avons dû modifier la méthodologie que nous souhaitions employer. Plutôt que de passer par ZAP pour projeter nos annotations du début à la fin du processus, nous avons décidé de ne l'utiliser que pour générer et récupérer les alignements phrase par phrase de nos deux corpus. En effet, qu'il s'agisse d'alignements anglais-français ou français-anglais, les résultats seraient sensiblement les mêmes. Dans un second temps, nous avons envisagé de créer notre propre programme qui nous permettrait tout d'abord d'enrichir les alignements de ZAP avec nos annotations des collocations, pour ensuite les projeter dans notre corpus anglais. Nous décrivons cette nouvelle méthodologie dans la sous-section suivante.

11.1.1. Génération et enrichissement des alignements lexicaux

ZAP dispose d'une table de traduction anglais-français générée avec Berkeley Aligner (Liang et al., 2006) tout à fait satisfaisante. Dès lors, nous n'avons pas jugé nécessaire d'en créer une similaire, d'autant plus que leur document a été réalisé grâce à l'alignement lexical de corpus parallèles de la plateforme OPUS, stratégie que nous comptions également employer[20]. Ces alignements, comme nous l'avons montré dans la section 10.2 décrivant le fonctionnement de ZAP, prennent la forme suivante (exemple de la phrase 195 du sous-corpus GV) :

[20] Nous utilisons cependant GIZA++ pour créer une table de traduction anglais-arabe (voir section 12.1.1).

```
{18  Dubai={20 Dubaï=1.0}, 7 blogger={6    blogueur=1.0}, 20 reputation={18
     réputation=1.0}, 13     n't={16   pas=1.0}, 10   the={17 la=1.0}, 2
     the={5    le=1.0}, 8    Ammaro={9 Ammaro=1.0}, 17 tarnish={15
ternira=1.0}, 11   case={13  affaire=1.0}, 4    hand={3   côté=1.0}}
```

Soit une liste dont chaque élément décrit un alignement lexical contenant, dans chacune des deux phrases alignées, l'identifiant et la forme du token anglais, puis, après le premier signe « égal », l'identifiant et la forme du token français, eux-mêmes suivis du score d'alignement après le second signe « égal ». La stratégie pour laquelle nous avons optée consistait ensuite à transformer et enrichir ces alignements avec nos propres annotations, de sorte à pouvoir ensuite les projeter.

Nous avons donc écrit un programme Python qui transforme et enrichit les alignements de ZAP. Pour chaque sous-corpus, nous avons récupéré les alignements générés par un court programme exploitant ZAP, auxquels nous avons associé l'identifiant de chaque phrase. Puis, notre programme Python transforme chaque alignement lexical de sorte qu'il prenne la forme suivante :

```
[identifiant_phrase, identifiant_token_anglais, forme_token_anglais,
identifiant_token_français, forme_token_français, score]
```

Chaque liste contenant un alignement lexical est inséré dans une liste de listes correspondant à l'ensemble des alignements pour une phrase du corpus. À la suite de cette transformation, notre programme parcourt le corpus français au format `cupt` entièrement annoté en collocations et lorsque les informations d'un token français (son identifiant et sa forme) dans l'alignement produit correspond à un token de la phrase en question dans le corpus, l'annotation dudit token est ajoutée aux éléments de la liste, qu'il s'agisse de l'astérisque signalant que le token ne fait pas partie d'une annotation ou des annotations des collocations à proprement parler. Ainsi, pour l'exemple mentionné précédemment, l'alignement enrichi prend la forme suivante :

```
[['195', '18', 'Dubai', '20', 'Dubaï', '1.0', '*'], ['195', '7', 'blogger',
'6', 'blogueur', '1.0', '*'], ['195', '20', 'reputation', '18',
'réputation', '1.0', '1'], ['195', '13', "n't", '16', 'pas', '1.0', '*'],
['195', '10', 'the', '17', 'la', '1.0', '*'], ['195', '2', 'the', '5',
'le', '1.0', '*'], ['195', '8', 'Ammaro', '9', 'Ammaro', '1.0', '*'],
['195', '17', 'tarnish', '15', 'ternira', '1.0', '1:COLL'], ['195', '11',
'case', '13', 'affaire', '1.0', '*'], ['195', '4', 'hand', '3', 'côté',
'1.0', '*']]
```

On peut remarquer que les deux tokens en gras contiennent une annotation car ils forment la collocation (ternir,réputation) dans le corpus français. Une fois que tous nos alignements ont été enrichis, notre programme parcourt le corpus anglais au format `cupt`, dont toutes les annotations sont au départ des astérisques. Pour chaque phrase, lorsque les informations d'un token anglais (son identifiant et sa forme) correspondent à un token dans les alignements enrichis, la 11^{e} colonne contenant les annotations est modifiée par le dernier élément de la liste enrichie, c'est-à-dire l'annotation. Toujours pour faire suite à notre exemple, la collocation (ternir,réputation) a bien été projetée dans le corpus anglais, avec l'équivalent direct (tarnish,reputation). Ci-après, nous détaillons les fichiers `cupt` français puis anglais, dont les tokens annotés sont en gras.

```
# source_sent_id = *** *** GV/GV.tri.fr::195
# text = De son côté, le blogueur de Bahrain Ammaro espère que l'affaire ne
ternira pas la réputation de Dubaï.
1    De   de   ADP  _    _    3    case _    _    *
2    son  son  DET  _
     Number=Sing|Number[psor]=Sing|Person[psor]=3|PronType=Prs    3    det
     _    _    *
3    côté côté NOUN _    Gender=Masc|Number=Sing 10   obl:mod  _    _    *
4    ,    ,    PUNCT  _    _    3    punct  _    _    *
5    le   le   DET  _    Definite=Def|Gender=Masc|Number=Sing|PronType=Art
     6    det  _    _    *
6    blogueur  blogueur  NOUN _    Gender=Masc|Number=Sing 10   nsubj    _
     _    *
7    de   de   ADP  _    _    8    case _    _    *
8    Bahrain   Bahrain   PROPN  _    _    6    nmod _    _    *
9    Ammaro    Ammaro    PROPN  _    _    8    flat:name _    _    *
10   espère    espérer   VERB _
     Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin    0    root _
     _    *
11   que  que  SCONJ  _    _    15   mark _    _    *
12   l'   le   DET  _    Definite=Def|Number=Sing|PronType=Art 13   det  _
     _    *
13   affaire   affaire   NOUN _    Gender=Fem|Number=Sing  15   nsubj    _
     _    *
14   ne   ne   ADV  _    Polarity=Neg  15   advmod   _    _    *
15   ternira   ternir    VERB _
     Mood=Ind|Number=Sing|Person=3|Tense=Fut|VerbForm=Fin     10   ccomp
     _    _    1:COLL
16   pas  pas  ADV  _    Polarity=Neg  15   advmod   _    _    *
17   la   le   DET  _    Definite=Def|Gender=Fem|Number=Sing|PronType=Art
     18   det  _    _    *
18   réputation    réputation    NOUN _    Gender=Fem|Number=Sing  15   obj
     _    _    1
19   de   de   ADP  _    _    20   case _    _    *
20   Dubaï     Dubaï     PROPN  _    _    18   nmod _    _    *
21   .    .    PUNCT  _    _    10   punct  _    _    *

# source_sent_id = GV.tri.en::195
# text = On the other hand, Bahraini blogger Ammaro said the case shouldn't
be used to tarnish Dubai's reputation.
1    On   on   ADP  IN   _    4    case _    _    *
2    the  the  DET  DT   Definite=Def|PronType=Art    4    det  _    _    *
3    other     other     ADJ  JJ   Degree=Pos    4    amod _    _    *
4    hand hand NOUN NN   Number=Sing    9    obl  _    _    *
5    ,    ,    PUNCT  ,    _    9    punct  _    _    *
6    Bahraini  Bahraini  ADJ  JJ   Degree=Pos    7    amod _    _    *
7    blogger   blogger   NOUN NN   Number=Sing    9    nsubj    _    _    *
8    Ammaro    Ammaro    PROPN  NNP  Number=Sing    7    flat _    _    *
9    said say  VERB VBD  Mood=Ind|Tense=Past|VerbForm=Fin 0    root _    _
     *
10   the  the  DET  DT   Definite=Def|PronType=Art    11   det  _    _    *
11   case case NOUN NN   Number=Sing    15   nsubj:pass     _    _    *
12   should    should    AUX  MD   VerbForm=Fin  15   aux  _    _    *
13   n't  not  PART RB   _    15   advmod   _    _    *
14   be   be   AUX  VB   VerbForm=Inf  15   aux:pass _    _    *
15   used use  VERB VBN  Tense=Past|VerbForm=Part|Voice=Pass   9    ccomp
     _    _    *
16   to   to   PART TO   _    17   mark _    _    *
17   tarnish   tarnish   VERB VB   VerbForm=Inf  15   xcomp    _    _
     1:COLL
```

```
18  Dubai     Dubai     PROPN     NNP  Number=Sing   20   nmod:poss _    _
    *
19  's   's   PART POS  _     18   case _       _     *
20  reputation     reputation     NOUN NN   Number=Sing   17   obj  _    _
    1
21  .    .    PUNCT     .    _     9    punct     _    _     *
```

La projection est alors terminée. Bien évidemment, toutes les projections ne trouvent pas un équivalent direct comme dans notre exemple. Dans les sections suivantes, nous en évaluons la qualité et dressons une typologie des différences observées avec le corpus français.

11.1.2. Evaluation standard

Avant de procéder à l'évaluation de la qualité de la projection, les résultats « bruts » de cette dernière sont présentés dans le tableau suivant.

Corpus	Tokens annotés (FR)	Tokens annotés (projection EN)	Différence
GV	3522	2224	-36,85%
TED	1565	1205	-23,00%
UN	5867	3596	-38,70%
WM	3735	2532	-32,21%
Tous	14 669	9557	**-34,85%**

Tableau 14 : Résultats bruts de la projection FR-EN

On constate qu'au cours de la projection, environ un tiers des tokens n'ont pas trouvé leur équivalent dans le corpus cible. Ce phénomène semble étroitement corrélé avec le nombre total de tokens annotés du sous-corpus source.

Pour évaluer la qualité de notre méthodologie de projection automatique, nous avons sélectionné aléatoirement 500 phrases du corpus anglais complet, pour lesquelles nous avons effectué une projection manuelle. En fonction des tokens annotés dans le corpus français, nous avons projeté les annotations sur les tokens anglais correspondants. Cette projection manuelle a été confrontée à celle effectuée automatiquement sur les mêmes données. Les résultats sont détaillés dans le tableau suivant.

	Précision	Rappel	F-mesure
EP	90 / 182 = **49,45**	90 / 108 = **83,33**	**62,07**
Token	192 / 299 = **64,21**	192 / 217 = **88,48**	**74,42**

Tableau 15 : Résultats de l'évaluation standard de la projection automatique FR > EN

Les résultats sont corrects, bien qu'on puisse constater qu'il existe une disparité notable selon qu'on se base sur une collocation complète (ligne EP) ou seulement sur les tokens (ligne Token). Ces 12 points de différence entre les F-mesures s'expliquent en partie par le fait qu'une collocation en français ne se traduise pas toujours par plusieurs tokens en anglais. Nous détaillerons les différences observées plus bas.

11.1.3. Correction / augmentation du corpus anglais

Au contraire de l'évaluation de l'annotation automatique, un nombre important de collocations n'étaient pas complètes, avec seulement un des deux éléments de la collocation annoté. Il a donc fallu effectuer un travail conséquent et chronophage, requérant une certaine rigueur, pour corriger et augmenter le corpus anglais. Ce travail a été mené en plusieurs étapes.

Tout d'abord, nous avons extrait tous les patrons morphosyntaxiques des tokens de collocations établis pendant la projection. Ceci nous a permis de nous rendre compte que des patrons autres que `(NOUN,VERB)` ou `(NOUN,ADP,VERB)` avaient été annotés au cours de la projection. Les patrons ne répondant pas aux contraintes morphosyntaxiques imposées ont été supprimés.

Ensuite, nous avons extrait tous les patrons de lemmes des collocations complètes du corpus, pour lesquels nous avons généré automatiquement des expressions régulières. Ces dernières nous ont permis de retrouver les expressions similaires qui n'auraient été annotées que partiellement ou pas du tout, à cause de phénomènes liés à la traduction, par exemple.

Une fois ces deux étapes effectuées, nous avons parcouru le corpus avec le script de validation `cupt` fourni par la *Shared Task* de PARSEME, en vue de détecter les annotations incomplètes qui n'auraient pas été extraites dans l'étape précédente. Ceci peut être dû au fait qu'aucune collocation répondant à ce patron n'était complète dans l'intégralité du corpus, ou tout simplement que la collocation française ne trouvait pas d'équivalent multi-tokens en anglais. Ce script nous a permis également de vérifier la numérotation des annotations, afin de ne pas avoir de conflit entre les annotations projetées automatiquement et celles ajoutées dans la deuxième étape du processus d'augmentation.

Arrivé à ce stade-là du nettoyage du corpus anglais, aucun patron morphosyntaxique interdit, aucune annotation incomplète, aucune erreur de numérotation ne subsistait. Dans un dernier temps, nous avons à nouveau extrait la liste des patrons de lemmes, afin cette fois de vérifier qu'il s'agissait bien de collocations acceptables en termes de mesure d'association. Pour ce faire, nous avons vérifié sur *Sketch Engine* que le score d'association des deux composants de la collocation proposé par l'outil *Word Sketch*[21] sur le corpus *English Web 2020 (enTenTen20)* était suffisamment élevé. Cet outil permet d'obtenir le score d'association *logDice* d'une paire de termes en se basant sur de très larges corpus. Nous avons déterminé qu'un seuil autour de 9 ou supérieur à cette valeur indiquait qu'il s'agissait d'une collocation. Pour prendre un exemple, le verbe *play* avec pour objet *role* présente un score *logDice* de 12,0. Les collocations dont le score était trop faible ont été retirées du corpus final.

Ces quelques paragraphes ne rendent pas justice au caractère particulièrement fastidieux de ce travail de vérification, notamment quand on prend en considération que le corpus global contient plus de 100 000 phrases. Dans la sous-section suivante, nous présentons la typologie des erreurs résultant de la projection automatique.

11.1.4. Typologie des erreurs commises par la projection automatique

Nous l'avons vu, environ un tiers des tokens annotés du corpus français ne trouvent pas leur équivalent après projection vers le corpus anglais (-34,85%). En outre, environ la même proportion des annotations dont au moins un token a bien été projeté semblent incomplètes (F-mesure 62,07). Nous tâcherons de dresser ici une typologie des causes de ces chiffres, et bien

[21] L'outil *Word Sketch* de *Sketch Engine* utilise la mesure d'association *logDice* (Rychlỳ, 2008), une variation du coefficient *Dice* permettant d'obtenir des nombres plus élevés (contrairement à ceux de *Dice*, jugés trop bas). La formule utilisée est la suivante : $logDice = 14 + \log_2 D = 14 + \log_2 \frac{2f_{xy}}{f_x + f_y}$

qu'il soit difficile de l'ignorer totalement, nous essaierons de n'entrer que superficiellement dans les détails relatifs à l'étude linguistique contrastive, qui sera abordée dans la section 13.

11.1.4.1. Encapsulation

Bien souvent, là où le texte français nécessitait l'utilisation de plusieurs mots pour exprimer le sens de l'expression, ce même sens se retrouvait encapsulé dans un token unique dans le texte anglais. Les exemples sont nombreux : par exemple, *poser|question* sera fréquemment traduit simplement par *ask* (GV:1674) :

FR : *Le blogueur Asad Ali Mohamadi écrit que ses voisins à Copenhague, au Danemark, lui **[posent]_1:COLL** des **[questions]_1** sur Neda, et que dès qu'on allume la télévision et Internet, on voit des informations sur l'Iran et Neda.*

EN : *Blogger Asad Ali Mohamadi, writes that his neighbours in Copenhagen, Denmark, are **asking** him about Neda, and that as soon as you turn on the television and internet you see news about Iran and Neda.*

Nous reviendrons plus longuement sur ces cas d'encapsulation.

11.1.4.1. Ellipse

Les cas d'encapsulation concernent les verbes, comme c'est le cas de *ask* vis-à-vis de *poser|question*. Le cas contraire est également récurrent, c'est-à-dire que c'est le token nominal qui est absent, s'apparentant plus exactement à une ellipse. Ces cas d'ellipse expliqueront la chute du nombre d'occurrences de certaines expressions omniprésentes en français, bien plus rares en anglais. L'exemple suivant illustre notre propos (UN:10514) :

FR *: Le Conseil de sécurité, au paragraphe 14 de sa résolution 687 (1991), a pris soin de noter que les **[mesures]_2:COLL** que **[prendrait]_2** l'Iraq pour satisfaire aux conditions du cessez-le-feu (…).*

EN : *The Security Council, in paragraph 14 of its resolution 687 (1991), was careful to note that **actions** by Iraq to comply with the terms of the cease-fire (…).*

Ici, la collocation *prendre|mesure* est rendue uniquement par le token *actions* en anglais. L'expression aurait pu être complète si le participe passé *taken* y avait été adjoint, ou encore si la syntaxe avait été respectée d'une langue à l'autre avec une proposition relative équivalente *(the actions that Iraq would take)*, mais l'anglais en fait l'économie.

11.1.4.2. Choix lexical divergent

Il arrive régulièrement que la projection fautive soit due à un choix lexical drastiquement différent entre les deux langues pour exprimer plus ou moins la même idée. Encore une fois, les exemples sont assez nombreux. En voici un, dont la collocation *suivre|cours* est rendue par *go|to|class* en anglais (GV:25473) :

FR *: La plupart du temps, les très jeunes employées de maison qui viennent de zones rurales pour se former et aider leurs familles doivent **[suivre]_1:COLL** des **[cours]_1** du soir (si on les y autorise).*

EN : *In most cases teenage maids that come from rural areas to help themselves and poor families are made to **go to** evening **class** (if ever allowed).*

Ces choix lexicaux divergents sont aussi régulièrement illustrés par l'utilisation de verbes faibles, synonymes d'appauvrissement. Nous y reviendrons.

11.1.4.3. Transfert des parties du discours

Notre méthodologie de projection n'inclut pas les parties du discours dans le choix final pour projeter ou pas une annotation. Nous pensons que cette contrainte supplémentaire augmenterait potentiellement la précision, au détriment cependant de l'observation de phénomènes de transfert. Il n'est pas rare en outre qu'un token étiqueté VERB en français (comme un participe passé) soit projeté sur un token étiqueté ADJ en anglais. Dans d'autres cas, le verbe dans la langue-source est un substantif dans la langue-cible, et vice versa. À cela peut s'ajouter des tournures euphémisantes, comme en témoigne l'exemple suivant (TED:5640) :

FR : *(…) et au bout de deux heures, il y a des* ***[besoins]_1:COLL*** *naturels à* ***[satisfaire]_1****, et tout le monde se lève (…).*

EN : *(…) and two hours in, there* ***needs*** *to be that bio break, and everyone stands up, (…).*

Le token *besoins*, naturellement étiqueté VERB, a été projeté sur le token *needs*, étiqueté NOUN. Evidemment, le token *satisfaire* n'a pas d'équivalent, la tournure de phrase anglaise étant bien plus directe que celle adoptée en français. Par ailleurs, les cas de nominalisation d'une collocation verbale en français ne sont pas rares, avec des exemples comme *résoudre*|*problème* devenant des expressions entièrement nominalisées telles que *resolution*|*issue*.

11.1.4.4. Reformulation complète

Parfois, la projection n'a pas lieu du tout car le segment en question a subi une transformation quasi-totale d'une langue à l'autre. C'est le cas de l'exemple suivant (UN:14965) :

FR : *(…) il a fallu* ***[payer]_1:COLL*** *237 448 dollars des Etats-Unis environ la* ***[facture]_1*** *du fret aérien à partir de l'Europe.*

EN : *Approximately $237,448 had to be appropriated to airfreight from Europe (…).*

Non seulement ni le verbe *payer* ni le substantif *facture* n'apparaissent dans la phrase anglaise, mais le segment, placé en fin de phrase en français, se retrouve en tête de phrase en anglais. L'ordre des tokens n'a en aucun cas une incidence sur la projection, mais les cas de reformulation drastique comme celui-ci illustrent parfaitement une des raisons pour lesquelles des tokens ne trouvent aucun équivalent.

11.1.4.5. Verbes à particules

Cette dernière catégorie ne représente pas une grande partie des cas, mais elle existe et fait baisser la qualité des résultats malgré tout. D'une part, les verbes à particules sont relativement difficiles à cibler correctement pour une raison simple : les alignements lexicaux utilisés ne concernent que des tokens uniques et non pas des *n-grams*. De plus, quand bien même lesdits alignements prendraient en considération les *n-grams*, notre méthodologie de projection ne considère les tokens que tels qu'ils sont séparés dans le fichier `cupt`. Ainsi, des expressions telles que celle présente dans l'exemple suivant (GV:15238) donnent rarement, si ce n'est jamais, des résultats complets :

FR : *Les pompiers ont tenté pendant des heures d'**[éteindre]_1:COLL** l'**[incendie]_1** du Musée National.*

EN : *Firefighters tried to **put out** the **fire** at the National Museum for hours.*

Dans ce cas précis, 2 tokens sur 3 ont été automatiquement annotés correctement (*out* et *fire*).

Dans cette section, nous avons détaillé notre méthodologie pour projeter les annotations du corpus français vers le corpus anglais, de la solution adoptée pour exploiter ZAP, en passant par notre méthode de génération et d'enrichissement des alignements lexicaux, jusqu'à l'évaluation de cette méthodologie sur un échantillon de 500 phrases, avant de discuter du nettoyage du corpus anglais, pour finalement terminer sur le dressage d'une typologie des erreurs emmenées par cette méthodologie. Dans la section suivante, nous présentons les résultats que nous avons obtenus sur le corpus complet et en proposons des interprétations.

11.2. Résultats et interprétations

Après avoir vérifié, corrigé et augmenté le corpus anglais, nous avons lancé une évaluation à la manière de celle que nous avions effectuée pour l'annotation automatique des collocations dans le corpus français. Nous avons mis en parallèle le corpus anglais annoté automatiquement avec notre méthodologie de projection avec celui entièrement revu. Il en ressort les résultats présentés dans le tableau ci-après[22]. Compte tenu du fait de la différence de résultats entre les scores des EP annotées et ceux des tokens annotés, nous avons inclus ces derniers ici, chose que nous n'avions pas faite pour l'annotation automatique, car les chiffres étaient quasiment identiques.

Corpus	Base	Précision	Rappel	F-mesure
GV	**EP**	573 / 1414 = *40,52*	573 / 1508 = 38,00	39,22
	Token	1240 / 2224 = 55,76	1240 / 3037 = *40,83*	*47,14*
TED	**EP**	388 / 717 = **59,69**	388 / 745 = **52,08**	**53,08**
	Token	831 / 1205 = **68,96**	831 / 1502 = **55,33**	**61,40**
UN	**EP**	849 / 2458 = 39,87	849 / 2277 = *37,29*	*35,86*
	Token	1979 / 3596 = *55,03*	1979 / 4594 = 43,08	48,33
WM	**EP**	675 / 1612 = 41,87	675 / 1446 = 46,68	44,15
	Token	1482 / 2532 = 58,33	1482 / 2912 = 50,89	54,45
Tous	**EP**	2485 / 6201 = 40,07	2485 / 5976 = 41,58	40,81
	Token	5532 / 9557 = 57,88	5532 / 12045 = 45,93	51,22

Tableau 16 : Evaluation de la projection automatique du corpus anglais après correction

Le nombre total de collocations annotées s'élève à 5976, soit une réduction de 17,13% par rapport au corpus français (7211 collocations annotées). On constate que les résultats sont moyens, chose qui, après avoir pris connaissance de la typologie des erreurs possibles engendrées par la projection automatique, apparaît normale. On peut d'ores et déjà réalisé cependant qu'il y a une corrélation assez évidente entre les résultats obtenus et le nombre de

[22] Les résultats les plus élevés sont en gras et les plus bas en italique. Ils sont considérés selon la base (EP, Token) et le type de score (Précision, Rappel, F-mesure).

collocations (voire la richesse du vocabulaire) dans chaque sous-corpus. En effet, c'est le corpus TED qui obtient les meilleurs scores.

À l'instar de ce que nous avons fait pour les résultats du corpus français, nous présenterons ici les 10 collocations les plus fréquentes pour le corpus complet, puis pour chacun des sous-corpus, en prenant en considération les patrons uniques et significatifs pour chacun de ces sous-corpus s'ils existent. Pour le moment, nous nous contenterons de faire des commentaires superficiels quand il s'agira d'étude linguistique contrastive, car nous en parlerons *in extenso* dans la section 13, mais nous comparerons inévitablement les résultats du corpus français avec ceux du corpus anglais.

Regardons tout d'abord les collocations les plus fréquentes sur le corpus dans son intégralité.

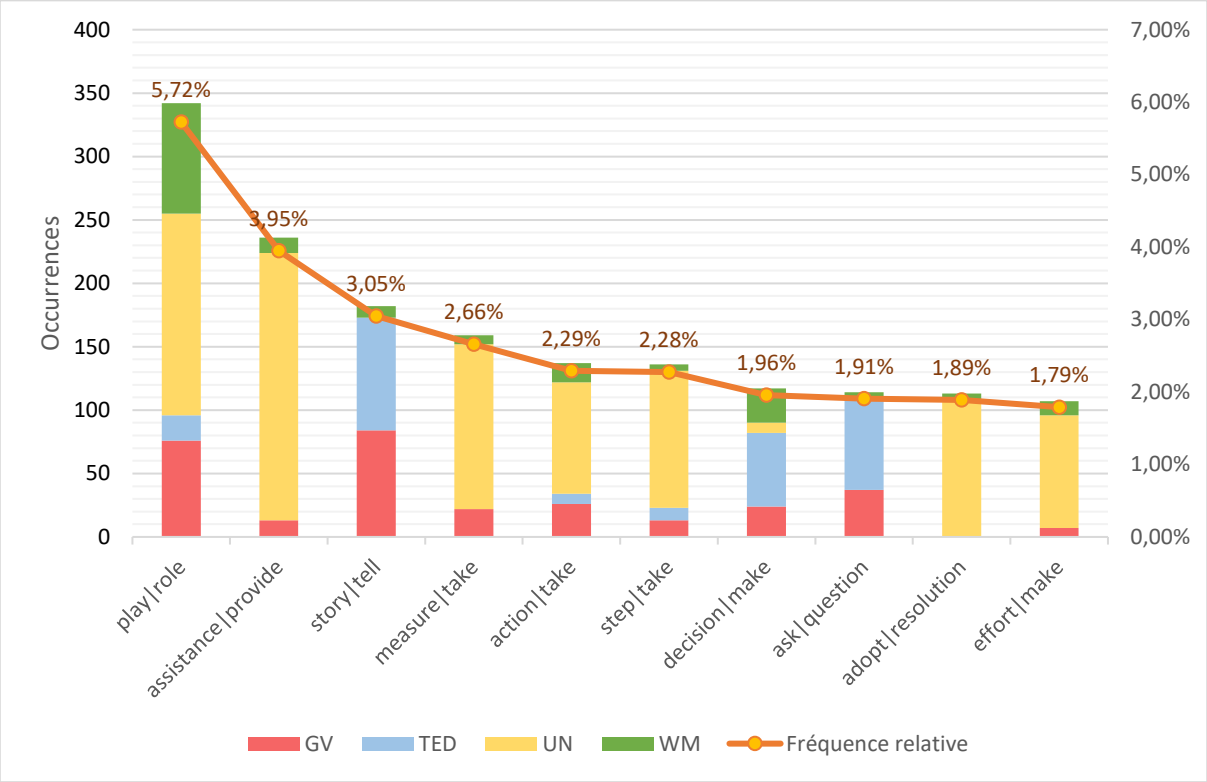

Figure 9 : Collocations anglaises les plus fréquentes et proportion par sous-corpus

Cette fois, il y a une très large différence entre la collocation la plus fréquente et le reste. En effet, *play|role* se détache largement avec 342 occurrences (fréquence relative de 5,72%), loin devant les 236 occurrences (fréquence relative de 3,95%) de *assistance|provide*, les 182 occurrences (fréquence relative de 3,05%) de *story|tell* et les 159 occurrences (fréquence relative de 2,66%) de *measure|take*. Ceci nous amène à penser une fois encore que cette expression est très largement usitée et ce, peu importe le genre textuel.

On constate également que, bien qu'elle soit la plus fréquente dans le corpus anglais, la collocation *play|role* a 106 occurrences de moins que *jouer|rôle* dans le corpus français (342 occurrences contre 448). En outre, l'équivalent de *assistance|provide* dans le corpus français, *apporter|aide*, n'apparaît que 49 fois contre les 236 occurrences dans le corpus anglais, soit quasiment 5 fois plus. Enfin, on remarque également que la collocation la plus fréquente du corpus français, c'est-à-dire *mesure|prendre* (485 occurrences pour une fréquence relative de

6,73%), est déclinée en 3 expressions synonymes en anglais, occupant la 4e, 5e et 6e places de ce classement : *measure|take* (159 occurrences, fréquence relative de 2,66%), *action|take* (137 occurrences, fréquence relative de 2,29%) et *step|take* (136 occurrences, fréquence relative de 2,28%), toutes concentrées majoritairement dans le corpus des Nations Unies. L'addition des occurrences de ces 3 expressions s'approche du total de *mesure|prendre*, et il est intéressant de constater que la fréquence d'usage d'une expression ou d'une autre est étroitement proche.

Concernant le sous-corpus GV, dont le contenu est à vocation journalistique et informative, voici les résultats obtenus.

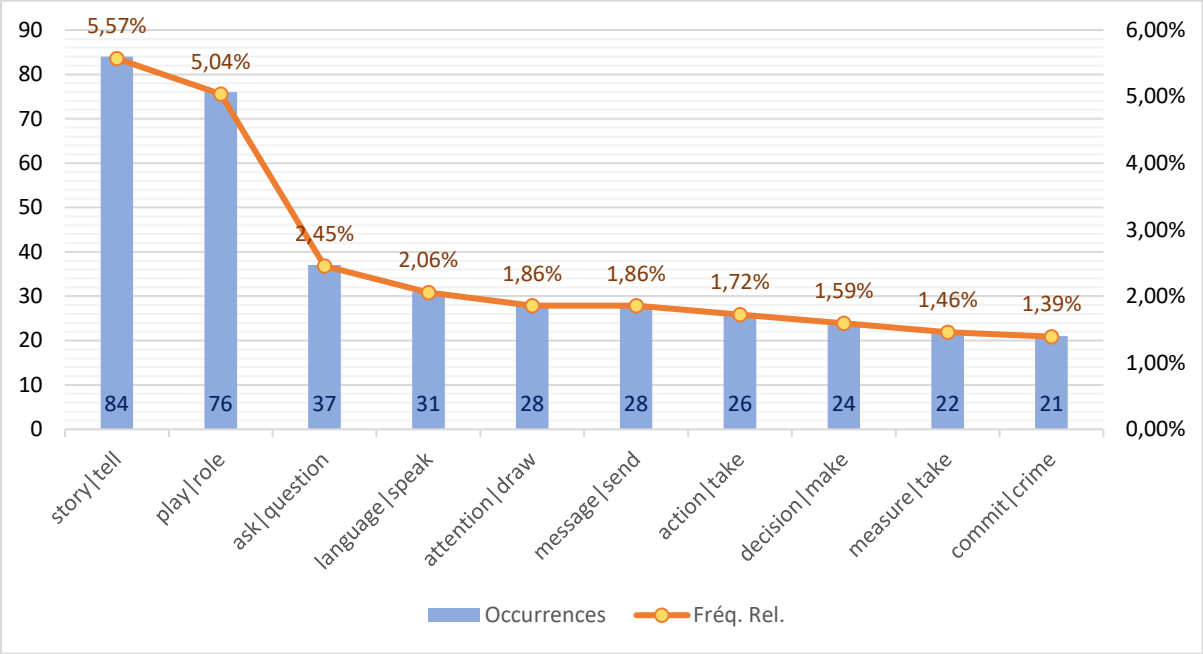

Figure 10 : Collocations anglaises les plus fréquentes (sous-corpus Global Voices)

Tout comme c'était le cas avec le corpus français, les collocations annotées dans le corpus GV anglais reflète assez fidèlement le genre de la presse, avec des expressions comme *story|tell*, *ask|question* ou encore *attention|draw*. La première de cette énumération est de loin la plus fréquente avec 84 occurrences (fréquence relative de 5,57%). Il est intéressant de remarquer que ce chiffre est quasiment égal à celui remonté dans le corpus GV français, dans lequel *histoire|raconter* était présent à 82 reprises. *A contrario*, bien qu'elle demeure dans les collocations les plus fréquentes avec 37 occurrences (fréquence relative de 2,45%), l'expression *ask|question* apparaît quasiment deux fois moins que dans le corpus français (64 occurrences, fréquence relative de 3,72%). Ceci est dû au phénomène d'encapsulation dont nous avons parlé dans la section précédente (voir section 11.1.4), où *ask* se suffit bien souvent à lui-même. Les phénomènes d'ellipse sont quant à eux responsables d'autres chiffres plus bas que dans le corpus français, comme pour *decision|make*, dont le verbe est souvent absent en anglais. Notons également que l'expression *hommage|rendre*, dont les 32 occurrences la faisait apparaître dans les 10 collocations les plus fréquentes du corpus GV français, a vu son équivalent anglais (*pay|tribute*) sortir de ce classement pour le corpus GV anglais avec seulement moitié moins d'occurrences (16 occurrences, fréquence relative de 1,06%).

Si l'on se tourne vers les collocations spécifiques à ce sous-corpus, à l'instar de ce que nous avions fait pour le corpus français, toujours en ne tenant compte que des collocations ayant au moins 5 occurrences[23], on observe que seules 2 collocations dépassent ce seuil : *face|threat* (11 occurrences, fréquence relative de 0,73%) et *chant|slogan* (9 occurrences, fréquence relative de 0,60%). Le caractère journalistique ressort une nouvelle fois, mais les remarques que nous avions faites avec les collocations spécifiques en français, à savoir que Global Voices était un journal engagé traitant de sujets se rapprochant bien souvent de conflits survenus dans le monde arabo-musulman, apparaît moins catégoriquement.

Nous avons obtenu les résultats suivants pour le corpus TED.

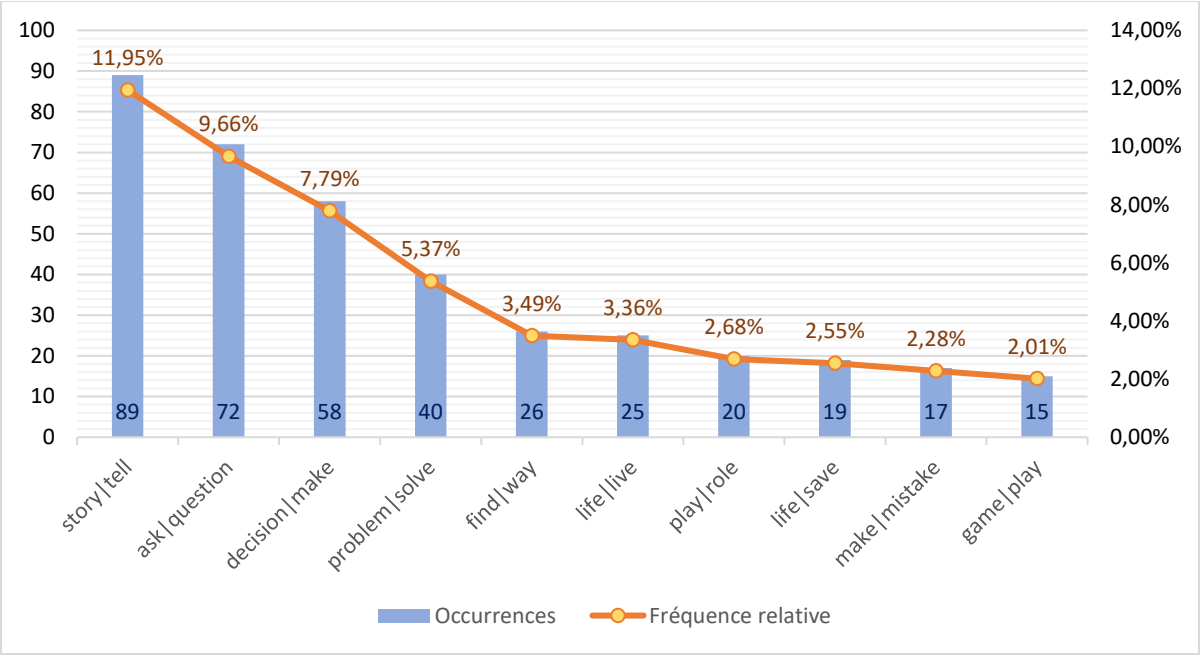

Figure 11 : Collocations anglaises les plus fréquentes (sous-corpus TED 2020)

Les différences entre les corpus TED français et anglais sont minimes : ce sont les mêmes collocations, à l'exception des deux dernières, qui se retrouvent dans les 10 collocations les plus fréquentes, sans qu'il y ait pour autant une réelle incidence sur l'interprétation que nous pourrions en donner. On constate cependant que le même phénomène d'encapsulation a lieu pour *ask|question*, dont l'équivalent français est utilisé à 27 reprises supplémentaires. Une fois encore, on remarque sans surprise que *story|tell* (89 occurrences, fréquence relative de 11,95%) et *ask|question* (72 occurrences, fréquence relative de 9,66%) sont les expressions les plus utilisées dans un contexte de conférences données par des orateurs, lesquels utilisent bien souvent des procédés rhétoriques pour capter l'attention de leur public avec des formules telles que *"Let me **tell** you a little **story** from my own negotiating experience"* (TED:2279) ou encore *"Let me just **ask** you one **question** first"* (TED:2298). En revanche, nous tourner vers les collocations spécifiques à ce corpus n'est en rien intéressant, car aucune ne dépasse le seuil fixé à 5 occurrences pour être significative.

[23] Pour rappel, nous avons choisi ce seuil de 5 occurrences plutôt que 10 car très peu de collocations spécifiques à un corpus apparaissent à cette fréquence. Selon le corpus, ce seuil de 10 n'est pas du tout atteint.

Pour ce qui est du corpus des Nations Unies, ce sont les résultats suivants qui sont ressortis de notre travail.

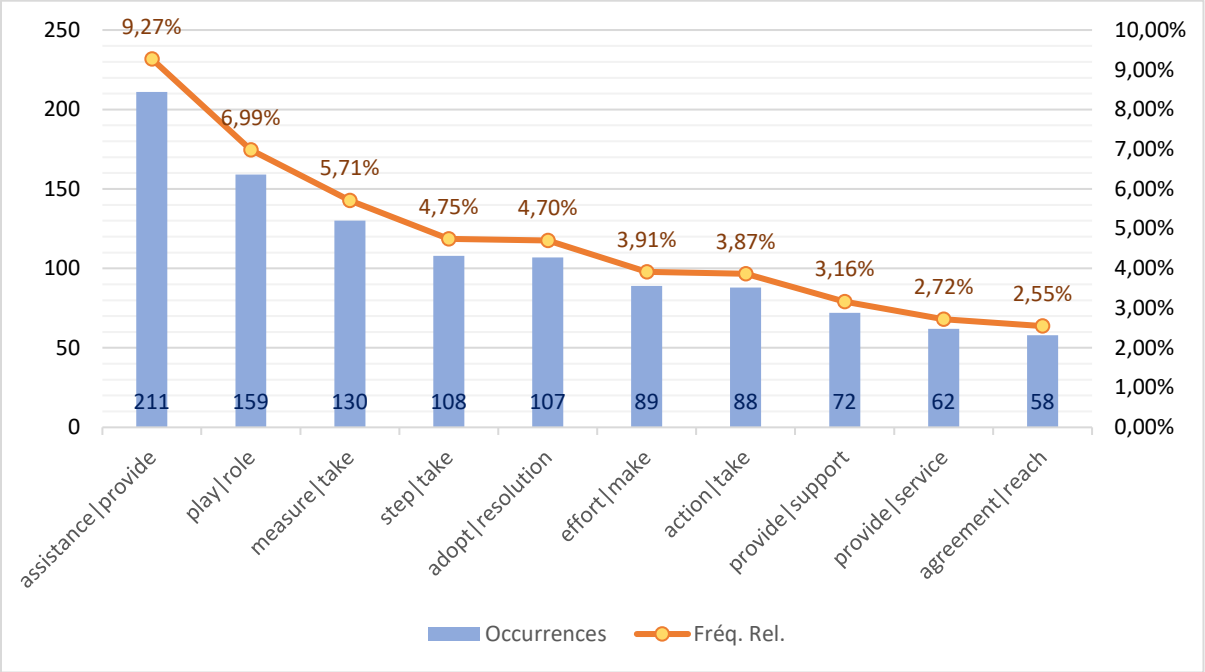

Figure 12 : Collocations anglaises les plus fréquentes (sous-corpus United Nations)

C'est dans ce sous-corpus que l'on peut remarquer les changements les plus radicaux. D'une part, quand la collocation la plus fréquente en français apparaissait 421 fois dans le corpus (14,64% de fréquence relative), la plus fréquente en anglais n'apparaît « que » 211 fois (fréquence relative de 9,27%). En outre, cette collocation n'est pas la même dans les deux langues : en français, il s'agissait de *mesure|prendre*, dont l'équivalent anglais est en fait décliné en 3 expressions synonymes, comme nous l'avons fait remarquer plus haut (*measure|take*, *step|take* et *action|take*) ; en anglais, c'est *provide|assistance* qui est la plus fréquente, tandis que son équivalent français n'apparaissait que 49 fois. Au-delà de ces différences notables, on constate que *play|role* (159 occurrences, fréquence relative de 6,99%), qui est pourtant la collocation la plus usitée sur le corpus anglais complet, a 86 occurrences de différence avec son équivalent français. Ceci est dû au phénomène d'ellipse du verbe, comme l'illustre l'exemple suivant : *"In recognition of the intergovernmental nature of the World Summit for Social Development, non-governmental organizations will have no negotiating* ***role*** *in the work of the Summit and its preparatory process"* (UN:457). Autrement, toutes ces expressions font ressortir très clairement que les thèmes juridico-politiques sont bien au centre des débats, et que la langue utilisée est assez normée pour que le nombre de collocations verbales soit important.

Du côté des collocations spécifiques à ce sous-corpus, on retrouve des expressions qui sont évidemment très proches des mêmes thèmes que nous venons d'évoquer, auxquels nous pouvons ajouter une dimension humanitaire : *objective|set* (12 occurrences, fréquence relative de 0,53%), *address|need* (10 occurrences, fréquence relative de 0,44%), *difficulty|experience* (7 occurrences, fréquence relative de 0,31%) et *authority|delegate* (7 occurrences, fréquence relative de 0,31%).

Enfin, le dernier sous-corpus, celui de WikiMatrix, a rendu les résultats présents dans la figure ci-après.

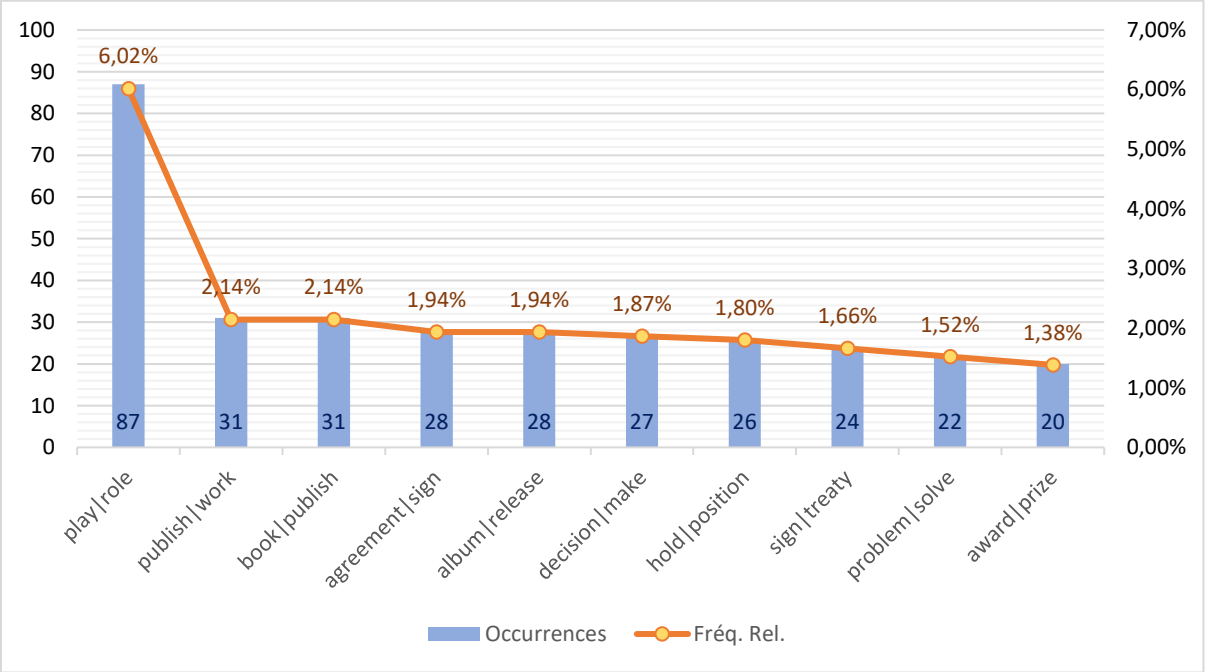

Figure 13 : Collocations anglaises les plus fréquentes (sous-corpus WikiMatrix)

Le sous-corpus WM est celui avec la plus grande variété de collocations, avec 37 patrons apparaissant au moins 10 fois dans le corpus. Outre *play*|*role* qui se détache largement (87 occurrences, fréquence relative de 6,02%), le reste est très équilibré et la courbe ne descend que très lentement. Une différence notable est celle concernant la collocation française *prix*|*recevoir* (79 occurrences, fréquence relative de 4,28%), qui est déclinée en plusieurs expressions synonymes dont *award*|*prize* (20 occurrences, fréquence relative de 1,38%), mais aussi *prize*|*receive*, *prize*|*win*, *award*|*win*, *prize*|*win*, toutes entre 9 et 18 occurrences. Au-delà de ça, la variété des expressions utilisées dans ce sous-corpus illustre bien les thèmes très éclectiques abordés par les articles Wikipedia.

Les collocations spécifiques à ce sous-corpus appuient un peu plus le dernier argument avancé : *album*|*release* (5^{e} collocation la plus fréquente, 28 occurrences et fréquence relative de 1,94%), *award*|*receive* (18 occurrences, fréquence relative de 1,24%), *record*|*set* (10 occurrences, fréquence relative de 0,69%), *victory*|*win* (7 occurrences, fréquence relative de 0,48%) et *bear*|*name* (6 occurrences pour les 3 expressions, fréquence relative de 0,33%).

11.3. Bilan

En résumé, dans cette section, nous avons traité de la méthodologie employée pour mener à bien la projection des annotations de notre corpus français vers son équivalent parallèle anglais. Pour ce faire, nous avons tiré profit de ZAP (Akbik & Vollgraf, 2018) pour générer des alignements lexicaux, que nous avons ensuite exploités avec un programme que nous avons spécifiquement développé (car la projection depuis ZAP ne se fait que depuis l'anglais vers une autre langue) pour projeter les annotations des tokens annotés en français dans une phrase-source dont un équivalent lexical aurait été trouvé dans la phrase-cible. Cette méthodologie a été testée sur un échantillon de 500 phrases pour lesquelles nous avions au préalable fait la

projection manuellement, ce qui nous a permis d'obtenir des résultats encourageants, surtout au niveau du token (F-mesure 74,42), plus qu'au niveau de l'EP (F-mesure 62,07). Nous avons pu remarquer cependant que sur un corpus complet, en l'occurrence de plus de 100 000 phrases, ces résultats chutaient.

Après la projection automatique effectuée, nous avons vérifié, corrigé et augmenté manuellement le corpus. Pour cela, nous avons extrait les patrons de collocations projetées automatiquement, normalisés avec les parties du discours pour nous débarrasser de toutes les annotations qui ne correspondaient pas à nos critères (VERB et NOUN), et également normalisés avec les lemmes de chaque annotation. Grâce à ces tuples de lemmes, nous avons généré des expressions régulières nous permettant de naviguer dans le corpus afin de retrouver les expressions qui n'auraient pas été complètement annotées, voire pas annotées du tout. Nous avons finalement vérifié le score d'association des composants de la collocation avec l'outil *Word Sketch* de *Sketch Engine*. Il s'agit là d'un processus long et fastidieux, mais somme toute inévitable pour obtenir un corpus qui ne soit pas incomplet ou fautif. En outre, une partie de ce dernier pourrait servir de corpus d'entraînement à VarIDE (ou un autre outil d'annotation automatique) pour annoter automatiquement les collocations verbales dans un autre corpus anglais.

Enfin, d'un point de vue linguistique, nous avons pu remarquer que, bien que le français et l'anglais soient des langues relativement proches, un certain nombre de différences notables dans la traduction ont fait que les résultats quantitatifs sont assez différents entre les deux corpus parallèles. D'un côté, le phénomène d'encapsulation, c'est-à-dire qu'un seul token ait un contenu sémantique équivalent à une expression polylexicale, y est pour beaucoup, notamment en ce qui concerne les collocations les plus représentées. C'est le cas, entre autres, de *poser|question* qui se retrouve très régulièrement traduit par *ask* uniquement. D'un autre côté, tous les cas d'ellipse sont également responsables d'un nombre important de changements. Nous pouvons citer les cas de *measure|take* ou surtout *play|role* pour lesquels l'anglais fait bien souvent l'économie du verbe. Outre cela, d'autres phénomènes linguistiques occasionnés au cours de la traduction (transposition catégorielle, reformulation, etc.) ont également pu être observés. Nous y reviendrons plus en détail dans la section 13 pour mener une étude contrastive trilingue.

Dans la section suivante, nous décrivons la deuxième projection effectuée du corpus anglais vers le corpus arabe. À quelques détails près, la façon de procéder a été similaire. Dans un premier temps, nous discutons de la création de la table de traduction bilingue anglais-arabe générée avec GIZA++ et des modifications apportées au code source de ZAP. Puis, nous abordons la génération et l'enrichissement des alignements obtenus avec ce dernier outil, avant d'évaluer la qualité de la projection automatique. Nous parlons ensuite de la correction / augmentation manuelle du corpus, avant de dresser une typologie des erreurs commises au cours de la projection, avant de terminer sur la présentation des résultats obtenus et leur interprétation.

12. PROJECTION DES ANNOTATIONS (ANGLAIS-ARABE)

12.1. Projection automatique

Contrairement au français, l'arabe n'est pas une des langues-cibles gérées par la version distribuée de ZAP. Ainsi, une fois le corpus anglais entièrement annoté en collocations, le travail réalisé pour la projection de l'anglais vers l'arabe nécessitait deux étapes préliminaires supplémentaires. Tout d'abord, il s'agissait de créer une table de traduction bilingue anglais-arabe avec GIZA++ que nous pourrions exploiter pour générer des alignements avec ZAP. Ensuite, des modifications mineures devaient être apportées à son code source de sorte à pouvoir effectivement nous appuyer sur cette table de traduction. Dans cette section, nous décrivons en premier lieu ces deux étapes, avant de détailler la projection des annotations en elle-même, de l'évaluer et de dresser une typologie des différences observées, avant de conclure sur une discussion des résultats obtenus après correction et augmentation de notre corpus arabe.

12.1.1. Création de la table de traduction bilingue et modification de ZAP

Pour créer notre table de traduction bilingue anglais-arabe, nous avons utilisé GIZA++. Nous avons fourni en entrée à l'outil un corpus parallèle de 12,5M de phrases alignées issues des corpus complets dont nous avons utilisé des échantillons pour notre projet (Global Voices, TED 2020, United Nations et WikiMatrix). Nous avons tout d'abord tokenisé l'ensemble des textes avec Camel Tools (Obeid et al., 2020), très efficient et rapide, notamment pour l'arabe, mais dont l'usage ne se limite pas à cette langue.

La suite correspond à l'usage standard de GIZA++, que nous avons décrit en détail dans la section 10.1, à savoir la génération des fichiers vocabulaire (`*.vcb`) et phrases (`*.snt`) à partir des corpus tokenisés, puis la génération des fichiers de cooccurrences (`*.cooc`) à partir des fichiers phrases, puis la génération des fichiers classes (`*.vcb.classes`) à partir des corpus tokenisés, avant de terminer avec l'alignement à proprement parler avec GIZA++ et l'obtention des dictionnaires bilingues (`*.actual.ti.final`) et des alignements Viterbi (`*.AA3.final`).

Pour pouvoir utiliser cette table de traduction bilingue pour l'arabe cependant, il était nécessaire de procéder à quelques ajustements mineurs dans le code source de ZAP. Tout d'abord, le fichier de la table de traduction devait être légèrement modifié (remplacement des espaces par des tabulations), puis renommé selon les conventions utilisées par l'outil pour les autres tables de traduction (`*-hmm.dict`) avant d'être placé dans le même dossier que les autres (`/src/main/resources/alignment`). De plus, afin de pouvoir initialiser une instance de l'aligneur heuristique pour l'arabe, il était nécessaire d'ajouter une constante `ARABIC` dans la classe de type `enum` appelée `Language`, ainsi qu'une ligne supplémentaire dans la méthode retournant le code de la langue nécessaire à la lecture du bon fichier `*-hmm.dict` lors de l'instanciation de l'aligneur heuristique. Dès lors, il nous était possible d'exécuter notre programme exploitant ZAP.

12.1.2. Génération et enrichissement des alignements

Pour générer et récupérer les alignements anglais-arabe, il nous a suffi de remplacer la ligne suivante par celle qui la suit dans notre programme :

```
HeuristicAligner aligner = HeuristicAligner.getInstance(Language.FRENCH);
```

```
HeuristicAligner aligner = HeuristicAligner.getInstance(Language.ARABIC);
```

La suite, à savoir l'enrichissement et la projection des annotations, s'est faite sensiblement de la même manière que celle décrite dans la section 11.1.1 pour le français vers l'anglais. Nous reprenons l'exemple que nous avions utilisé pour illustrer ce processus pour l'anglais vers l'arabe, à savoir la phrase 195 du sous-corpus GV.

L'alignement généré par notre programme exploitant ZAP pour cette phrase prend la forme suivante :

```
195 {15 used=20 ,{1.0=استخدام 13} reputation=3 ,{1.0=سمعة 18} other= 3}
7 ,{1.0=أخرى blogger=9 ,{1.0=المدون 5} said=13 ,{1.0=قال 8} n't= 11}
4 ,{1.0=لا hand=12 ,{1.0=ناحية 2} should=11 ,{1.0=أن 9} case= 15}
17 ,{1.0=القضية tarnish=14 ,{1.0=تشويه 17} be={1.0=يجب 12}}
```

Bien que l'affichage soit faussé du fait de la bidirectionnalité du texte, la forme est la même qu'avec les alignements français-anglais. Puis, une fois enrichi avec les annotations grâce à la première partie de notre programme Python, ce même alignement prend la forme suivante :

```
[['195', '15', 'used', '13', '1.0','استخدام', '*'], ['195', '20', 'reputation',
'18', '1','1.0','سمعة'], ['195', '3', 'other', '3', '1.0','أخرى', '*'], ['195',
'7', 'blogger', '5', '1.0','المدون', '*'], ['195', '9', 'said', '8', '1.0','قال',
'*'], ['195', '13', "n't", '11', '1.0','لا', '*'], ['195', '4', 'hand', '2',
'1.0','ناحية', '*'], ['195', '12', 'should', '9', '1.0','أن', '*'], ['195', '11',
'case', '15', '1.0','القضية', '*'], ['195', '17', 'tarnish', '17', '1.0','تشويه',
1':COLL'], ['195', '14', 'be', '12', '1.0','يجب', '*']]
```

Nous pouvons à nouveau remarquer que deux alignements lexicaux, correspondant aux tokens de la collocation anglaise (tarnish,reputation), contiennent une annotation. L'équivalent repéré est donc (تشويه,سمعة) *(tašwīh,sumʿa)*. Ainsi, l'étape de projection de notre programme retourne le fichier `cupt` annoté suivant pour l'arabe (les tokens concernés sont en gras) :

```
# source_sent_id = GV/v5/GV.tri.diac.ar::195
# text = مِن ناحِيَةٍ أُخْرَى, المُدَوَّن البَحْرَيْنِيّ عمارو قالَ أَنَّهُ لا يَجِب اِسْتِخْدام هٰذِهِ القَضِيَّةِ لِتَشْوِيهِ سُمْعَةِ دُبَيّ.
1    مِن   مِن   ADP  P---------    AdpType=Prep  2    case _    _    *
2    ناحية    نَاحِيَة    NOUN N------S2I    Case=Gen|Definite=Ind|Number=Sing
     8    obl  _    _    *
3    آخَر أخرى ADJ  A-----FS2I
     Case=Gen|Definite=Ind|Gender=Fem|Number=Sing   2    amod _    _    *
4    ,    ,    PUNCT    G---------    _    3    punct    _    _    *
5    المدون    مُدَوَّن NOUN N------S1D    Case=Nom|Definite=Def|Number=Sing8
     nsubj    _    _    *
6    البحريني  بَحرَيْنِيّ    X    U---------    _    7    nmod _    _    *
7    عمارو    عمارو    X    U---------    _    5    nmod _    _    *
8    قَال قال  VERB VP-A-3MS--
     Aspect=Perf|Gender=Masc|Number=Sing|Person=3|Voice=Act   0    root _
     _    *
9-10 أنه  _    _    _    _    _    _    _    *
9    أن   أَنَّ   SCONJ    C---------    _    12   mark _    _    *
10   ه    هُوَ   PRON SP---3MS4-
     Case=Acc|Gender=Masc|Number=Sing|Person=3|PronType=Prs   9    fixed
     _    _    *
11   لا   لَا   PART F---------    _    12   advmod    _    _    *
12   وَجَب يجب  VERB VIIA-3MS--
     Aspect=Imp|Gender=Masc|Mood=Ind|Number=Sing|Person=3|VerbForm=Fin|Voic
e=Act     8    ccomp    _    _    *
13   استخدام   اِستِخدَام   NOUN N------S1R
     Case=Nom|Definite=Cons|Number=Sing    12   nsubj    _    _    *
14   هذه هٰذا  DET  SD----FS2-
     Case=Gen|Gender=Fem|Number=Sing|PronType=Dem   15   det  _    _    *
15   القضية    قَضِيَّة NOUN N------S2D    Case=Gen|Definite=Def|Number=Sing13
     nmod _    _    *
16-17    لتشويه    _    _    _    _    _    _    _    *
16   ل    لِ   ADP  P---------    AdpType=Prep  17   case _    _    *
17   تشويه    تَشوِيه    NOUN N------S2R
     Case=Gen|Definite=Cons|Number=Sing    13   nmod _    _    1:COLL
18   سُمعَة سمعة NOUN N------S2R    Case=Gen|Definite=Cons|Number=Sing    17
     nmod _    _    1
19   دبي  دبي  X    U---------    _    18   nmod _    _    *
20   .    .    PUNCT    G---------    _    8    punct    _    _    *
```

De prime abord, la projection semble avoir été effectuée correctement. Or, (سمعة,تشويه) *(tašwīh,sumʿa)* n'est pas une collocation verbo-nominale, au sens où تشويه (*tašwīh*, « le fait de ternir ») est un *maṣdar* (ou nom d'action) du verbe de forme II شوّه (*šawwaha*, « ternir »), qui est de facto étiqueté NOUN. En l'occurrence, il s'agit d'une annexion indéfinie qu'on traduirait par *ternissement de réputation*. Techniquement, la projection a été réalisée avec succès, car les annotations des tokens de la langue-source ont été transférés aux tokens alignés de la langue-cible. Au niveau linguistique cependant, c'est une différence dépendante de la langue arabe que nous devrons prendre en considération dans notre analyse. Dans les sections suivantes, nous en évaluons la qualité et dressons une typologie des erreurs commises lors de la projection.

12.1.3. Evaluation standard

Avant de procéder à l'évaluation de la qualité de la projection, les résultats « bruts » de cette dernière sont présentés dans le tableau suivant.

Corpus	Tokens annotés (EN)	Tokens annotés (projection AR)	Différence
GV	3030	1576	-47,99%
TED	1494	655	-56,16%
UN	4567	3087	-32,41%
WM	2899	1943	-32,98%
Tous	14 088	7261	-39,44%

Tableau 17 : Résultats bruts de la projection EN > AR

Dans l'ensemble, il semblerait que la projection de l'anglais vers l'arabe soit environ 5% moins bonne que celle réalisée du français vers l'anglais. Les chiffres sont relativement équilibrés pour les corpus UN et WM et similaires à la première projection, avec environ un tiers de perte dans la projection, mais les corpus GV et TED souffrent d'une perte assez conséquente. Cela pourrait s'expliquer par un déséquilibre dans le corpus ayant servi à créer la table de traduction anglais-arabe.

À des fins expérimentales, nous avons pris l'initiative de procéder à une « double projection », c'est-à-dire que nous souhaitions tester d'ajouter aux tokens déjà annotés de l'anglais vers l'arabe ceux qui ne l'auraient pas été depuis le français vers l'arabe. Pour ce faire, nous avons créé une nouvelle table de traduction français-arabe en utilisant l'anglais comme langue pivot : si un token anglais était aligné avec un token français dans la table de traduction français-anglais de ZAP et si le même token anglais était aligné avec un token arabe dans notre table de traduction, alors les tokens français et arabe correspondants à ce token anglais pouvaient eux-mêmes être alignés. Pour éviter tout doublon, les tokens déjà annotés par la projection simple sont ignorés au cours de la seconde projection. Nous anticipions, grâce à cette double projection, sans doute perdre en précision mais potentiellement gagner en rappel. Pour référence, le tableau suivant détaille la différence entre la projection simple et la projection double en termes de tokens annotés.

Corpus	Tokens annotés (simple projection)	Tokens annotés (double projection)	Différence
GV	1576	1804	+12,64%
TED	655	753	+13,01%
UN	3087	4077	+24,28%
WM	1943	2590	+24,98%
Tous	7261	9882	+26,52%

Tableau 18 : Comparaison du nombre de token annotés avec simple et double projection

L'écart se creuse finalement un peu plus entre les sous-corpus qui avaient déjà une couverture correcte et ceux qui souffraient d'une perte conséquente. Néanmoins, cette double projection a le mérite d'augmenter un peu plus le nombre de tokens annotés, et la comparaison des résultats des deux projections lors de l'évaluation nous permettra de savoir si oui ou non elle s'avère utile.

Pour en évaluer la qualité, nous avons procédé de la même manière que pour la projection anglais-français, à savoir que nous avons sélectionné aléatoirement 500 phrases du corpus arabe complet, pour lesquelles nous avons effectué une projection manuelle. En fonction des tokens annotés dans le corpus anglais, nous avons projeté les annotations sur les tokens arabes

correspondants. Cette projection manuelle a été confrontée à celles effectuées automatiquement sur les mêmes données (simple et double). Les résultats sont détaillés dans le tableau suivant.

Projection	Base	Précision	Rappel	F-mesure
Simple	**EP**	84 / 405 = 20,74	84 / 211 = 39,81	**27,27**
	Token	280 / 609 = 45,98	280 / 440 = 63,64	**53,38**
Double	**EP**	84 / 409 = 20,54	84 / 211 = 39,81	**27,10**
	Token	282 / 618 = 45,63	282 / 440 = 64,09	**53,31**

Tableau 19 : Résultats de l'évaluation standard des projections simple et double vers le corpus arabe

Les résultats sont beaucoup moins bons que ceux de la projection du français vers l'anglais. Nous en détaillons toutes les raisons dans les sous-sections suivantes. Bien que seul un quart des collocations complètes ait correctement été projeté, on constate qu'au niveau du token, plus de la moitié d'entre eux ont bien trouvé un équivalent, ce qui reste encourageant. On remarque cependant que la double projection n'a que très peu d'incidence sur les résultats finaux, du moins sur cet échantillon.

12.1.4. Correction / augmentation du corpus arabe

À l'instar de ce qui a été fait pour le nettoyage du corpus anglais après projection automatique, nous avons procédé en suivant les mêmes étapes nécessaires que nous avions identifiées et décrites à la section 11.1.3. Cependant, dès les premières étapes, nous avons eu quelques surprises.

En premier lieu, nous avons extrait tous les patrons morphosyntaxiques de collocations établis pendant la projection. Même si nous en avions l'intuition, d'autant plus après la découverte de la projection de *tarnish*|*reputation* en سُمعَةتَشوِيه (*tašwīh*|*sumʿa*, « ternissement|réputation »), un nombre important des collocations complètes annotées *via* la projection avaient pour patron morphosyntaxique `NOUN,NOUN`. Ce que nous n'avions pas anticipé en revanche, c'était la proportion de ces patrons-là : plus de 45% des annotations complètes étaient concernées par ce patron-là (exactement 47,03% pour la projection simple et 46,55% pour la projection double). Cette découverte était à la fois inquiétante d'un point de vue métriques de performance de la projection (voir Tableau 19), mais extrêmement intéressante d'un point de vue linguistique. Plus préoccupant cependant, 84 annotations projetées complètes contenaient un token étiqueté `X`, signifiant que `stanza` n'était pas parvenu à un traitement CoNLL aussi performant que celui obtenu sur le corpus anglais. Malheureusement, cela laissait présager qu'une étape supplémentaire de normalisation pourrait être nécessaire. En effet, si l'étiquetage en parties du discours était fautif, tout autre traitement (tokenisation, lemmatisation) pouvait l'être également, allongeant encore un peu plus cette étape de nettoyage du corpus.

Après avoir supprimé toutes les annotations dont le patron morphosyntaxique ne respectait pas nos contraintes, nous avons extrait tous les patrons de lemmes des collocations complètes du corpus, pour lesquels nous avons généré automatiquement des expressions régulières. Comme pour le corpus anglais, ces dernières nous ont permis de naviguer dans le corpus pour retrouver facilement les expressions similaires au patron lemmatisé que la projection automatique n'aurait pas retrouvées. C'est au cours de cette opération que nous avons pu

détecter les tokens qui étaient mal étiquetés / lemmatisés de manière récurrente, et nous avons remédié à cela au fil de l'eau.

Une fois ce travail effectué, nous avons lancé un script créé entre temps par nos soins nous permettant de renuméroter les annotations résultant de l'augmentation manuelle. Ceci nous a fait gagner un temps conséquent en résolution manuelle de conflits de numérotation, compte tenu du nombre important d'annotations que nous avons ajoutées. Ensuite, nous avons tiré profit du script de validation `cupt` de PARSEME pour détecter les erreurs qui subsistaient.

À ce stade, le nettoyage du corpus sur le plan de la forme était complète. Dans un dernier temps, nous avons extrait une nouvelle fois les patrons de lemmes de collocations et les avons soumis à l'avis d'un expert en langue arabe[24] afin qu'il puisse trancher quant à leur acceptabilité. Nous n'avons pas pu utiliser *Sketch Engine* comme pour le corpus anglais, car les résultats des requêtes sur le corpus *Arabic Web 2012 (arTenTen12)* nous sont apparus très étranges : scores très bas, pas de relations syntaxiques mais des positions vis-à-vis du verbe, etc. En outre, ne connaissant aucune autre ressource pour mesurer le score d'association pour les collocations arabes et ce, malgré nos recherches, nous avons dû nous fier à cet avis, qui n'en demeure pas moins précieux. À la suite de ce retour, nous avons amendé une ultime fois le corpus.

Encore une fois, cette étape a été particulièrement fastidieuse, chronophage et même éreintante. Elle constitue très clairement l'aspect le plus important à améliorer selon nous. Dans la sous-section suivante, nous présentons la typologie des erreurs résultant de la projection automatique.

12.1.5. Typologie des erreurs commises par la projection automatique

Bien que nous nous soyons appuyé à la fois sur les corpus anglais et français pour mener à bien cette projection, nous ne prendrons comme point de référence initial que le corpus anglais pour ne pas empiéter sur l'étude linguistique contrastive trilingue menée plus loin (voir section 13). Dans la typologie suivante, que la projection se soit faite partiellement ou entièrement, elle est fautive dans tous les cas.

12.1.5.1. *Maṣdar* + nom

Nous l'avons dit, plus de 45% des annotations projetées complètes étaient entièrement nominales, la plupart d'entre elles ayant un *maṣdar* (ou nom d'action) à la place du verbe. Le fait qu'une telle proportion de verbes trouvent un équivalent nominal dans la projection nous met sur une piste fort intéressante, qui est que l'arabe remplace très volontiers les verbes à forme non finie par leur équivalent nominal. Prenons un des exemples les plus fréquents pour la collocation *mesure*|*prendre* (UN:3385) :

EN : *The meeting reviewed progress made at the mid-point of the preparatory phase of the Year and identified further **[measures]_1:COLL** to be **[taken]_1**.*

[24] Il s'agit de Frédéric Imbert, professeur des universités et agrégé de langue arabe, co-directeur de ce travail.

AR[25] : المتعين وَأَسْتَعْرِض الاجتماع التَقَدُّمِ المُحْرِز فِي مُنْتَصَفِ المَرْحَلَةِ التَحْضِيرِيَّةُ لِلسَنَةِ وَحَدَّدَ **التَدابِيرِ** الإضافية **إِتِّخاذُها**.

Les tokens en gras, (تَدبِير,اِتِّخَاذ) *(tadbīr,ittiẖāḏ)*, se traduisent par *(mesure,prise)*. À la fin du segment, une traduction littérale possible serait « et il a identifié les mesures supplémentaires nécessaires à leur prise ». Ces exemples sont très nombreux.

12.1.5.2. Tournures passives avec تَمَّ *(tamma)*

Rejoignant un peu le patron *maṣdar* + nom évoqué précédemment, les tournures passives avec تَمَّ *(tamma)* amènent à une projection erronée. En effet, la forme canonique (et recommandée) de la voix passive en arabe est le changement de vocalisation par rapport à celles utilisées à la voix active. Ainsi, كُتِبَ (*kutiba*, « on a écrit » ou « il a été écrit ») est la voix passive de كَتَبَ (*kataba*, « il a écrit »). Cependant, cet usage cède de plus en plus de terrain à une autre tournure, sous l'influence de la langue utilisée dans la presse, elle-même influencée par certaines formulations utilisées par les langues occidentales, en particulier l'anglais, tournure dans laquelle on utilise le verbe تَمَّ *(tamma)* suivi du *maṣdar*, qui fait office de sujet. Alors, ce qui aurait pu être une collocation verbo-nominale n'en est plus une. Ces cas sont nombreux. En voici un exemple (WM:13428) :

EN : *In the event that the orders were to be carried out, the **[action]_1:COLL [taken]_1** could be the last official act of Her Majesty's Government.*

AR : فِي حالَةِ تَنْفِيذِ الأَوامِرَ يُمْكِن أَنَّ يَكُون **الإِجْراءَ الَّذِي تَمَّ اِتِّخاذُهُ** هُوَ العَمَلِ الرَسْمِيّ الأَخِير لِحُكُومَةِ صاحِبَةِ الجَلالَةِ.

Le segment en gras (الإِجْراءَ الَّذِي تَمَّ اِتِّخاذُهُ, *al-ʾiğrāʾa al-lāḏī tamma ittiẖāḏa-hu*) traduit littéralement donnerait « la mesure qui a été la prise d'elle ». Au niveau morphosyntaxique, nous avons donc, si on omet le pronom relatif الَّذِي *(al-lāḏī*, « qui »*)* et le pronom de rappel ه *(-hu*, « lui »*)*, on obtient un patron verbe-nom-nom (ou plus précisément verbe-nom-*maṣdar*). Dans tous les cas, l'expression ne peut prétendre au statut de collocation verbo-nominale.

12.1.5.3. Participes passés étiquetés ADJ

C'est une différence de traitement (justifiée, selon nous) entre l'anglais (et le français) et l'arabe. Le participe passé en anglais et en français est très majoritairement étiqueté VERB, mais le participe passé (اسم المفعول *ism al-mafʿūl*, « participe passif ») arabe est quant à lui étiqueté ADJ. Ainsi, plus de 130 annotations projetées complètes se retrouvent avec un composant adjectival. Parmi ces erreurs de projection, on trouve l'exemple suivant (UN:14239) :

EN : *(…) the **[decisions]_1:COLL [taken]_1** by the Security Council and other competent organs of the United Nations and its specialized agencies aimed at the prevention, suppression and punishment of the crime of apartheid, in accordance with article VI of the Convention.*

[25] Afin de ne pas alourdir la lecture plus que nécessaire et comme la phrase-source en anglais est déjà présente, nous ne proposerons pas de translittération complète pour les phrases citées en exemples ni ne proposerons de traduction en français. En revanche, les tokens faisant l'objet de l'exemple seront toujours à la fois translittérés et traduits. Le segment intéressant vis-à-vis de notre argumentation sera également traduit littéralement en français pour illustrer nos propos.

AR : **وللقرارات المتخذة** من قبل مجلس الأمن وغيره من أجهزة الأمم المتحدة المختصة ووكالاتها المتخصصة والرامية إلى منع جريمة الفصل العنصري وقمعها والمعاقبة عليها، وفقا للمادة السادسة من الاتفاقية.

Les tokens en gras, lemmatisés en (مُتَّخَذ,قَرَار *qarār,muttaḫaḏ)*, se traduisent par *(décision,prise)*. La traduction du début du segment n'est pas différente de celle du segment anglais (« les décisions prises par le Conseil de sécurité et d'autres organes compétents de l'ONU […] »), mais le participe passé est étiqueté ADJ et ne peut donc pas être considéré comme un élément d'une collocation verbo-nominale.

12.1.5.4. Encapsulation

À l'instar de ce dont nous avons parlé pendant la première projection, une collocation anglaise peut ne pas trouver son équivalent polylexical en arabe, *via* l'encapsulation du contenu sémantique de l'EP dans un token unique. C'est un phénomène relativement récurrent en arabe compte tenu du contenu sémantique intrinsèque véhiculé par les formes augmentées (voir section 3.2.3.2). L'exemple suivant en est l'illustration (GV:17225) :

EN : *In an interview with Rupert Wingfield-Hayes of BBC News Tokyo, Miyamoto mentions one of her friends growing up was unable to cope with the treatment he from his peers and **[committed]_1:COLL [suicide]_1**.*

AR : فِي مُقابَلَةٍ مَعَ رُوبِرْت وينغفيلد هايز مِن بِي بِي سِي نِيُوز طُوكِيُو، ذَكَرَت ميامونو أَحَدُ أَصْدِقائها الَّذِي نَشَأَ غَيْر قادِرٌ عَلَى تَحْمِل المُعامَلَةِ السَيِّئَةِ مِن زُمَلائِهِ **فَانتَحَر**.

Le token en gras (انتَحَر *intaḥara*, « se suicider ») encapsule en un seul token tout le contenu sémantique de la collocation anglaise *commit|suicide*.

12.1.5.5. Ellipse

Pareillement, les cas d'ellipse sont possibles de l'anglais à l'arabe. Le couple de phrases suivant en est un exemple (GV:25081) :

EN : *Kermeki believe that this, along with social media campaigns, **[played]_2:COLL** a **[role]_2** in his release.*

AR : يَعْتَقِد كيرميكي بِأَنَّ هٰذا الأَمْرُ، بِالإضافَةِ لِحَمَلات وَسائِلِ الإعْلامِ الاِجْتِماعِيَّةِ، كانَ لَهُم **دَوْر** كَبِيرٍ فِي إطْلاقِ سَراحِ توكلى.

Le token en gras دَوْر (*dawr*, « rôle ») est employé après كانَ لَهُم (*kāna la-hum*, « ils ont eu »), faisant donc le choix de ne pas employer le verbe لَعِبَ (*laʿiba*, « jouer »).

12.1.5.6. Choix lexical divergent

Tout comme la première projection, les choix lexicaux d'une langue à l'autre peuvent être tout à fait différents. Bien souvent, faute d'un meilleur verbe, le choix du traducteur se porte sur un verbe faible comme قام بِـ (*qāma bi-*, « faire / effectuer »), appauvrissant l'expression et même la dénuant de son statut de collocation verbale selon notre définition. L'exemple suivant illustre notre propos (UN:8057) :

EN : *The meeting made possible a better understanding of the **[role]_1:COLL [played]_1** by national institutions in the specific domain of action to combat racism and racial discrimination.*

AR : بزيادة التعريف **بالدور** الذي **تقوم به** المؤسسات الوطنية في المجال الدقيق المتمثل في وسمح الاجتماع مكافحة العنصرية والتمييز العنصري.

Les tokens en gras (قَام،بِ،دَور) *(dawr,bi-,qāma)* se traduiraient littéralement par *(rôle,entreprendre)*. Bien que la traduction naturelle serait évidemment la collocation *jouer|rôle*, le choix lexical fait pencher la balance et l'association n'est plus une collocation au sens où on l'entend, à savoir avec un verbe qui ne soit pas un verbe faible.

12.1.5.7. Reformulation complète

Comme la plupart des phénomènes intervenant au cours de la traduction d'un document d'une langue à l'autre, les reformulations totales sont parfois privilégiées de l'anglais vers l'arabe pour éviter tout calque ou tournure qui serait inacceptable dans la langue-cible. L'exemple que nous proposons maintenant fait partie de cette catégorie (WM:19489) :

EN : *Nicholas was rushed back to Kyoto, where Prince Kitashirakawa Yoshihisa ordered that he be taken into the Kyoto Imperial Palace to rest, and **[messages]_1:COLL** were **[sent]_1** to Tokyo.*

AR : تَوَجَّهَ نيقولا عائِداً إِلَى كِيُوتُو حَيْثُ أَمْرُ الأَمِيرِ كيتاشيراكاوا يوشيهيسا بِنَقْلِه إِلَى قَصْرِ كِيُوتُو الإمبراطوري لِلراحَةِ وَ**أَعْلَم الإمبراطور ميجي بِالأَمْرِ فِي طُوكِيُو**.

Le segment en gras (وَأَعْلَم الإمبراطور ميجي بِالأَمْرِ فِي طُوكِيُو, *wa ʾaʿlama al-ʾimbirāṭūr Mayǧī bi-l-ʾamri fī Ṭūkyū*, « et il informa l'empereur Meiji à Tokyo de l'affaire ») est préféré à la formulation anglaise (qui est la même en français, soit dit en passant). On pourrait considérer qu'il s'agit là d'une encapsulation du sens dans un seul token, mais le fait que ce qui vient après le verbe soit explicité de la sorte, contrairement à ce qui était fait dans la langue-source, nous pousse à dire qu'il s'agit plutôt d'une reformulation.

12.1.5.8. Verbes à particules

La particule d'un verbe à particule anglais trouve rarement son équivalent parfait en arabe. Ainsi, si le verbe arabe en question appelle une particule radicalement différente de celle du verbe anglais, la projection ne se fera pas. En outre, si le verbe anglais n'est pas un verbe à particule et que son équivalent arabe en appelle une, le token-cible ne pourra pas être annoté. Voici un exemple appartenant cette catégorie-là (WM:18797) :

EN : *Morton **[received]_1:COLL** a number of **[awards]_1** during her career, including the Joan of Arc medal, and Rosalie Morton Park in Belgrade is named in her honor.*

AR : **حصلت** مورتون **على** بعض **الجوائز** خلال مسيرة عملها، منها ميدالية جوان دارك، ويوجد متنزه باسمها في بلغراد.

L'équivalent arabe de *receive|award* est un verbe dont le régime est indirect et nécessitant la particule عَلَى *(ʿalā)*, particule dont le sens le plus commun est sans doute « sur » lorsqu'il n'est pas employé avec un verbe dont il modifierait le sens. Dans ce cas précis, la projection se fait correctement sur les parties nominale et verbale de la collocation-cible, mais la particule est omise.

12.1.5.9. Erreurs de traitement `stanza`

Comme nous l'avons mentionné dans la section précédente (voir section 12.1.4), un certain nombre d'annotations n'ont pas pu être projetées efficacement à cause de la transformation du texte brut au format CoNLL avec la librairie Python `stanza`. En effet, des tokens n'ont pas fait l'objet d'un traitement précis. Il en a résulté des erreurs de tokenisation, de lemmatisation et d'étiquetage en parties du discours. Un token mal tokenisé ou mal lemmatisé ne pouvait pas trouver d'équivalent dans la table de traduction, ainsi l'alignement lexical ne pouvant pas se faire, la projection de l'annotation non plus.

Les erreurs de tokenisation que nous avons pu remarquer sont régulièrement le fait de l'agglutination de diverses particules à un substantif ou à un nom. Par exemple فسيكتبها (*fa-sa-yaktubu-hā*, « donc il l'écrira ») pris en un seul token alors qu'il est composé du « *fā*ʾ (فَـ) de conséquence », du préfixe سَـ *(sa-)* marquant le futur, du verbe à l'inaccompli يَكتُبُ (*yaktubu*), et du pronom personnel objet suffixe ـها *(-hā)*, soit 4 tokens.

Quand elles ne sont pas dues à une mauvaise tokenisation, les erreurs de lemmatisation sont parfois des pluriels brisés dont le singulier n'a pas été retrouvé, comme حُلول (*ḥulūl*, « solutions ») pas lemmatisé en حَلّ (*ḥall*, « solution »), ou une forme conjuguée qui n'a pas été traitée efficacement, comme نَتَّخِذُ (*nattaḫiḏu*, « nous prenons ») pas lemmatisée en اِتَّخَذَ (*ittaḫaḏa*, « prendre »). La difficulté réside dans le fait que ces erreurs ne sont pas régulières, et peuvent tantôt avoir été commises, tantôt pas.

Toutes ces erreurs ont nécessité une normalisation des lemmes fautifs et, le cas échéant, des étiquettes de parties du discours. La tokenisation a été ignorée dans le sens où seul le lemme du token principal a été retenu dans la normalisation (c'est-à-dire que les morphèmes agglutinés à ce dernier n'ont pas été retenus dans la normalisation), d'une part car elle aurait trop compliquée à corriger à la main (ajout de lignes dans le fichier `cupt`, nouveau calcul des relations de dépendance, etc.), et d'autre part pour éviter d'avoir des problèmes au cours de l'exécution du script d'évaluation qui requiert que les corpus source et cible soient de longueur égale.

12.2. Résultats et interprétations

Après le nettoyage du corpus arabe, nous avons effectué une évaluation similaire à celle que nous avions lancée pour la projection du français vers l'anglais. Nous avons confronté le corpus arabe annoté automatiquement avec notre méthodologie de projection et le corpus entièrement revu. Il en ressort les résultats présentés dans le tableau ci-après[26]. Nous avons inclus les résultats pour les projections simple et double, afin de pouvoir les comparer un peu plus en détail que sur l'échantillon de 500 phrases de l'évaluation standard.

[26] Comme précédemment, les résultats les plus élevés sont en gras et les plus bas en italique. Ils sont considérés selon la base (EP, Token), le type de score (Précision, Rappel, F-mesure).

Corpus	Base	Projection	Précision	Rappel	F-mesure
GV	**EP**	*Simple*	168 / 1124 = 14,95	168 / 1238 = 13,57	14,23
		Double	169 / 1344 = 12,57	169 / 1238 = 13,65	13,09
	Token	*Simple*	620 / 1577 = 39,32	620 / 2585 = 23,98	29,79
		Double	664 / 1805 = 36,79	664 / 2585 = 25,69	30,25
TED	**EP**	*Simple*	59 / 485 = 12,16	59 / 510 = *11,57*	11,86
		Double	60 / 574 = *10,45*	60 / 510 = 11,76	*11,07*
	Token	*Simple*	233 / 657 = 35,46	233 / 1051 = *22,17*	27,28
		Double	242 / 756 = *32,01*	242 / 1051 = 23,03	*26,78*
UN	**EP**	*Simple*	364 / 1953 = 18,64	364 / 1985 = 18,34	18,49
		Double	376 / 2833 = 13,27	376 / 1985 = 18,94	15,61
	Token	*Simple*	1191 / 3107 = 38,33	1191 / 4128 = 28,85	32,92
		Double	1374 / 4100 = 33,51	1374 / 4128 = **33,28**	33,40
WM	**EP**	*Simple*	296 / 1256 = **23,57**	296 / 1609 = 18,40	**20,66**
		Double	323 / 1820 = 17,75	323 / 1609 = **20,07**	18,84
	Token	*Simple*	940 / 1950 = **48,21**	940 / 3460 = 27,17	34,75
		Double	1092 / 2598 = 42,03	1092 / 3460 = 31,56	**36,05**
Tous	**EP**	*Simple*	887 / 4818 = 18,41	887 / 5342 = 16,60	17,46
		Double	928 / 6571 = 14,12	928 / 5342 = 17,37	15,58
	Token	*Simple*	2984 / 7291 = 40,93	2984 / 11224 = 26,59	32,23
		Double	3372 / 9259 = 36,42	3372 / 11224 = 30,04	32,92

Tableau 20 : Evaluation de la projection automatique du corpus arabe après correction

Globalement, nous pouvons faire deux constats. Le premier est que les résultats, à l'instar des autres évaluations sur corpus complet, se dégradent en comparaison de l'évaluation standard menée sur un échantillon de 500 phrases. En effet, projection simple et double confondues, seulement environ 16% des collocations sont correctement projetées entièrement et 32% des tokens trouvent un équivalent. Après avoir établi la typologie des erreurs remarquées au cours de la projection, on comprend tout à fait pourquoi. Le deuxième constat que l'on peut faire est que ce que nous avions anticipé avec la double projection était vrai : il y a immanquablement une perte de précision et un gain en rappel. Malheureusement, la perte de précision est trop élevée tandis que le gain en rappel est trop faible. De fait, bien que les F-mesures soient légèrement meilleures au niveau du token avec la double projection (sauf dans le cas du corpus TED), la projection simple obtient toujours une F-mesure supérieure au niveau de la collocation. Ainsi, la projection simple reste malgré tout plus intéressante.

Regardons maintenant les collocations les plus fréquentes dans le corpus complet.

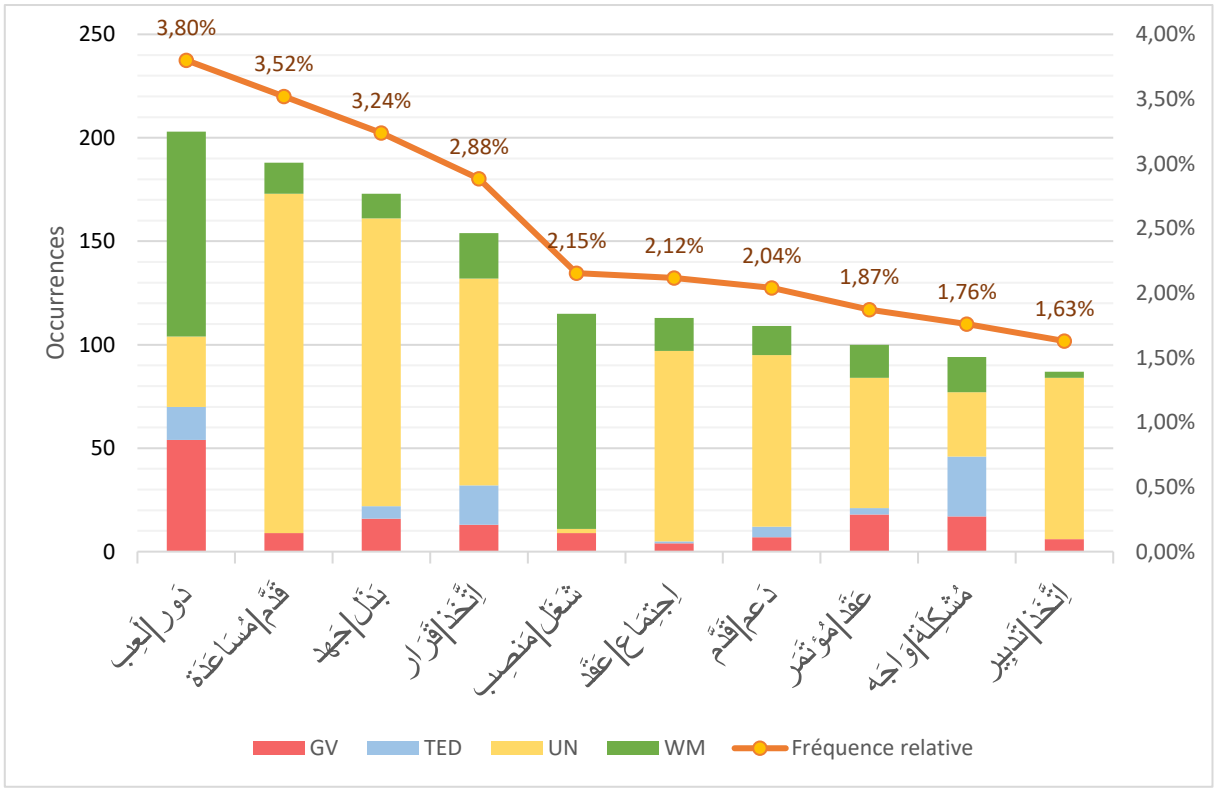

Figure 14 : Collocations arabes les plus fréquentes et proportion par sous-corpus

Comme pour le corpus anglais, c'est la collocation دَور|لَعِبَ (*dawr|laʿiba*, « *rôle|jouer* »)[27] qui est la plus fréquente. C'est par ailleurs la seule à dépasser les 200 occurrences (203 occurrences, fréquence relative de 3,80%). La courbe descend de manière régulière pour les 3 suivantes : قَدَّم|مُسَاعَدَة (*qaddama|musāʿada*, « *fournir|aide* ») avec 188 occurrences (fréquence relative de 3,52%), بَذَل|جَهد (*baḏala|ǧahd*, « *dépenser|effort* ») avec 173 occurrences (fréquence relative de 3,24%) et اِتَّخَذ|قَرَار (*ittaḫaḏa|qarār*, « *prendre|décision* ») avec 154 occurrences (fréquence relative de 2,88%). Ces trois patrons sont massivement présents dans le corpus UN. Le deuxième de cette énumération est intéressant car il entre dans la catégorie que (Brashi, 2005) a qualifié de *strong collocations*[28].

Globalement et c'est normal compte tenu du fait que ce soit le corpus contenant le moins d'annotations de nos 3 corpus parallèles (seulement 5342 collocations annotées contre 5976 pour l'anglais et 7211 pour le français), seules les 8 premiers patrons ont au moins 100 occurrences. Dès le 9e, c'est-à-dire à مُشكِلَة|وَاجَه (*muškila|wāǧaha*, « *problème|faire front* », on chute à 94 occurrences, soit une fréquence relative de 1,76%. En outre, il faut garder en mémoire qu'un nombre important des collocations en anglais et en français sont traduites par une annexion (donc un syntagme nominal) en arabe, ce qui est notamment vrai pour les collocations les plus fréquentes.

[27] Toutes les traductions fournies ici sont des traductions aussi littérales que possible pour que le lecteur non-arabophone puisse se faire une idée de l'expression arabe. Il est généralement aisé d'en déduire un équivalent idiomatique en français.

[28] Dans son étude, les *strong collocations* sont les patrons dont les composants ont été associés à une fréquence supérieure à 80% par un groupe de contrôle arabophones apprentis traducteurs et professionnels de la traduction arabophones.

Concernant le corpus GV, voici les résultats obtenus.

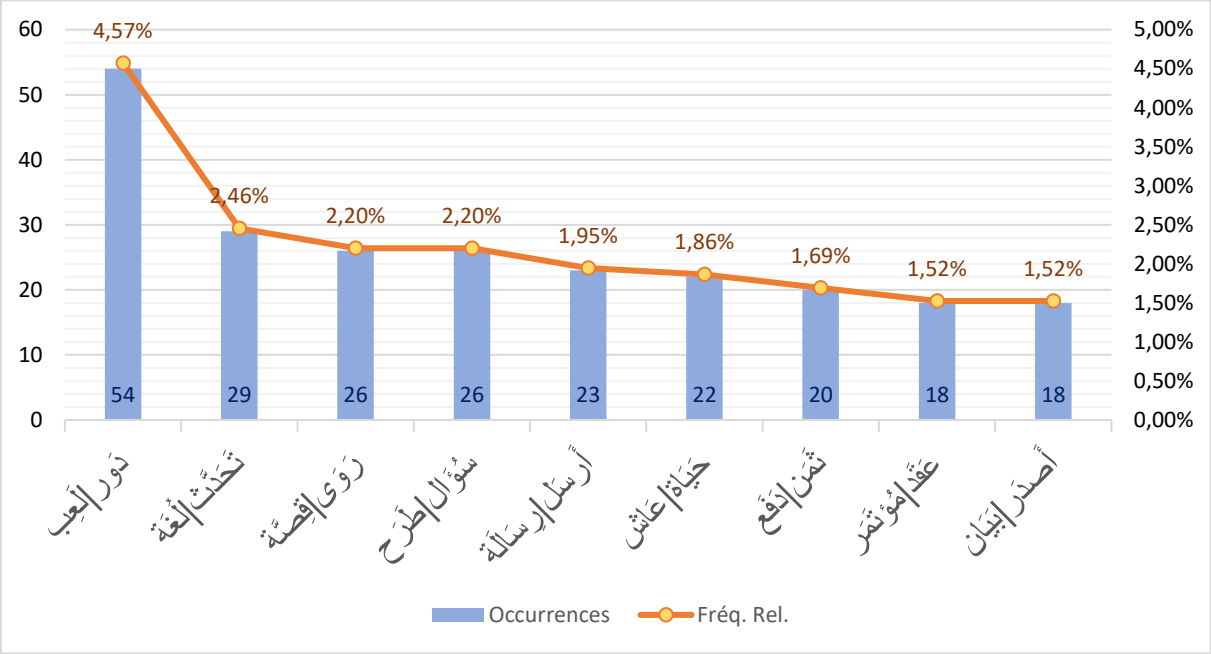

Figure 15 : Collocations arabes les plus fréquentes (sous-corpus Global Voices)

La collocation la plus fréquente du corpus complet se détache du reste : دَور|لَعِبَ (*dawr|laʿiba*, « *rôle|jouer* ») avec 54 occurrences (fréquence relative de 4,57%). On tombe quasiment à moitié moins d'occurrences dès le deuxième patron تَحَدَّثَ|لُغَة (*taḥaddaṯa|luġa*, « *parler|langue* ») avec 29 occurrences (fréquence relative de 2,46%) et le troisième patron رَوَى|قِصَّة (*rawā|qiṣṣa*, « *raconter|histoire* ») avec 26 occurrences (fréquence relative de 2,20%). On remarquera dans cette liste une première collocation dont les composants partagent la même racine trilittère : أَرسَلَ|رِسَالَة (*ʾarsala|risāla*, « *adresser|lettre* ») dont les deux composants sont formés à partir de la racine ر س ل *(rāʾ, sīn, lām)*. Nous y reviendrons, mais d'autres collocations sont formées de la même manière, parfois au détriment d'un verbe unique qui a pu être en usage auparavant, parfois à la limite du pléonasme. Nous tourner vers les collocations spécifiques à ce sous-corpus ne nous apprendra rien de plus, car aucun patron ne dépasse le seuil de 5 occurrences que nous avions fixé pour les autres langues.

Pour ce qui est du sous-corpus TED, nous avons obtenu les résultats suivants.

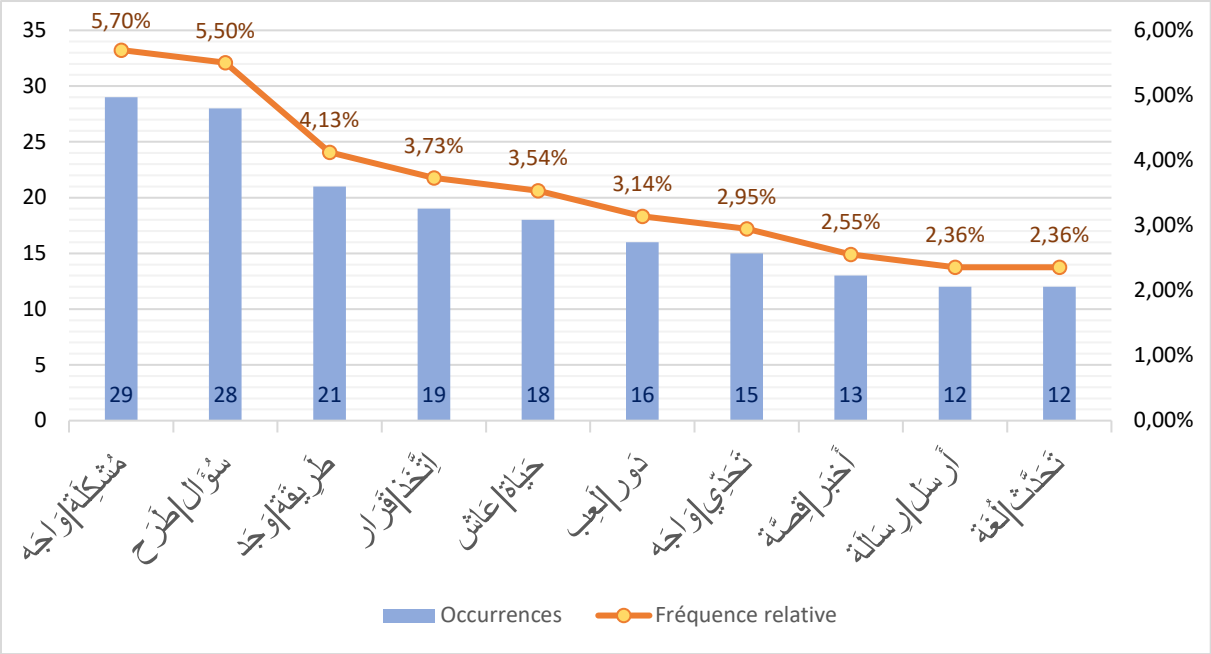

Figure 16 : Collocations arabes les plus fréquentes (sous-corpus TED 2020)

On constate tout de suite que le corpus contient peu d'annotations et que seuls 3 patrons apparaissent plus de 20 fois. Le premier est مُشكِلَة|وَاجَه (*muškila|wāğaha*, « *problème|faire front* » avec 29 occurrences (fréquence relative de 5,70%). Le deuxième est سُؤَال|طَرَح (*ṭaraḥa|suʾāl*, « *question|jeter* ») avec 28 occurrences (fréquence relative de 5,50%). Le troisième est طَرِيقَة|وَجَد (*ṭarīqa|wağada*, « *voie|trouver* ») avec 21 occurrences. La courbe décroît lentement et régulièrement ensuite.

Du côté des patrons spécifiques à TED, il n'y en a qu'un qui dépasse le seuil de 5 occurrences (fréquence relative de 0,98%), mais il est intéressant pour trois raisons. Il s'agit de la collocation قَصّ|قِصَّة (*qaṣṣa|qiṣṣa*, « *raconter|histoire* »). La première raison est qu'il s'agit d'une expression très ancienne issue du Coran (7e siècle, aux alentours de 610-632 pour la période de la révélation). Cela rend l'expression quelque peu *consacrée*, bien que des variantes synonymiques soient en usage. La deuxième raison découle de la première, à savoir que cette expression entre dans la catégorie des *strong collocations* de Brashi. Enfin, la troisième raison est celle que nous avons mentionné plus tôt, à savoir que les deux composants de la collocation sont formés sur la même racine trilittère ق ص ص *(qāf ṣād ṣād)*. Lorsque nous parlions de pléonasme, une expression comme celle-ci, bien qu'elle soit tout à fait correcte en arabe, s'apparenterait presque à une formule comme *conter un conte* en français. Du point de vue du contenu du sous-corpus TED en revanche, elle ne nous apprend rien que nous ne sachions pas déjà à ce stade.

En ce qui concerne celui des Nations Unies, ce sont les résultats suivants qui ont primé.

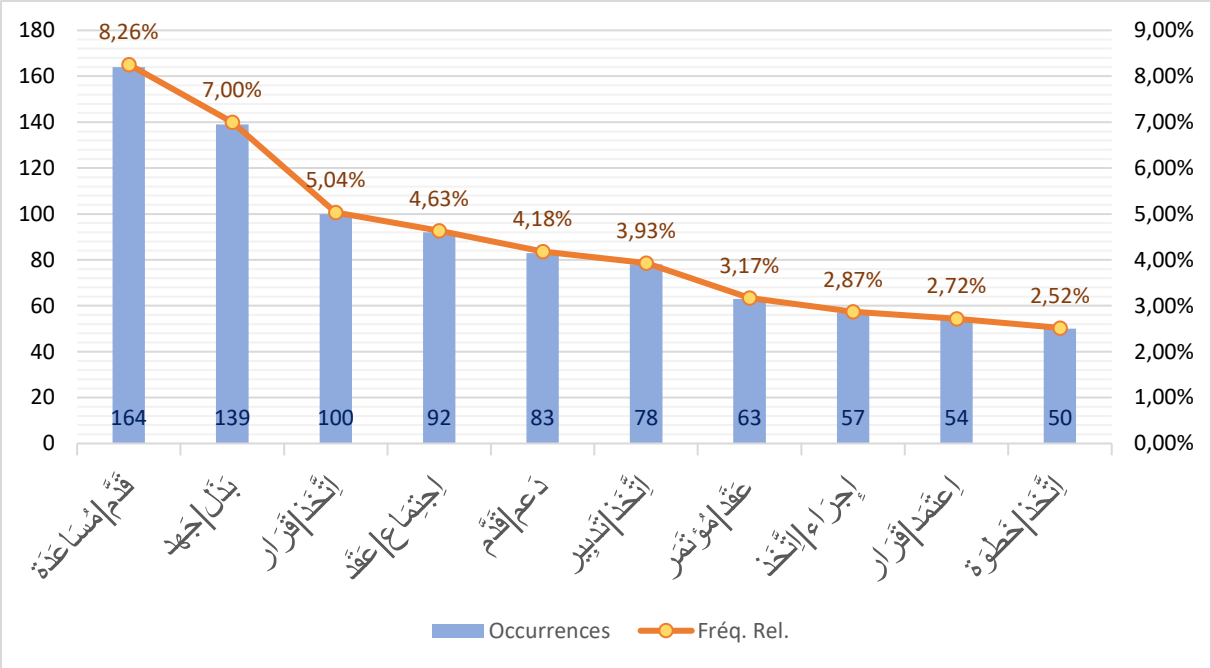

Figure 17 : Collocations arabes les plus fréquentes (sous-corpus United Nations)

Sans grande surprise, le sous-corpus UN est à nouveau celui qui présente le plus grand nombre d'annotations avec 1986 collocations annotées, soit plus de 37% du total sur le corpus arabe. Dix patrons en tout ont au moins 50 occurrences : c'est le seul des quatre sous-corpus à avoir plus de quatre patrons dépassant ce seuil. La fréquence relative des deux premiers patrons est également supérieure à celle de toutes les autres collocations, tous corpus confondus. Il s'agit de قَدَّم|مُسَاعَدَة (*qaddama|musāʿada*, « *fournir|aide* ») avec 164 occurrences (fréquence relative de 8,26%) et بَذَل|جَهد (*baḏala|ǧahd*, « *dépenser|effort* ») avec 139 occurrences (fréquence relative de 7,00%). Cette dernière, nous l'avons déjà précisé, fait partie des collocations appartenant à la catégorie des *strong collocations* de Brashi. Il est à nouveau intéressant de noter que l'équivalent de *mesure|prendre*, à l'instar de ce que nous avions observé pour l'anglais, se décline en trois variantes synonymiques : اِتَّخَذ|تَدبِير (*ittaḫaḏa|tadbīr*, « *prendre|mesure* ») avec 78 occurrences (fréquence relative de 3,93%), إِجرَاء|اِتَّخَذ (*ʾiǧrāʾ|ittaḫaḏa*, « *mesure|prendre* ») avec 57 occurrences (fréquence relative de 2,87%) et اِتَّخَذ|خَطوَة (*ittaḫaḏa|ḫaṭwa*, « *prendre|mesure* ») avec 50 occurrences (fréquence relative de 2,52%). Si ces variations lexicales tout à fait interchangeables se retrouvaient en un seul patron, ce dernier constituerait la collocation la plus fréquente du corpus complet avec quasiment 300 occurrences.

Si l'on examine les collocations spécifiques à ce sous-corpus, on retrouve toujours les thèmes juridiques, économiques et politiques associés à ce sous-corpus. La plus fréquente est خَصَّص|مَورِد (*ḫaṣṣaṣa|mawrid*, « *allouer|ressource* ») avec 12 occurrences (fréquence relative de 0,60%). On trouve également جَزَاءَة|فَرَض (*faraḍa|ǧazāʾa*, « *sanction|imposer* ») avec 9 occurrences (fréquence relative de 0,45%) ou encore حِصَار|فَرَض (*faraḍa|ḥiṣār*, « *embargo|imposer* ») avec 8 occurrences (fréquence relative de 0,40%).

Enfin, voici les résultats obtenus sur le sous-corpus WM.

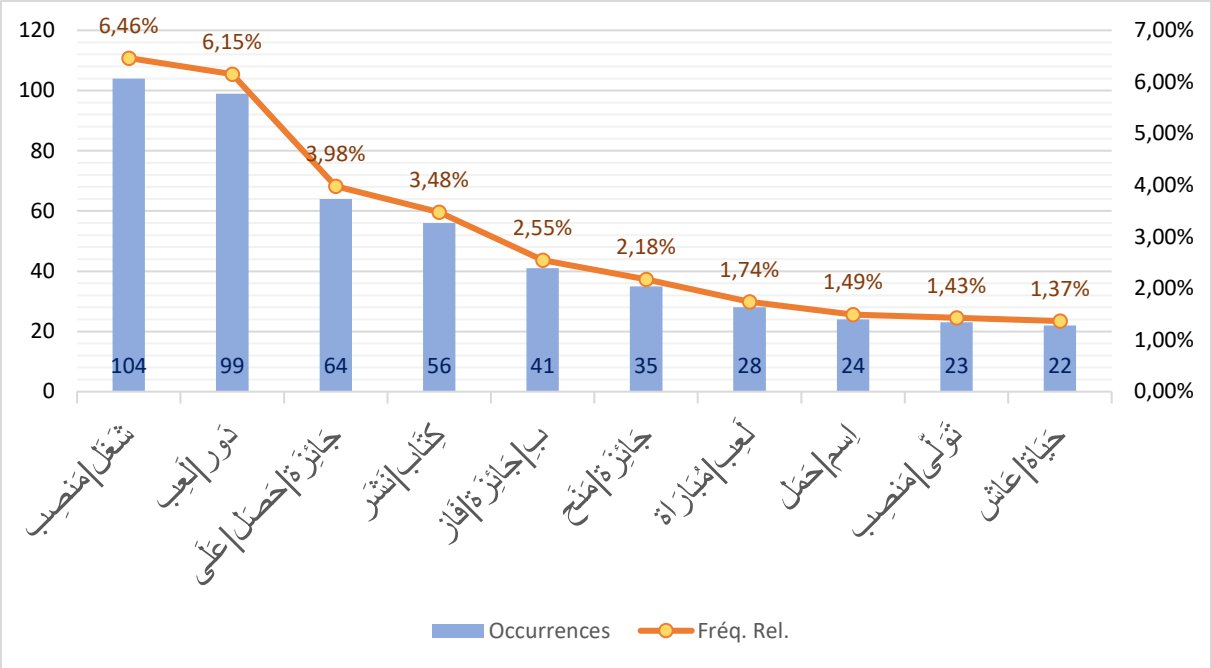

Figure 18 : Collocations arabes les plus fréquentes (sous-corpus WikiMatrix)

Le sous-corpus WM présente une fois encore la plus grande diversité de patrons avec 339 patrons différents, témoignant de la diversité des sujets abordés. La collocation la plus fréquente est différente de celles observées dans les deux autres langues : شَغَلَ|مَنصِب (*šaġala|manṣib*, « *occuper|position* ») avec 104 occurrences (fréquence relative de 6,46%). L'équivalent français *(occuper|poste)* n'apparaît qu'à 45 reprises et l'équivalent anglais *(hold|position)* qu'à 26 reprises. Notons par ailleurs que le 9e patron le plus fréquent, تَوَلَّى|مَنصِب (*tawallā|manṣib*, « *occuper|poste* ») est une variante synonymique. Les patrons complétant les 3 collocations les plus fréquentes sont دَور|لَعِبَ (*dawr|laʿiba*, « *rôle|jouer* ») avec 99 occurrences (fréquence relative de 6,15%) et جَائِزَة|حَصَلَ|عَلَى ((*ğāʾiza|ḥaṣala|ʿalā*, « *recevoir|prix* ») avec 64 occurrences (fréquence relative de 3,98%). Cette dernière a également un équivalent quasi-synonymique avec le 5e patron le plus fréquent : بِـ|جَائِزَة|فَازَ (bi-|ğaʾiza|fāza, « *remporter|prix* ») avec 41 occurrences (fréquence relative de 2,55%).

Du côté des collocations spécifiques à ce sous-corpus, on n'en retrouve qu'une seule significative : أَلبوم|أَصدَر (*ʾaṣdara|ʾalbūm*, « *émettre|album* ») avec pas moins de 17 occurrences pour une fréquence relative de 1,06%.

12.3. Bilan

En résumé, dans cette section, nous avons traité de la méthodologie employée pour mener à bien la projection des annotations de notre corpus anglais vers son équivalent parallèle arabe. Elle a été sensiblement la même que celle employée pour la première étape de projection, à quelques différences près. Nous avons dû tout d'abord nous servir de GIZA++ pour générer une table de traduction bilingue anglais-arabe, car une telle ressource n'était pas fournie par ZAP. Ce dernier outil, après avoir effectué quelques modifications nécessaires dans le code source, nous a à nouveau permis de générer des alignements lexicaux, eux-mêmes exploités pour projeter les annotations du corpus anglais vers le corpus arabe lorsqu'un équivalent était trouvé. À des fins expérimentales, nous avons généré une deuxième table de traduction bilingue

français-arabe en utilisant l'anglais comme langue pivot, afin d'augmenter le nombre de tokens projetés sur le corpus arabe : si un token n'avait pas trouvé son équivalent de l'anglais vers l'arabe, peut-être qu'il en trouverait un du français vers l'arabe. Cette manipulation nous a permis de comparer les deux méthodes, et de savoir si une projection nourrie depuis deux sources pouvait être bénéfique. Il s'avère que les résultats obtenus sur un échantillon de 500 phrases pour lesquelles nous avons effectué une projection manuelle ne sont pas meilleurs avec une double projection : nous anticipions une perte en précision et un gain en rappel, mais la perte est plus importante que le gain, rendant le résultat final moins bon. La projection simple obtient une F-mesure de 27,27 au niveau de l'EP (contre 27,10 pour la double) et 53,38 au niveau du token (contre 53,31 pour la double). Comme pour la projection depuis le français, les résultats chutent encore sur le corpus complet, creusant un peu plus l'écart entre les deux types de projection.

Le nettoyage du corpus avec les annotations projetées a été très similaire à la première, mais a entraîné quelques surprises. Nous avons tout d'abord extrait les patrons de parties du discours des annotations projetées pour ne conserver que les patrons verbo-nominaux. Il s'avère que presque la moitié des projections complètes avaient un patron entièrement nominal. En effet, beaucoup des collocations verbo-nominales en français et en anglais sont traduites par une annexion (syntagme nominal avec complément du nom) en arabe. Nous détaillons ce phénomène dans la section 13.2.1. En outre, un certain nombre de collocations comportaient un élément dont l'étiquette du partie du discours était `X`. Ces étiquettes étaient le fruit d'erreurs de traitement dans la conversion du texte brut au format CoNLL (mauvaise tokenisation, mauvaise lemmatisation, etc.). Ce constat a donné lieu à une étape de normalisation supplémentaire. Nous avons par ailleurs extrait les patrons de collocations (sous forme de tuples de lemmes), puis généré des expressions régulières afin de retrouver les expressions qui n'auraient pas été complètement annotées, voire pas annotées du tout. Nous aurions voulu ensuite vérifier le score d'association des composants des collocations annotées, mais l'outil *Word Sketch* de *Sketch Engine* pour l'arabe nous fournissait des résultats difficilement exploitables. À la place, nous avons fait appel à l'avis d'un expert en langue arabe pour repérer ce qui pouvait ou ne pouvait pas constituer une collocation verbo-nominale. Ce retour nous a permis de corriger une ultime fois le corpus. Les mêmes remarques que précédemment peuvent être faites : le processus de nettoyage de corpus est extrêmement chronophage et fastidieux. Avec le recul, il aurait fallu que nous anticipions mieux les écueils possibles liés au traitement automatique de la langue arabe, en effectuant par exemple un pré-traitement plus complet du corpus arabe dès le départ (nous avons enlevé les diacritiques du corpus, mais une pré-tokenisation aurait pu sans doute aider), ainsi que sur le corpus ayant servi à créer la table de traduction. Un post-traitement sur le fichier `cupt` aurait dû être envisagé pour nettoyer toutes les erreurs de lemmatisation et d'étiquettes morphosyntaxiques.

Enfin, d'un point de vue linguistique, nous avons pu confirmer qu'un nombre assez important de différences existaient entre l'arabe et les deux autres langues de notre projet. Certains phénomènes liés au processus de traduction demeurent : des occurrences d'encapsulation sont observables, des occurrences d'ellipse également, ou encore de transposition, de reformulation, etc. Ce qui reste le plus frappant cependant est la propension à l'usage des catégories nominales plutôt que verbales, avec une grande tendance à utiliser des

annexions ayant un *maṣdar* comme premier terme lorsque l'équivalent anglais ou français utilise un verbe à une forme non finie. Qui plus est, parmi ces formes non finies, les participes passés anglais et français traduits par un participe passif en arabe ne portent pas une étiquette VERB mais ADJ (voir section suivante).

Dans la section suivante, nous utiliserons toutes les données collectées jusqu'ici grâce à notre corpus parallèle trilingue entièrement annoté en collocations pour mener une étude linguistique contrastive. Dans un premier temps, nous l'aborderons sous l'angle quantitatif, en commentant les résultats obtenus en calculant les diverses distances entre les composants des collocations et en calculant la proportion des expressions continues et discontinues. Dans un second temps, nous l'aborderons sous l'angle qualitatif en nous penchant notamment sur le caractère nominal de l'arabe en comparaison des deux autres langues, avant de discuter des phénomènes liés au processus de traduction, d'une part quand cette dernière est de bonne qualité (encapsulation, ellipse, etc.), et d'autre part quand elle est de moins bonne qualité (calque, appauvrissement lexical, etc.).

13. ETUDE LINGUISTIQUE CONTRASTIVE TRILINGUE MULTI-GENRE

Arrivé au bout de ce travail d'annotation, nous disposons de trois corpus parallèles entièrement annotés en collocations verbo-nominales. Ces supports vont finalement nous permettre de mener une étude linguistique contrastive trilingue multi-genre. Dans cette partie, nous tâcherons de comparer à la fois les différences dans l'usage des collocations entre les langues, mais également entre les genres. Dans un premier temps, nous aborderons cette comparaison de manière quantitative, en nous attachant notamment à comparer dans quelle mesure la distance entre les composants d'une collocation peut être variable et dans quelle proportion les collocations repérées sont continues ou discontinues. Dans un second temps, nous adopterons une approche plus qualitative en tâchant de comparer les principales différences observées tout au long de ce projet.

13.1. Etude quantitative

Il nous a semblé intéressant, avec toutes les données dont nous disposons à la fin de ce processus d'annotation, de mesurer principalement deux choses : à quel point la distance entre les composants d'une collocation variait d'une langue à l'autre et d'un genre à l'autre, mais aussi comparer l'usage des collocations continues et discontinues.

13.1.1. Distance entre les composants d'une collocation

Au cours du travail d'annotation, que ce soit dans le cadre de l'annotation entièrement manuelle du corpus d'entraînement initial ou dans celui du nettoyage de chacun des trois corpus après annotation / projection automatique, nous avons pu nous rendre compte que la distance entre les composants d'une collocation semblait varier grandement. Nous nous sommes donc intéressé à ces chiffres-là afin de les comparer pour cette dernière étape du projet. Le tableau suivant résume les résultats obtenus, en proposant la distance[29] minimum, maximum et moyenne pour chaque genre (c'est-à-dire chaque sous-corpus) et chaque langue :

Corpus	**GV**			**TED**			**UN**			**WM**			**Complet**		
Stats / Langue	**Min**	**Max**	**Moy**	**Min**	**Max**	**Moy**	**Min**	**Max**	**Moy**	**Min**	**Max**	**Moy**	**Min**	**Max**	**Moy**
FR	1	**23**	**2,80**	1	**29**	**2,82**	1	*35*	**3,36**	1	**26**	**2,99**	1	*35*	**3,08**
EN	1	20	*2,54*	1	15	2,71	1	**42**	*2,91*	1	*25*	2,93	1	**42**	*2,80*
AR	1	*17*	2,77	1	*11*	*2,07*	1	39	3,30	1	**26**	*2,82*	1	39	2,92
TRI	1	23	2,70	1	29	2,53	1	**42**	**3,19**	1	26	2,91	1	42	2,93

Tableau 21 : Distances minimum, maximum et moyenne entre les composants d'une collocation

Plusieurs choses peuvent être dites au sujet de ces chiffres. Tout d'abord, tous les corpus dans l'ensemble des langues possèdent *a minima* une collocation continue, c'est-à-dire avec

[29] Pour les collocations à 2 tokens, la distance a été calculée en soustrayant l'indice de position du composant 2 à celui du composant 1. Pour le segment *Il a commis un horrible crime*, le premier token ayant pour indice 0, le calcul serait 5 *(crime)* – 2 *(commis)* = 3, donc la collocation est discontinue.
Pour les collocations à 3 tokens (comme les verbes à particules), la distance a été calculée en soustrayant l'indice de position du composant 3 à celui du composant 1, différence à laquelle nous avons soustrait encore 1. Ainsi, pour le segment *He was entrusted with tasks he never could have imagined*, le premier token du segment ayant pour indice 0, le calcul serait 4 *(tasks)* – 2 *(entrusted)* – 1 = 1, donc la collocation est continue.

deux tokens directement contigus. Nous verrons plus loin dans quelle proportion ces expressions sont utilisées.

De plus, en ce qui concerne les distances maximum, on remarque sans grande surprise que c'est le corpus UN qui présente la collocation dont les composants sont les plus éloignés l'un de l'autre, avec une distance de 42 tokens dans le corpus anglais. C'est bien supérieur aux valeurs maximales des autres corpus, ce qui peut s'expliquer de plusieurs façons. D'une part, les phrases de ce corpus sont bien plus longues que celles des autres corpus, avec une moyenne de 38,61 tokens (les 3 langues confondues) contre 30,34 tokens pour GV, 30,11 tokens pour WM, et seulement 19,66 tokens pour TED. D'autre part, les phrases sont beaucoup plus alambiquées, avec des incises et des relatives beaucoup plus nombreuses et beaucoup plus longues que dans les phrases des autres corpus. Pour référence, voici l'exemple dont la distance entre les tokens composant la collocation est de 42 (UN:13394) :

> *Although no specific [**action]_1:COLL** to review the fundamental factors that negatively affect the observance of the principles of national sovereignty and non-interference in the internal affairs of States in their electoral processes, as requested by the General Assembly in its resolution 47/130, was [**taken]_1** by the Commission at that session, references were made, in a number of resolutions, to the issue of elections in the context of guaranteeing the free expression of the will of peoples and ensuring respect for national sovereignty and non-interference in the internal affairs of the States concerned.*

C'est toujours ce même corpus qui a la distance moyenne la plus élevée entre les quatre sous-corpus, avec notamment une distance moyenne de 3,36 tokens pour le français, la plus élevée de toutes les langues, tous genres confondus. Cette observation-là peut se justifier par le fait que le français est la langue la moins synthétique des trois langues considérées au cours de ce projet. En effet, sur l'ensemble des corpus parallèles, le corpus français a environ 13% de tokens de plus que l'anglais et 13,5% de plus que l'arabe (voir Tableau 9). *De facto*, il semble logique que ce soit le français qui ait la moyenne la plus élevée des trois langues sur la distance moyenne. Il est cependant intéressant de noter que l'anglais est la langue dont la distance moyenne est la plus basse. En effet, considérant que l'article défini arabe n'est jamais compté comme un token à part entière (الـ *al-* est toujours suffixé au mot qu'il définit) et que le corpus arabe compte légèrement moins de tokens que son équivalent anglais, cette observation peut sembler paradoxale. On peut imaginer qu'un des facteurs responsables de cette différence soit l'ordre VSO de la phrase en arabe, augmentant automatiquement la distance entre le verbe et son objet en comparaison de l'ordre SVO canonique en anglais.

Enfin, d'après ce tableau, on est en droit de supposer que le genre oral des conférences TED est celui faisant l'usage le plus direct des collocations verbo-nominales. En effet, avec une distance moyenne de 2,53 toutes langues confondues, elle est plus basse que celle du corpus journalistique (GV) et de celle du corpus encyclopédique (WM), sans compter le corpus juridique (UN) dont nous avons déjà parlé.

Ces observations sont intéressantes à plus d'un titre. Par ailleurs, si l'on considère que toutes les moyennes de distance sont supérieures à 2,5 tokens, on peut supposer que les collocations

sont massivement discontinues. Dans la section suivante, nous présentons des résultats quantitatifs relatifs à cette question de discontinuité.

13.1.2. Proportion de collocations continues / discontinues

Par intuition et à la suite de la consultation des résultats de la sous-section précédente, on peut considérer d'emblée que la majorité des collocations sont discontinues. Nous considérons une collocation comme continue lorsque ses composants sont directement contigus. Ne serait-ce qu'en français, on trouvera plus souvent la collocation *poser|question* avec un déterminant quelconque (*une, la, les, cette,* etc.) au milieu. Certaines configurations donneront lieu à des expressions continues, comme *cette affaire pose question* ou *voici la question posée*. Mais qu'en est-il réellement ? Dans le tableau suivant, nous présentons les résultats obtenus pour l'usage des collocations continues et discontinues et la proportion des premières entre les langues et les genres :

Corpus	**Langue** / **Type**	**FR**	**EN**	**AR**	**Complet (moyenne)**
GV	*Continues*	194	362	462	339,33
	Discontinues	1527	1146	776	1149,67
	Proportion continues	*11,27%*	24,01%	**37,32%**	**22,79%**
TED	*Continues*	28	102	242	124,00
	Discontinues	740	643	267	550,00
	Proportion continues	*3,65%*	13,69%	**47,54%**	**18,40%**
UN	*Continues*	382	597	559	512,67
	Discontinues	2493	1680	1427	1866,67
	Proportion continues	*13,29%*	26,22%	**28,15%**	**21,55%**
WM	*Continues*	133	248	572	317,67
	Discontinues	1714	1198	1037	1316,33
	Proportion continues	*7,20%*	17,15%	**35,55%**	**19,44%**
Complet	*Continues*	737	1309	1835	1293,67
	Discontinues	6474	4667	3507	4882,67
	Proportion continues	*10,22%*	21,90%	**34,35%**	20,95%

Tableau 22 : Proportion des collocations continues et discontinues

Le nombre absolu de collocations importe peu dans cette comparaison. En revanche, les lignes indiquant la proportion des expressions continues est celle qui nous intéresse. On constate que les résultats sont réguliers et clivants, surtout entre les langues. Pour les premières, c'est le français qui utilise le moins de collocations continues (10,22% sur l'ensemble du corpus), suivi de l'anglais qui en utilise deux fois plus (21,90%), lui-même suivi de l'arabe, pour qui les expressions continues constituent plus d'un tiers de toutes les collocations annotées (34,35%). En ce qui concerne les genres, c'est le corpus journalistique qui fait l'usage le plus massif des collocations continues avec presque une collocation continue toutes les quatre collocations annotées, suivi de près par le corpus juridique, lui-même talonné respectivement par le corpus encyclopédique puis le corpus de conférences, tous deux avec un peu moins de 20% de collocations continues.

Le premier constat que l'on peut faire est que les collocations sont très majoritairement discontinues, mais le constat le plus intéressant à faire ici est sans nul doute les différences existant entre les langues. On remarque que les différences de proportion d'usage des collocations continues sont stables entre les genres : le français les utilise trois fois moins que l'arabe sur tous les genres (excepté pour TED où les chiffres sont encore inférieurs) et deux fois moins que l'anglais sur tous les genres (même remarque concernant TED). Comment justifier une telle disparité ?

Plusieurs raisons peuvent être évoquées pour ça. D'une part, le déterminant zéro est rare en français et une collocation verbo-nominale sera le plus souvent accompagnée d'un déterminant quelconque. En revanche, il n'est pas rare du tout en anglais, notamment lorsque le nom est indéfini et pluriel (comparons par exemple *to solve problems* et *résoudre des problèmes*) ou encore quand le nom est indénombrable ou *singulare tantum* (comparons par exemple *to provide assistance* et *apporter une / de l'aide*). De plus, le déterminant zéro n'est pas rare non plus en arabe en ce sens que les déterminants indéfinis n'existent pas, l'indéfinitude étant marquée par la désinence casuelle (comparons par exemple حَلَّ مُشكِلَةً *ḥalla muškilat^{an}* et *résoudre un problème*).

En outre, nous l'avons dit, le déterminant défini الـ *(al-)* est toujours préfixé au nom qu'il définit et n'est jamais compté comme un token à part entière. De fait, حَلَّ المُشكِلَة (*ḥalla al-muškila*, « résoudre le problème ») est constitué formellement de deux tokens et non de trois. Dans le même esprit, les modifieurs adjectivaux en arabe sont toujours rejetés après le nom, et même si le français a une grande tendance à faire la même chose, ce n'est pas toujours vrai. Quant à l'anglais, c'est le contraire, et l'adjectif est quasiment toujours inséré avant le nom qu'il modifie (comparons par exemple *to play a crucial role* et لَعِبَ دَورًا حَاسِمًا, *laʿiba dawr^{an} ḥāsim^{an}*, litt. « jouer rôle crucial »). Ainsi, même modifié par un adjectif, les composants de la collocation demeurent contigus.

Enfin, même dans une phrase relative dont le composant nominal de la collocation constituerait l'antécédent, si ce nom est indéfini, l'expression demeurera continue, au contraire des deux autres langues. En effet, en arabe, le pronom relatif n'est présent que lorsque l'antécédent est défini. De fait, quand le français dira *une décision qu'il a prise* et l'anglais *a decision (that) he made*, l'arabe dira قَرَارًا اِتَّخَذَهُ (*qarār^{an} ittiḫaḏa-hu*) : les composants de la collocation sont contigus. L'ordre VSO de la phrase arabe garantit encore un peu plus cette contiguïté, car il devient OVS dans une relative dont l'antécédent est indéfini : جَرِيمَةً اِرتَكَبَهَا الرَئِيس (*ǧarimat^{an} irtakaba-hā al-raʾīs*, « un crime qu'a commis le président »).

En ce qui concerne les disparités continues / discontinues entre les genres, nous pouvons justifier les chiffres importants du corpus des Nations Unies car le style employé use et abuse des participes passés à valeur adjectivale. La voix passive en général est énormément usitée dans ce corpus, mais les tokens ne sont alors plus contigus, ni en français, ni en anglais, ni même en arabe lorsque c'est la tournure « impropre » avec تَمَّ *(tamma)* qui est utilisée (voir section 12.1.5.2). En revanche, les participes passés à valeur adjectivale, notamment employés dans des tournures avec compléments d'agent, se placent tout de suite après le verbe, comme *the resolutions adopted by X* ou *le rôle joué par Y*. Nous l'avons vu, les expressions similaires

en arabe ont le participe passé (ou plutôt « participe passif » dans le cas de l'arabe) étiqueté ADJ (voir section 12.1.5.3) et n'entrent donc pas en considération.

A contrario, le discours des conférences TED est beaucoup plus ancré dans la voix active et le niveau de langue est plus relâché que celui des autres corpus. La dimension orale du canal de communication fait que les phrases sont plus courtes, plus directes, et construites beaucoup plus simplement que peuvent l'être les rapports des Nations Unies. Après avoir éprouvé quantitativement nos données, attelons-nous maintenant à une analyse qualitative.

13.2. Etude qualitative

Dans cette section, qui sera la dernière de notre étude, nous discuterons principalement des différences que nous avons déjà pu observer et desquelles nous avons déjà discuté brièvement dans les typologies dressées après chacune des projections effectuées (voir sections 11.1.4 et 12.1.5). Cette fois-ci, nous tâcherons d'entrer un peu plus dans les détails en nous basant cette fois sur les trois langues de ce projet.

13.2.1. Caractère nominal de l'arabe

Peut-être que l'élément le plus frappant dans ce projet trilingue a été de voir à quel point l'arabe a tendance à utiliser beaucoup plus de substantifs que le français et l'anglais. En effet, nous l'avons constaté après la projection des annotations du corpus anglais vers le corpus arabe, bon nombre des collocations verbo-nominales du premier trouvaient un équivalent entièrement nominal dans le second. Comme nous l'avons montré, après la projection automatique, plus de 45% des collocations projetées complètement avaient un patron entièrement nominal. En s'y penchant un peu plus près, ceci est notamment vrai lorsque le verbe dans la langue-source (dans ce cas-là, l'anglais) est à une forme non finie, notamment à l'infinitif. Considérons l'exemple suivant (GV:10640) :

FR : *Dans les monarchies du Golfe, on n'a pas fini de **[répondre]_1:COLL [à]_1** cette **[question]_1**.*

EN : *In the case of the Gulf monarchies, we might need a bit of time to **[answer]_1:COLL** such a **[question]_1**.*

AR : فِي حالَةِ دُوَلِ الخَلِيجِ قَد نَحْتاج بِعَضِّ الوَقْتِ **لِلإِجابَة عَلَى هٰذا السُؤالُ**.

En français et en anglais, les collocations sont annotées car elles répondent toutes les deux aux critères que nous nous sommes imposés. Dans les deux cas, le verbe est à l'infinitif et la proposition dont ils font partie sont des subordonnées infinitives : la première est une complétive, objet du verbe *finir*, tandis que la seconde est une circonstancielle, complément de but du verbe *need*. Ceci nous amène à conclure que fondamentalement, ces propositions jouent un rôle nominal dans la phrase. Ainsi, en arabe, cette ambiguïté de statut n'existe plus : le segment en gras لِلإِجابَة عَلَى هٰذا السُؤالُ (*li-l-ʾiǧāba ʿalā hāḏā al-suʾāl*, « pour la réponse à cette question ») utilise directement le *maṣdar* (pour rappel, le nom d'action) associé au verbe concave de forme IV أَجَابَ عَلَى (*ʾaǧāba ʿalā*, « répondre à »). Cependant, cet usage s'explique relativement aisément du fait que l'infinitif n'existe pas en arabe : c'est en effet la forme conjuguée à l'accompli de la 3e personne du singulier qui est donné par convention, notamment dans les ouvrages de grammaire et les dictionnaires, pour traduire un infinitif. Dans un cas

comme l'exemple précédent, il est donc bien plus naturel d'utiliser le *maṣdar* plutôt que de reprendre le sujet de la proposition principale (c'est-à-dire la 1ère personne du pluriel), ce qui donnerait لِنُجيبَ عَلَى هٰذا السُؤالُ (*li-nuǧība ʿalā hāḏā al-suʾāl*, « pour que nous répondions à cette question »).

Cet usage ne se limite pas qu'aux infinitifs, mais également aux formes participiales, qui sont également des formes non finies du verbe. Prenons l'exemple suivant[30] (WM:2564) :

FR : *Par exemple, il a trouvé que les garçons étaient « nettement meilleurs » en raisonnement arithmétique, tandis que les filles étaient « supérieures » sur les questions de compréhension.*

EN : *For example, he found boys were "decidedly better" in arithmetical reasoning, while girls were "superior" at* ***[answering]_1:COLL*** *comprehension* ***[questions]_1****.*

AR : على سبيل المثال، وجد أن الأولاد **"**أفضل بشكل كبير**"** في التفكير الحسابي، في حين كانت الفتيات **"**متفوقة**"** **في الإجابة على أسئلة الفهم**.

Remarquons tout d'abord que le français fait tout bonnement l'économie du verbe *répondre*, mais ce qui nous intéresse en premier lieu est la comparaison anglais / arabe. En anglais, le verbe *answer* est dans une forme participiale *(gerund)*, lui conférant un statut proche de celui du nom car, à l'instar de l'infinitif, les participes présents peuvent remplir la fonction de sujet ou d'objet dans une phrase. Dans ce cas précis, la proposition subordonnée participiale joue le rôle de complément de l'attribut *superior*. Une fois encore, fondamentalement, le verbe a un comportement nominal : cette ambiguïté n'est pas présente en arabe. À nouveau, le segment en gras الإجابة على أسئلة الفهم (*al-ʾiǧāba ʿalā ʾasʾilati l-fahm*, « pour la réponse aux questions de compréhension ») se contente d'utiliser le *maṣdar* plutôt qu'utiliser une tournure en reprenant le sujet الفتيات (*al-fatayāt*, « les filles »), lourde et moins naturelle.

Pour terminer sur l'usage du *maṣdar*, nous reprenons un exemple déjà mentionné, dans lequel le *maṣdar* se retrouve dans une annexion (un syntagme nominal avec un complément du nom, appelé الإضَافَة, *al-ʾiḍāfa*) en tant que premier terme (المُضَاف, *al-muḍāf*) complété par un second (المُضَاف إلَيهِ, *al-muḍāf ʾilayh*). Cet usage est sensiblement proche des précédents, si ce n'est qu'il n'y a pas de préposition entre le *maṣdar* et le nom qui le complète (GV:195) :

FR : *De son côté, le blogueur de Bahrain (sic) Ammaro espère que l'affaire ne* ***[ternira]_1:COLL*** *pas la* ***[réputation]_1*** *de Dubaï.*

EN : *On the other hand, Bahraini blogger Ammaro said the case shouldn't be used to* ***[tarnish]_1:COLL*** *Dubai's* ***[reputation]_1****.*

AR : مِن ناحِيَةٍ أُخْرَى, المُدَوَّن البَحْرَيْنِيّ عمارو قالَ أَنَّهُ لا يَجِب اِسْتِخْدامِ هٰذِهِ القَضِيَّةِ **لِتَشْوِيهِ سُمْعَةِ دُبَيّ**.

Une fois encore, on constate que le verbe en anglais est à l'infinitif. En revanche, le verbe en français est bel et bien conjugué, car la traduction n'est pas entièrement fidèle au texte original (*Ammaro said the case shouldn't be used to tarnish* aurait très bien pu être traduit *Ammaro a dit que l'affaire ne devrait pas être utilisée pour ternir*, auquel cas le verbe *ternir*

[30] Les segments en arabe ne sont pas traduits à nouveau ni translittérés. En revanche, lorsqu'il s'agit de les analyser, nous fournissons toujours la translittération, ainsi que la traduction nécessaire à la compréhension pour les lecteurs non-arabophones.

n'aurait pas été conjugué). Le segment en gras تَشْوِيهِ سُمْعَةِ دُبَيّ (*tašwīhi sumʿa Dubayy*, « le ternissement de la réputation de Dubaï ») est un syntagme nominal composé de deux annexions.

Ces cas, où un verbe non fini et son objet sont traduits par un *maṣdar* et un nom dans une annexion, sont les plus nombreux. Cependant, le caractère plus nominal de l'arabe en comparaison du français et de l'anglais ne s'arrête pas là. En effet, un autre cas, relevé dans la typologie des différences observées après la projection automatique de l'anglais vers l'arabe, se rapproche de cette idée. Il inclut une fois encore des verbes non finis dans les deux langues occidentales : les participes passés. Nous l'avons dit, ces formes-là sont étiquetées VERB et s'ils sont associés à un substantif, ils peuvent être candidat au statut de collocations verbo-nominales. En revanche, ces mêmes participes passés sont traduits en arabe par des « participes passifs », étiquetés ADJ. Or, dans la grammaire arabe, même s'il remplit la fonction d'adjectif ou d'attribut, le participe passif est fondamentalement un nom, comme sa dénomination اِسمُ المَفعُول *(ism al-mafʿūl)* l'indique, اِسم *(ism)* signifiant « nom ». Ces traductions ne sont pas systématiques, mais elles arrivent, comme c'est le cas pour le triplet suivant (UN:6062) :

FR : *35. Malgré les **[efforts]_1:COLL** inlassablement **[déployés]_1**, seuls 8 des 26 Etats membres ont versé tout ou partie de leurs quotes-parts au titre de la période de quatre ans allant jusqu'à décembre 1992.*

EN : *35. For the period of four years up to December 1992, despite vigorous **efforts**, remittances in full or part payment of assessed contributions to the Institute have only been received from eight member States out of 26.*

AR : ٣٥ - وَبِالنِسْبَةِ لِفَتْرَةِ السَنَواتِ الأربع حَتَّى كانُونِ الأول/دِيسَمْبِر ١٩٩٢، فَإِنَّهُ بِالرَغْمِ مِن **الجُهُودِ** النَشِيطَة **المَبْذُولَةِ**، لَم يَتِمّ تُلْقِي مَبالِغَ بِالكامِلِ أَو مَدْفُوعاتِ جُزْئِيَّةٍ مِن الأنصبة المُقَرَّرَةِ لِلمَعْهَد إلا مِن ثَمانِي مِن الدُوَلِ الأعضاء المُشارَكَةِ فِيهِ، البالِغِ مَجْمُوعُها ٢٦.

On remarque qu'en anglais, aucun verbe n'est utilisé avec *efforts* : il constitue un cas d'ellipse très récurrent. Cependant, ce qui nous intéresse ici est la comparaison français / arabe. On constate que le participe passé *déployés* est utilisé comme un adjectif qualificatif malgré son étiquette de VERB ; en arabe, le participe مَبذُولَة (*mabḏūla*, « déployés ») est bien placé après le nom qu'il qualifie, à savoir جُهُود (*ǧuhūd*, « efforts »). Il remplit bien la fonction d'adjectif qualificatif, malgré son statut ambigu entre le nom et l'adjectif. Cependant, aucun doute sur le fait que cette association ne peut prétendre au statut de collocation verbo-nominale.

Pour terminer, rappelons rapidement, même si les exemples sont nombreux, que la tournure passive avec تَمّ *(tamma)* renforce encore un peu plus le caractère nominal lié à l'arabe, en témoigne le très bref exemple suivant (TED:14824) :

FR : *Donc, en plein dans le mille, le **[problème]_1:COLL** a été **[résolu]_1**.*

EN : *So bingo, **[problem]_1:COLL [solved]_1**.*

AR : وَعَلَيهِ فَقَد وَجَدَها وَتَمَّ **حَلّ المُشْكِلَةِ**

Le segment en gras est en fait une annexion définie dont le premier terme est le *maṣdar* de première forme حَلّ (*ḥall*, « résolution ») dérivé du verbe حَلَّ (*ḥalla*, « résoudre ») et le second terme le nom مُشكِلَة (*muškila*, « problème »). Le verbe تَمّ *(tamma)* rend la tournure à la voix passive, qu'on pourrait traduire littéralement par « a eu lieu la résolution du problème ».

La conclusion que l'on peut en tirer, c'est que tous ces triplets sont les illustrations que ce qui est un verbe dans une forme non finie pour le français ou l'anglais n'est autre qu'un nom pour l'arabe. Ce statut de verbe pour les deux premières langues apparaît finalement ambigu tant leurs fonctions syntaxiques sont fondamentalement celles que remplissent habituellement les noms et les adjectifs. Cette ambiguïté n'existe pas en arabe et, bien que ce soit un tout autre sujet, cette approche de la grammaire rejoint les théories en grammaire cognitive, notamment celles de Langacker, qui considère qu'un verbe appartient proprement à la catégorie des verbes uniquement lorsque son contenu permet de capter un événement sous l'angle séquentiel ; lorsque le même contenu s'avère sommatif, ce verbe a cessé d'en être un et appartient à une tout autre catégorie, c'est-à-dire celle des formes non finies (Langacker, 2008). Ceci justifierait et expliquerait la facilité avec laquelle ces formes verbales occupent des fonctions nominales ou adjectivales.

13.2.2. Phénomènes liés au processus d'une traduction de qualité

Nous l'avons vu précédemment, mais le processus de traduction crée inévitablement des dissonances entre le contenu du texte-source et celui du texte-cible. C'est notamment vrai pour les traductions de qualité, le rendu idiomatique d'une langue n'étant que rarement similaire à celui d'une autre, surtout si lesdites langues appartiennent à des familles différentes. Bien évidemment, les collocations verbo-nominales ne dérogent pas à la règle. Ainsi, des phénomènes tels que l'encapsulation (le contenu sémantique d'une expression polylexicale dans une langue est rendu par un token unique dans une autre), l'ellipse (un des composants de la collocation dans une langue est omis dans la traduction sans perte de sens, sans pour autant que l'élément restant encapsule la sémantique complète de l'expression originale), la transposition (la catégorie grammaticale d'un mot change du texte-source au texte-cible) ou encore la reformulation (l'énoncé-cible a le même sens que l'énoncé-source, mais le reformule partiellement ou entièrement) ne sont pas rares. Certains cas sont réguliers, tandis que d'autres semblent aléatoires.

13.2.2.1. Encapsulation

L'encapsulation, dans ce projet, concerne les collocations qui ne trouvent pas d'équivalent polylexical dans la langue alignée. Elle concerne plus précisément les verbes qui portent la sémantique entière de la collocation verbo-nominale de la langue alignée, qu'elle soit la langue de départ ou la langue d'arrivée. Dans notre triplet de langues, nous pouvons mentionner plusieurs exemples. En voici un premier (TED:23234) :

FR : *Ou si vous êtes paraplégique, comme j'ai **[rendu]_1:COLL [visite]_1** aux gens de chez Berkley Bionic (sic), qui ont développé eLEGS.*

EN : *Or if you're a paraplegic -- I've **visited** the folks at Berkeley Bionics -- they've developed eLEGS.*

AR : أَو إذا كُنْت مَشْلُولاً-- مِثْلَ الأَصْدِقاء الَّذِينَ **زُرْتُهُم** فِي بيركلي لِلتِكْنُولُوجِيا الحَيَوِيَّةِ-- الَّذِينَ طَوَّرا تَقْنِيَّةٍ eLEGS.

La collocation *rendre|visite* en français ne trouve jamais d'équivalent polylexical en anglais et en arabe. En effet, chacune de ces deux langues possèdent un verbe qui encapsulent la sémantique de l'expression française. En anglais, *j'ai rendu visite* est traduit par *I've visited (to*

visit), et en arabe par زُرْتُ (*zurtu*, du verbe concave de forme simple زَار, *zāra*). Chacun de ces deux verbes se suffit à lui-même pour exprimer l'idée de la collocation française.

Ce phénomène ne concerne évidemment pas que le français. Dans l'exemple suivant, ce sont les collocations anglaise et arabe qui sont rendues par un verbe unique en français (GV:23621) :

FR : *Avec d'un côté l'Iran, à majorité chiite, soutenant les Houthis, tandis que l'Arabie saoudite, à majorité sunnite, les* ***bombardent****, les médias ont adopté une vue simpliste d'un conflit qui se réduirait à des conflits inter-religieux entre les deux plus grandes communautés de l'islam.*

EN : *But with Shia Muslim-majority Iran supporting the Houthis and Sunni Muslim-majority Saudi Arabia* ***[dropping]_1:COLL [bombs]_1*** *on them, the press has pushed a simplistic narrative that the conflict boils down to the religious divides between the two largest denominations of Islam.*

AR[31] : لَكِن بِما أَنَّ الأَغْلَبِيَّة الشِيعِيَّةِ المُسْلِمَةِ المُتَمَثِّلَةِ فِي إيرانَ تُسانِد الحوثيين وَالأَغْلَبِيَّة السُنِّيَّةِ المُسْلِمَةِ المُتَمَثِّلَةِ فِي المَمْلَكَةِ العَرَبِيَّةِ السَعُودِيَّةِ **تُلْقِي** عَلَيهِم **القَنابِلِ**، اِتَّخَذَت الصَحافَةِ طَرِيقَةِ سَهْلَةً لِتُقَدِّم فِيها الصِراعِ عَلَى أَنَّهُ نَتِيجَةَ لِإنْقِسامات دِينِيَّةٍ بَيّنَ اِثْنَيْنِ مِن المَذاهِبِ الرَئِيسِيَّةِ فِي الإسْلامِ.

Alors que l'anglais utilise la collocation *drop|bomb* et l'arabe son équivalent أَلقَى|قُنبُلَة *(qunbula|alqā*), le français préfère le verbe *bombarder* qui encapsule bien la sémantique de ces expressions. Ceci n'est cependant pas systématique.

L'exemple suivant a déjà été mentionné, mais il est récurrent et symbolise bien le phénomène d'encapsulation et met en lumière encore une autre différence entre les 3 langues (WM:3351) :

FR : *De là, ils ont cru voir le château d'Aizuwakamatsu en flamme, et* ***se sont suicidés*** *par désespoir.*

EN : *From Iimori Hill they thought they saw Tsuruga Castle on fire, and* ***[committed]_1:COLL [suicide]_1*** *in despair.*

AR (sic) : وَمِنها اِعْتَقَدُوا أَنَّهُم راووا قَلْعَةِ تسوروغا تَشْتَعِل، **فَإِنْتَحَرُوا** فِي يَأْسٍ.

Ici, une collocation verbo-nominale anglaise *(commit|suicide)* est traduite en français par un verbe pronominal réfléchi *(se suicider)* et en arabe par une verbe de forme VIII (إِنتَحَرَ, *intaḥara*).

Nous en avons déjà parlé, mais le phénomène d'encapsulation concerne principalement, dans notre projet, la collocation française *poser|question*, compte tenu de la fréquence importante à laquelle elle est utilisée. Assez souvent, et l'anglais et l'arabe utiliseront un seul verbe pour traduire cette expression : *ask* pour la première et سَأَلَ *(saʾala*, « demander ») pour la seconde. Cependant, comme en témoigne l'exemple suivant (TED:27268), ces changements ne sont pas forcément obligatoires :

FR : *Alors finalement, je vais terminer cette dernière minute en* ***[posant]_1:COLL*** *des* ***[questions]_1*** *sur les compagnies.*

[31] Les tokens en gras sont annotés et constituent une collocation verbo-nominale. Les marqueurs *[terme]_1:COLL [terme]_1* ne sont pas notés ici à cause de la bidirectionnalité du texte.

EN : *So lastly, I'm going to finish up in this last minute or two **asking** about companies.*

AR : أَخِيراً, عَلَى أَنَّ أَنْهَى بِالدَقِيقَة أَو الإِثْنَتَيْنِ المُتَبَقِّيَةُ **بِالسُؤالِ** عَن الشَرِكاتِ

On remarque un phénomène intéressant entre les 3 langues. La langue de départ est l'anglais, et le conférencier utilise le verbe *ask* dans sa forme participiale *asking*. Le passage au français se fait en « désencapsulant » ce verbe pour le décomposer en *poser* et *question*, le verbe demeurant dans sa forme participiale. En revanche, et cela rejoint notre argumentaire plus haut concernant le statut ambigu des formes verbales non finies, l'arabe se contente d'utiliser le nom سُؤَال (*suʾāl*, « question ») sans mentionner de verbe. Nous avons donc affaire d'une part à une collocation verbo-nominale (en français), d'autre part à un verbe encapsulant le sens de cette expression (en anglais) et un phénomène d'ellipse du verbe dont l'économie est tout à fait justifiée car n'altérant pas le sens de la phrase (en arabe). C'est justement ce phénomène d'ellipse verbale, après avoir vu quelques exemples où le substantif n'était pas obligatoire, que nous abordons dans la sous-section suivante.

13.2.2.2. Ellipse

Les cas d'ellipse, qui concerne, dans le cas des collocations verbo-nominales, l'omission du composant verbal, ne sont pas rares. Ils ne concernent par ailleurs pas une langue en particulier et ont lieu dans tous les cas de figure concernant les trois langues de ce projet. Le premier exemple ci-après illustre une collocation française dont le verbe est omis à la fois en anglais, mais aussi en arabe (UN:2951) :

FR : *A ce propos, nous rappelons que plusieurs **[accords]_1:COLL** ont été **[conclus]_1** concernant la sécurité et la protection des civils palestiniens dans ces camps de réfugiés.*

EN : *In this regard, we recall several previous **agreements** on the security and safety of the Palestinian civilians in those refugee camps.*

AR : وَنَحْنُ نُشِير فِي هٰذا الصَدَدِ إِلَى العَدِيدَ مِن **الاتفاقات** السابِقَةِ المُتَعَلِّقَةِ بِأَمْنِ وسلامة المَدَنِيّينَ الفِلَسْطِينِيّينَ فِي مُخَيَّماتِ اللاجئين تِلْكَ.

Les substantifs *agreements* et اتفاقات *(ittifāqāt)* ne sont accompagnés d'aucun verbe. Ce phénomène n'est pas rare et justifie en partie le nombre d'annotations supérieur en français par rapport aux deux autres langues. C'est notamment le cas des collocations très fréquentes comme *jouer|rôle* et ses équivalents anglais et arabe, d'où le nombre d'occurrences drastiquement différent (448 en français, 342 en anglais et 202 en arabe). En voici un exemple (UN:2977) :

FR : *Nous nous déclarons en outre fermement résolus à œuvrer pour que cette dernière **[joue]_1:COLL** le **[rôle]_1** qui lui revient dans cette nouvelle phase des relations internationales, dans le domaine de la paix et de la sécurité comme dans la promotion du développement économique et social de tous les peuples.*

EN : *We also declared our readiness to cooperate fully in enabling the United Nations to find its appropriate **role** in the new era of international relations, with regard to peace and security as well as economic and social development.*

AR : كَما أَعْلَنّا اِسْتِعْدادِنا لِلتَعاوُنِ تَعاوُناً تامّاً فِي تَمْكِينِ الأمم المُتَّحِدَةِ مِن أَنَّ تَجِد **دَوْرِها** المُناسِبِ فِي العَصْرِ الجَدِيدِ للعلاقات الدُوَلِيَّةِ فِيما يَتَعَلَّق بِالسِلْم والأمن وَكَذٰلِكَ التَنْمِيَةِ الاقتصادية والاجتماعية.

Le segment anglais comme le segment arabe ne jugent pas utiles de mentionner le verbe associé aux substantifs *role* et دور *(dawr)*. Bien que ce cas soit le plus répandu, ce n'est cependant pas systématique et le contraire est vérifiable également, en témoigne le court exemple qui suit (TED:22050) :

FR : *Le climat n'a aucun **rôle**.*

EN : *That climate **[plays]_1:COLL** no **[role]_1**.*

AR[32] : أَنَّ المَناخ لا **يَلْعَب دَوْراً** فِي تَحْدِيدِ مُسْتَوَى السَعادَةِ

Cette fois-ci ce sont les segments anglais et arabes qui mentionnent les verbes *play* et لَعِبَ *(laʿiba)* avec leurs substantifs respectifs (*role* et دور *dawr*), tandis que le français emploie le verbe *avoir*. Il ne s'agit pas d'une ellipse *stricto sensu*, mais cet exemple illustre le fait que les ellipses ou phénomènes similaires ne sont pas l'apanage du passage du français à l'anglais ou l'arabe.

13.2.2.3. Transposition

La transposition se caractérise par le changement des parties du discours d'un ou plusieurs lexèmes entre la langue-source et la langue-cible. Ces cas de transposition sont nombreux, empêchant bien souvent une collocation verbo-nominale dans la langue-source d'en être une également dans la langue-cible. En voici un premier exemple (WM:5673) :

FR *: Son expertise de recherche comprend le développement de microbiocides (sic) luttant contre les **maladies** sexuellement **transmissibles** et le développement de vaccin contre le VIH.*

EN : *Her research expertise involves developing microbicides for sexually **[transmitted]_1:COLL [diseases]_1** and HIV vaccines.*

AR : وَتَشْمَل خِبْرَتَها البَحْثِيَّة تَطْوِير مُبِيدات ميكروبيه لِلأَمْراض المَنْقُولَةِ عَن طَرِيقِ الإِتِّصالِ الجِنْسِيّ ولقاحات لِفَيْرُوس نَقْصِ المَناعَةِ البَشَرِيَّةِ.

Ici, la collocation anglaise *transmit|disease* est bel et bien verbo-nominale, bien que le participe passé ait une fonction adjectivale par rapport au substantif. En revanche, le français utilise cette fois un véritable adjectif qualificatif *(transmissible)* et l'arabe un participe passif (مَنقُولَة *manqūla*, « transmise »). On remarquera la distinction sémantique, légère certes, entre l'adjectif français et les participes anglais et arabes. Toujours est-il qu'un phénomène de transposition est en effet observable. Voici un autre exemple témoignant du même phénomène (UN:2814) :

FR : *C'est pourquoi il importe d'examiner l'attitude de la communauté internationale face à ces **[problèmes]_1:COLL** urgents et d'envisager de nouveaux moyens de les **[résoudre]_1**.*

EN : *There is therefore a need to review the international response to these urgent **problems** and to discuss new approaches to their **solution**.*

AR : فَهُناكَ إِذْنٍ حاجَةٍ إِلَى إعادَةِ النَظَرِ فِي الاستجابة الدُوَلِيَّةِ لِهٰذِهِ **المَشاكِلِ** المُلِحَّةِ وَمُناقَشَة النَهْجِ الجَدِيدَةِ **لِحَلّها**.

[32] Les tokens en gras sont annotés et constituent une collocation verbo-nominale. Les marqueurs *[terme]_1:COLL [terme]_1* ne sont pas notés ici à cause de la bidirectionnalité du texte.

Cette fois-ci, c'est le verbe de la collocation française *résoudre|problème* qui trouve un équivalent nominal en anglais et en arabe (المَشَاكِل لِحَلِّهَا *al-mašākil li-ḥallihā*, litt. « les problèmes pour leur résolution »). Ces cas de transposition sont relativement nombreux, notamment, nous l'avons vu, dans le cas de l'utilisation du *maṣdar* en arabe. En revanche, la configuration la plus répandue est sans doute la reformulation pure et simple.

13.2.2.4. Reformulation

À divers degrés, les segments d'une langue peuvent se trouver formulés différemment dans une autre. Ces reformulations sont évidemment toujours le fait d'un choix de la personne en charge de la traduction, mais ces choix peuvent avoir diverses motivations. Voici une première illustration (GV:9787) :

FR : *Un de ces centres est à proximité d'où je vis et je sais que des hommes de mon village **[rendent]_1:COLL [visite]_1** aux prostituées là bas (sic). Il me l'ont dit.*

EN : *One of these centers is very close to me and I know that men from my village **visit** prostitutes there. They've told me.*

AR : أَحَدُ هٰذِهِ الدور قَرِيبٍ جِدّاً مِن مَسْكَنِي وَأَنا أَعْرِف أَنَّ بِعَضّ الرجالُ مِن قَرْيَتَيْ **يَذْهَبُونَ** هُناكَ، هُوَ قالُوا لِيَ ذٰلِكَ.

Ici, ce n'est pas tant l'encapsulation déjà discutée plus haut de *visit* pour la collocation française *rendre|visite* qui nous intéresse, mais plutôt le segment arabe. Une traduction possible pour le segment complet serait la suivante : « Une de ces maisons est très proche de chez moi et je sais que certains hommes de mon village y vont, ils me l'ont dit ». Finalement, les groupes verbaux *rendent visite aux prostituées* et *visit prostitutes* ne sont restitués en arabe que par le verbe يَذْهَبُون (*yaḏhabūna*, « ils vont »). Deux hypothèses peuvent être formulées : d'un côté, la mention des *prostituées* s'est faite dans le cotexte précédent ce segment, auquel cas la personne en charge de la traduction n'a pas jugé utile de reprendre ce substantif ; d'un autre côté, il est possible que le traducteur ait fait le choix de préserver son lectorat de cette mention jugée potentiellement subversive pour le public arabe. Par ailleurs, ces reformulations peuvent s'avérer le fruit d'un choix purement stylistique (GV:8909) :

FR : *Alors que la révolution syrienne **se noie dans le sang**, de nombreux internautes se sont mis à leurs claviers pour souhaiter que ce soit le dernier.*

EN : *With the Syrian Revolution **[reaching]_1:COLL** its bloodiest **[peak]_1**, many netizens took to their keyboards wishing it would be his last birthday.*

AR : وَمَعَ **بُلُوغِ** الثَّوْرَةِ السُورِيَّةِ **أَوْجِ** الدَمَوِيَّةِ، أَتَّجه العَدِيدَ مِن مُسْتَخْدَمِي الإِنْتِرْنِت لِلَوْحاتٍ الكَمْبيُوتِر لِيَتَمَنُّوا لَهُ أَنَّ يَكُون هٰذا آخَر ذِكْرَى مِيلاد لَهُ.

Quand l'anglais et l'arabe usent de formules similaires, bien que l'arabe utilise une fois de plus une annexion et non pas une collocation verbo-nominale, toutes deux signifiant littéralement « atteindre un sommet sanglant », le français fait un choix stylistique drastiquement différent. Plutôt que de rendre de manière un peu calquée la traduction littérale que nous venons de citer, la personne en charge de la traduction lui a préféré la métaphore *se noyer dans le sang*.

Voici un dernier exemple de reformulation (GV:16812) :

FR : *Dans le passé, un certain nombre **d'attaques** avait visé les propriétés possédées par des Ahmadis. La publication de la liste des adresses met donc les membres de cette communauté en grand danger.*

EN : *In the past, a number of **[attacks]_1:COLL** have been **[launched]_1** on properties owned by Ahmadis. So the list of addresses puts the community members at a great risk.*

AR : فِي الماضِي **حَدَثَت** عِدَّةِ **هَجَماتِ** عَلَى مُمْتَلَكاتٍ لِأَعْضاءِ الطائِفَةِ الأَحْمَدِيَّة، مِمّا يَجْعَل مِن وُجُودِ عَناوِينهم فِي تِلْكَ القَوائِمِ خَطَراً كَبِيراً عَلَيهِم.

Ce cas est un peu particulier. D'un côté, l'anglais utilise la collocation verbo-nominale *launch|attack*. De l'autre, le français fait cette fois l'ellipse (voir section 13.2.2.2) d'un verbe équivalent à *launch* (p. ex. *lancer*) et transpose la préposition anglaise *on* en un verbe à part entière *(avait visé)*, le rendant par la même plus central. Et en ce qui concerne l'arabe, la formulation est beaucoup plus générique avec le choix du verbe حَدَثَ (*ḥadaṯa*, « avoir lieu »), appauvrissant quelque peu le segment. Après avoir fait un tour d'horizon des phénomènes traductionnels ayant lieu au cours d'une traduction de bonne qualité, ceci nous mène directement à observer d'autres phénomènes liés à une traduction de moindre qualité.

13.2.3. Phénomènes liés au processus d'une traduction de qualité moindre

La traduction est un exercice délicat qui nécessite des connaissances approfondies à la fois dans la langue-source et dans la langue-cible, à telle enseigne que les erreurs et les mauvais choix sont récurrents. Dans cette section, nous donnerons quelques exemples d'appauvrissement lexical, de calques et d'omissions.

13.2.3.1. Appauvrissement lexical

On caractérisera l'appauvrissement lexical par l'utilisation d'un verbe faible au lieu d'un verbe plein qui aurait pu donner lieu à une collocation verbo-nominale courante. Ces cas sont évidemment nombreux, et quand bien même ils ne seraient pas toujours synonymes d'une traduction fautive, ils transmettent cependant un message quelque peu appauvri. Voici un exemple (WM:21777) :

FR : *Les auteurs Marc Weinberg et Leonard Maltin d'Orange Coast Magazine ont critiqué le choix de **faire** un **film** avec de vrais acteurs.*

EN : *Orange Coast Magazine writer Marc Weinberg and Leonard Maltin criticized the decision to **[shoot]_1:COLL** the **[film]_1** in live action.*

AR : وَاِنْتَقَدَ الناقِدانِ مارْك واينبرغ وليونارد ماتلين، مِن مَجَلَّةِ اورنج كوست، قَرارِ **تَصْوِيرَ الفِيلمِ** بِشَكْلٍ حَيّ.

L'arabe n'a pas d'annotation car la collocation anglaise *shoot|film* est rendue par l'annexion تَصوِير الفِيلم (*taṣwīr al-fīlm*, « le tournage du film ») : il ne s'agit donc pas d'un quelconque appauvrissement lexical. En revanche, la même chose ne peut pas être dite en ce qui concerne le français. Quand la collocation *réaliser|film* est tout à fait correcte et envisageable, le traducteur a opté pour le verbe faible *faire*. De fait, le message transmis s'en retrouve quelque peu appauvri.

Si l'appauvrissement du segment original est possible, le contraire l'est tout autant. L'exemple suivant en est la preuve (TED:26938) :

FR : *Ce que nous avons fait ensuite, ou les **[défis]_1:COLL** que nous avons **[relevés]_1**, c'était de coordonner ce mouvement.*

EN : *So, the next thing we did, or the **challenges** we **did**, was to coordinate this movement.*

AR[33] : هٰكَذا، وَالشَيْء التالِي الَّذِي قُمْنا بِهِ، أَو **التَحَدّيات** الَّتِي **واجَهْنا** كانَ لِتَنْسِيقِ هٰذِهِ الحَرَكَةِ.

Le segment original anglais utilise le verbe faible *do* avec pour objet *challenges*. Cet usage est typique de l'oralité des conférences TED et on constate que cette fois la traduction n'appauvrit pas l'original, mais l'enrichit dans les deux langues-cibles : avec *relever*|*défi* en français et وَاجَهَ|تَحَدٍّ (*wāğaha*|*taḥadd*^{in}, « *défi*|*faire front* »).

13.2.3.2. Calques

Les calques sont très nombreux et il est difficile de jauger quand il s'agit d'une mauvaise traduction et quand il s'agit d'une traduction correcte tant ils tendent à devenir la norme. C'est notamment le cas pour l'arabe, dont énormément de patrons de collocations sont des calques du français ou de l'anglais mais qui, à force d'être utilisés, deviennent canons. Historiquement, des verbes existaient pour exprimer une idée, mais ils ont petit à petit été remplacés par des expressions polylexicales calquées sur d'autres langues. Si l'on prend les exemples de قَدَّم|مُسَاعَدَة (*qaddama*|*musāʿada*, « *fournir*|*aide* ») et اِتَّخَذَ|قَرَار (*ittaḫaḍa*|*qarār*, « *prendre*|*décision* », deux collocations très présentes dans le corpus complet (respectivement 188 et 154 occurrences), les verbes سَاعَدَ (*sāʿada*, « aider ») et قَرَّرَ (*qarrara*, « décider ») sont tout à fait valables pour véhiculer la même idée. Sous l'influence de l'anglais notamment, les expressions polylexicales correspondantes tendent à les remplacer.

Dans d'autres cas cependant (et pas forcément qu'en arabe), le recours au calque apparaît clairement comme fautif. C'est le cas de l'exemple suivant (WM:18512) :

FR : *Les efforts de Mirza pour **résoudre** les **disputes** sanglantes en envoyant une pétition par une délégation de deux de ses hommes à Lénine n'ont pas eu de résultat.*

EN *: Mirza's efforts to **[resolve]_1:COLL** the bloody **[disputes]_1** by sending a petition through a delegate of two of his men to Lenin did not result in a resolution.*

AR : لَم تُسْفِر الجُهُودِ الَّتِي بَذَلَها كوجاك خان فِي إطارِ **حَلِّ النِزاعاتِ** الدَمَوِيَّةِ مِن خِلالَ إرْسالُ اِلْتِماسٍ عَن طَرِيقِ وَفْدٌ مُكَوّن مِن رِجْلَيْنِ إِلَى لَيْنَيْنِ عَن تَسْوِيَةٍ.

Alors qu'on a la collocation *dispute*|*resolve* en anglais, on a une traduction mot à mot en français avec l'expression *résoudre les disputes*. Des choix plus idiomatiques auraient pu être faits avec la collocation *différend*|*régler*, par exemple. On remarque par ailleurs qu'en arabe on a la collocation بَذَلَ|جَهد (*baḏala*|*ğahd*, « *dépenser*|*effort* ») en début de segment, alors que l'anglais et le français font l'ellipse du verbe. Concernant le calque mentionné plus tôt, l'arabe n'est pas concerné car l'expression verbale est rendue par l'annexion حَلّ النزاعات (*ḥall al-nizāʿāt*, « résolution des conflits »).

L'exemple suivant est similaire en ce qu'il illustre un calque de l'anglais vers le français (TED:6242) :

[33] Les tokens en gras sont annotés et constituent une collocation verbo-nominale. Les marqueurs *[terme]_1:COLL [terme]_1* ne sont pas notés ici à cause de la bidirectionnalité du texte.

FR : *Par ailleurs, cet appel a eu lieu après qu'il ait (sic)* ***servi*** *sa* ***peine****, il était donc dehors et avait un emploi et prenait soin de sa famille -- il a dû retourner en prison.*

EN : *And by the way, this appeal went through after he had finished* ***[serving]_1:COLL*** *his* ***[sentence]_1****, so he was out and working at a job and taking care of his family and he had to go back into jail.*

AR : وَبِالمُناسَبَة قَد وَضَعَ الِاسْتِئْنافِ بَعدَ أَنَّ كانَ مِيشِيل **قَد أَتِمّ حُكْمه** السابِق وَكانَ قَد بَدَأَ فِي العَمَلِ فِي وَظِيفَتِهِ الجَدِيدَةِ وَالِاعْتِناء بِعائِلَته وَقَد أُعِيد إِلَى السِجْنِ

La collocation anglaise *sentence|serve* est traduite littéralement par *peine|servir*. Un choix plus judicieux aurait pu être l'expression *peine|purger*, bien plus idiomatique. On remarquera par ailleurs l'usage fautif (mais commun) du subjonctif après la conjonction de subordination *après que* au lieu de l'indicatif. L'arabe quant à lui passe par une paraphrase avec le segment قَد أَتِمّ حُكْمه (*qad ʾatamma ḥukmahu,* « il avait déjà terminé sa peine »), ne commettant donc pas de calque, bien qu'il ne s'agisse pas d'une collocation.

13.2.3.3. Omissions

Parfois, pour éviter tout écueil de traduction, la personne en charge de la traduction préfère omettre tout bonnement un segment du contenu source. Ces cas sont rares dans notre corpus, aussi nous contenterons-nous d'un unique exemple. Dans ce dernier, la collocation utilisée en anglais et en arabe est omise de la traduction française (WM:368) :

EN : *During its history, many* ***[efforts]_1:COLL*** *were* ***[made]_1*** *to impede the spread of Islam in Rwanda.*

FR : *Pendant son histoire, beaucoup d'éléments ont entravé l'expansion de l'islam au Rwanda.*

AR : الجَماعاتِ التَبْشِيرِيَّةِ خِلالَ تارِيخِها **بُذِلَت جُهُود** كَثِيرَة لِعَرْقَلَة اِنْتِشارِ الإِسْلامِ فِي رُوانْدا.

Nous sommes à la limite de la reformulation, mais la perte d'information est telle que nous estimons qu'il s'agit là d'une omission. En effet, remplacer les segments *many efforts were made* et بُذِلَت جُهُودِ كَثِيرَة (*buḏilat ǧuhūd kaṯīra,* « beaucoup d'efforts ont été déployés ») par *beaucoup d'éléments* s'apparente à une manière de contourner la traduction. Le message est altéré, car l'agentivité derrière la tentative d'entrave est fortement réduite voire ignorée.

14. CONCLUSION

Dans cette partie, nous avons présenté la méthodologie que nous avons adoptée pour la projection des annotations de notre corpus français vers les équivalents parallèles anglais et arabe. Après avoir décrit le fonctionnement des outils tiers qui nous ont servi pour ce processus de projection, à savoir GIZA++ (Och & Ney, 2003) pour la génération de tables de traduction multilingues et ZAP (Akbik & Vollgraf, 2018) pour la génération des alignements lexicaux, nous avons entrepris d'exposer les solutions envisagées pour pouvoir projeter à proprement parler les annotations du corpus français.

Tout d'abord, cette projection devait avoir lieu du corpus français vers le corpus anglais. Initialement, nous avions prévu d'utiliser ZAP pour faire l'intégralité du travail de projection. Or, nous avons constaté que l'outil ne projetait d'annotations que depuis l'anglais vers une autre langue parmi un panel de langues relativement restreint. En outre, ZAP, dans sa version publiée, ne permet pas de projeter autre chose que des annotations standard. Ceci nous a mené à concevoir une solution pour pallier ce problème-là et réaliser notre projection différemment. Pour ce faire, nous avons exploité la table de traduction bilingue anglais-français fournie avec ZAP, ainsi que son aligneur heuristique pour réaliser des alignements lexicaux entre chaque phrase CoNLL de nos corpus parallèles français et anglais. Pour chaque phrase CoNLL, si un token anglais dans la phrase-cible avait un équivalent dans la table de traduction pour le token français dans la phrase-source, nous créions l'alignement. Tous ces alignements étaient ensuite enrichis d'informations supplémentaires (comme l'identifiant des tokens) pour pouvoir projeter les annotations du token-source vers le token-cible dans le fichier CoNLL.

Le résultat final de cette première projection automatique était encourageant. Après avoir choisi 500 phrases aléatoirement, nous avons réalisé manuellement la projection telle qu'elle aurait dû se faire afin d'en évaluer la qualité. Bien qu'environ un tiers des tokens ne trouvaient pas d'équivalent au cours de la projection, cette dernière obtient une F-mesure de 62,07 au niveau de la collocation et de 74,42 au niveau du token. Nous avons ensuite dressé une typologie des différences observées entre le corpus français et le corpus anglais concernant les projections « fautives », afin de déterminer quels étaient les facteurs menant à une projection erronée ou incomplète. Nous avons ensuite abordé notre méthodologie de nettoyage du corpus projeté automatiquement (correction, augmentation, etc.), présenté les différents résultats obtenus pour chaque sous-corpus et en avons tiré des conclusions.

Ensuite, nous avons fait sensiblement la même chose pour la projection vers le corpus arabe, à quelques détails près. Tout d'abord, nous avons cette fois dû générer une table de traduction anglais-arabe avec GIZA++ et adapter le code-source de ZAP, afin qu'il puisse traiter l'arabe. Une fois ces deux étapes complètes, nous avons réalisé le même travail d'alignement lexical pour chacune des phrases CoNLL de nos corpus parallèles, alignements ensuite enrichis pour finalement projeter les annotations sur le corpus arabe. Dans un but expérimental, nous avons réalisé une « double projection » : pour ce faire, nous avons généré une table de traduction français-arabe en utilisant l'anglais comme langue pivot (à l'aide des tables de traduction français-anglais et anglais-arabe), puis nous avons généré des alignements français-arabe que nous avons enrichis pour finalement projeter les annotations du corpus français qui n'auraient

pas déjà été projetées *via* le corpus anglais. Nous anticipions une perte en précision mais un gain en rappel avec cette méthodologie.

Les résultats des deux projections ont ensuite été évalués de la même manière que nous l'avions fait pour la précédente, et les scores obtenus sont finalement très proches. Bien qu'environ 40% des tokens n'étaient pas projetés pendant la projection simple contre 26% pendant la projection double, les F-mesures sont aux alentours de 27 au niveau de la collocation et 57 au niveau du token. Globalement, les résultats de la projection vers l'arabe étaient moins bons que la projection du français vers l'anglais, et cela s'est expliqué par les différences notables entre ces langues. Nous avons à nouveau discuté de la méthodologie de nettoyage du corpus et les différences observées ont été listées dans une typologie. Puis, nous avons présenté les différents résultats obtenus pour chacun des sous-corpus et en avons tiré des conclusions intermédiaires, avant de mener une étude linguistique contrastive trilingue.

Nous avons séparé cette dernière étude, dernière entreprise de notre projet, en deux parties : une discutant d'aspects quantitatifs concernant nos données et une autre abordant plus spécifiquement les aspects qualitatifs de ces dernières. Pour la première, nous avons calculé les distances qui existaient entre les deux composants des collocations annotées. Ces distances (minimum, maximum et moyenne) ont été calculées en vue de les comparer au niveau des trois langues et au niveau des quatre genres associés à chacun des sous-corpus. Il s'est avéré que la distance minimum était équivalente pour toutes les langues et tous les genres (distance 1, soit des collocations continues), mais les distances maximum et moyenne sont globalement plus importantes en français et dans le genre juridique. Concernant la langue, l'anglais et l'arabe sont bien plus synthétiques que le français et ont besoin de moins de tokens pour exprimer la même chose. Concernant le genre, les textes juridiques usent de formulations souvent alambiquées, avec parfois des incises et des relatives très longues, en atteste la distance « record » de 42 tokens pour une des phrases du corpus UN anglais.

Nous nous sommes également intéressé à déterminer la proportion d'usage des collocations continues et discontinues entre les genres et les langues, et les résultats sont similaires : le genre juridique a tendance à privilégier les collocations continues, et ces dernières sont les plus nombreuses en arabe. En effet, environ 30% des collocations arabes sont continues, contre 20% des collocations anglaises et seulement 10% des collocations françaises. Encore une fois, cela peut s'expliquer par le caractère synthétique des deux premières langues. En outre, l'arabe ne compte jamais l'article défini comme un token et l'article indéfini n'existe pas, rendant automatiquement un grand nombre de collocations continues.

Pour la deuxième partie de notre étude linguistique, nous avons abordé des aspects plus qualitatifs pour bien comprendre les divergences, nombreuses, que peuvent connaître les trois langues de notre projet. Nous avions pu constater bon nombre de ces différences au cours de nos travaux de projection et en avions esquissé les prémisses dans les différentes typologies dressées. Nous avons pu identifier que des phénomènes liés au processus traductionnel avaient lieu dans le passage d'une de ses langues vers les autres. Certains de ces derniers sont dus à une traduction de bonne qualité, tandis que d'autres sont le fruit d'une traduction de moindre qualité. La différence la plus frappante cependant demeure dans le caractère nominal prédominant en arabe en comparaison du français et de l'anglais. En effet, de manière quasi-automatique, les

verbes de forme non finie (participes, infinitifs) dans ces deux langues trouvaient un équivalent nominal en arabe, bien souvent un *maṣdar*. Ainsi, bon nombre des collocations verbo-nominales françaises et anglaises étaient en fait des annexions, c'est-à-dire des syntagmes nominaux avec complément du nom. Ceci expliquait en outre les résultats assez bas de la projection vers l'arabe.

En ce qui concerne les autres phénomènes de traduction, les collocations verbo-nominales ne sont pas en reste vis-à-vis d'autres faits de langue et nous pouvons en citer quelques-uns prédominants. Tout d'abord, les phénomènes d'encapsulation et d'ellipse sont récurrents : le premier concerne les verbes encapsulant la sémantique d'une collocation dans une autre langue (comme *ask* pour *poser une question*), le second concerne les verbes tout simplement omis car parfois inutiles à répéter (comme *role* pour *jouer un rôle*). Ce sont deux phénomènes qu'on observe régulièrement, avec bien souvent le français qui ressent le besoin d'utiliser plusieurs tokens tandis que l'anglais et l'arabe n'en utilisent qu'un seul. L'encapsulation, par exemple, est possible en arabe grâce aux diverses formes augmentées des verbes (comme اِنتَحَرَ *intaḥara* pour *to commit suicide*). L'ellipse est régulière également en anglais et en arabe, notamment dans le cas des collocations les plus fréquentes. Par ailleurs, les phénomènes de transposition sont aussi récurrents. Outre l'usage massif des annexions en arabe, dont le composant verbal de la collocation française ou anglaise trouve un équivalent de fait nominal, d'autres transferts de parties du discours ont pu être observés. Par exemple, des participes passés dans une langue trouvent parfois un équivalent purement adjectival dans l'autre : c'était le cas de *transmitted diseases* en anglais, dont la traduction française était *maladies transmissibles*. Cependant, les cas les plus fréquents sont liés à une reformulation partielle ou complète du segment. Ces choix sont parfois justifiés par une approche cibliste (pour la préservation ou le respect des mœurs du public visé) ; ils le sont parfois par des effets de style, avec des métaphores préférées à des formulations plus pragmatiques ; ou plus généralement, ils le sont pour se détacher du segment original et le moduler, pour éviter tantôt le calque, tantôt l'omission ou la mauvaise traduction.

Ce sont ces derniers « vices » de traduction, auxquels nous avons ajouté un appauvrissement lexical récurrent, qui ont conclu notre étude contrastive. Parmi ces écueils, nous avons pu constater la présence importante de choix lexicaux appauvrissant le message original (ou parfois l'enrichissant, selon le sens de traduction). Bien qu'une expression idiomatique existe, la traduction choisit de remplacer le verbe plein de la collocation originale par un verbe faible ou un verbe faible. Cela arrive dans tous les sens de traduction, aussi bien en français avec des verbes comme *faire* (et ses proches synonymes), *être* ou *avoir*, qu'en anglais avec *do*, *be* ou *have*, et en arabe avec différents verbes signifiant « faire », « réaliser » ou « effectuer » : قَامَ بِـ *(qāma bi-)*, فَعَلَ *(faʿala)* ou encore أَجرَى *(ʾağrā)*. Dans une moindre mesure, certaines traductions étaient difficilement acceptables voire fautives à cause de calques, comme **servir une peine* plutôt que *purger une peine* pour l'expression anglaise *serve a sentence*. Nous avons cependant fait remarquer que de nombreuses collocations arabes, devenues canoniques et relativement bien présentes dans notre corpus, n'étaient au départ que des expressions calquées sur des langues « occidentales ». Dans une moindre mesure encore, nous avons constaté que la personne en charge de la traduction pouvait parfois omettre tout simplement la collocation du segment original.

En conclusion, le travail de projection n'est pas une mince affaire. Bien qu'elle soit utile, la méthodologie que nous avons utilisée n'en demeure pas moins perfectible sur bien des points. Cependant, il apparaît évident, après avoir mené cette modeste étude linguistique contrastive, que l'étape de projection ne pourrait de toute façon pas être parfaite. En effet, qu'il s'agisse des différences intrinsèques aux langues étudiées ou des procédés traductionnels entrant en jeu au cours du passage d'une langue à l'autre, la projection reste entièrement dépendante du contenu des textes parallèles. Il n'en demeure pas moins que cette projection faillible soit paradoxalement une des clés pour observer les différences existant entre les langues. C'est en effet en comparant ce qui a été annoté après projection ou pas qui nous a permis de dresser des typologies et d'en déduire des récurrences que nous pouvions par la suite analyser. Malgré cela, nous sommes bien conscient que la portée de ce projet reste relativement limitée à plusieurs égards. Dans la dernière partie de ce travail, nous concluons en résumant l'intégralité du projet avant d'en aborder les apports, puis les limites et les perspectives d'amélioration.

IV. CONCLUSION

15. RESUME COMPLET

Arrivé à la fin de ce mémoire, nous pouvons dresser un bilan complet de ce que nous avons proposé à travers ce travail. Tout d'abord, la première phase a consisté en la constitution d'un un corpus parallèle trilingue (français, anglais et arabe) multi-genre. Pour ce faire, nous avons exploité les corpus suivants : le corpus *Global Voices v2018q4* pour des textes issus d'articles journalistiques, le corpus *TED2020 v1* pour des textes issus de conférences TED, le corpus *United Nations Parallel Corpus v1.0* pour des textes issus de rapports des Nations Unies, et le corpus *WikiMatrix v1* pour des textes issus d'articles Wikipedia. De ces corpus, nous avons extrait en moyenne 25 000 phrases alignées dans les trois langues afin de constituer des « tritextes », pour un total légèrement à 100 000 phrases.

La deuxième phase de ce travail consistait à utiliser un outil d'annotation automatique des expressions polylexicales verbales pour annoter les collocations dans notre corpus. Pour réaliser cette opération, nous avons repris le corpus d'entraînement utilisé pendant la *Shared Task 1.1* de PARSEME (Ramisch et al., 2018), corpus que nous avons modifié pour que les annotations qui y figuraient soient remplacées par les nôtres. Ces dernières devaient respecter le cadre théorique dans lequel nous avons inscrit notre travail, ainsi que les directives présentes dans le guide d'annotation rédigé en amont (voir Annexe A). Ce dernier se base à la fois sur les travaux de PARSEME (Savary et al., 2015) et sur ceux du projet SimpleApprenant (Todirascu & Cargill, 2019). Nous nous sommes concentré exclusivement sur les collocations verbo-nominales. Ces dernières doivent avoir au moins un élément verbal (autre qu'un verbe faible comme *faire*) et un élément nominal (comme *jouer + rôle*), ces deux éléments doivent être en relation de dépendance syntaxique (comme *jouer un rôle*, en relation verbe-objet), l'expression doit pouvoir être mise à la voix passive (comme *un rôle est joué*), le déterminant doit pouvoir être modifié (comme *jouer le rôle*), un modifieur adjectival et / ou adverbial doit pouvoir être inséré (comme *jouer un rôle bouleversant*), et l'information mutuelle des deux composants doit être assez élevée. Le corpus d'entraînement de 17 225 phrases, une fois modifié, comportait un total de 958 collocations annotées.

L'outil que nous avons utilisé pour l'annotation automatique a été VarIDE (Pasquer et al., 2018), notamment parce que cet outil a obtenu de bons résultats pour le français au cours de la *Shared Task 1.1* de PARSEME. Pour évaluer la qualité de l'annotation automatique, nous avons soustrait 500 phrases aléatoires du corpus d'entraînement afin de constituer un jeu de test, puis nous avons comparé la sortie annotée automatiquement avec la sortie annotée manuellement. Sur cet échantillon, le système obtient une très bonne précision (89,47), un rappel un peu plus faible (73,91), pour une F-mesure de 80,95. Une fois notre corpus français complètement annoté automatiquement, une longue phase de nettoyage de corpus a été nécessaire : correction des annotations incorrectes ou incomplètes, ajout des annotations manquantes, etc. Finalement, ce corpus contient 7211 annotations au total, réparties inéquitablement entre les quatre sous-corpus. Nous avons remarqué qu'au plus la langue utilisée était normée, au plus les collocations étaient nombreuses, ce qui fait que le sous-corpus des Nations Unies contient bien plus d'annotations que les autres sous-corpus. *A contrario*, la dimension orale des conférences TED réduit considérablement la fréquence d'usage des collocations. Deux collocations se détachaient largement du reste : *mesure*|*prendre* et *jouer*|*rôle*, la première majoritairement

présente dans les rapports juridiques du sous-corpus des Nations Unies, la seconde avec une distribution moins contrastée entre les sous-corpus.

La troisième phase de ce travail, et non la moindre, concernait la projection des annotations obtenues semi-automatiquement pour le français vers les corpus parallèles anglais et arabe. Pour réaliser cette opération, nous avons exploité GIZA++ (Och & Ney, 2003) pour créer des tables de traduction bilingues et ZAP (Akbik & Vollgraf, 2018) pour générer des alignements lexicaux entre les phrases des corpus parallèles. Nous avons commencé par la projection du français vers l'anglais, en utilisant une table de traduction fournie par ZAP. Nous anticipions de pouvoir projeter nos annotations directement avec ZAP, mais cet outil ne fait de projection que depuis l'anglais vers une autre langue, et ne projette pas d'annotations autres que les annotations classiques (étiquettes morphosyntaxiques, relations de dépendance syntaxique, cadres sémantiques, etc.). Nous avons donc utilisé ZAP uniquement pour générer des alignements lexicaux phrase à phrase entre les corpus parallèles français et anglais, alignements que nous avons enrichis afin d'effectuer la projection avec un programme que nous avons créé. Pour en évaluer la qualité, nous avons sélectionné aléatoirement 500 phrases, avons effectué une projection manuelle et l'avons comparée à la sortie de la projection automatique. Les performances au niveau du token sont satisfaisantes (précision de 64,21, rappel de 88,48, F-mesure de 74,42), mais elles le sont un peu moins au niveau de l'expression complète (précision de 49,45, rappel de 83,33, F-mesure de 62,07). Une fois les annotations projetées sur le corpus anglais complet, s'en est suivie une nouvelle phase de nettoyage du corpus : correction des annotations incorrectes ou incomplètes, ajout des annotations manquantes, vérification du score d'association des composants de la collocation, etc. Le corpus final contient 5976 collocations annotées. Plusieurs phénomènes liés au processus de traduction ont pu être observés et expliquent la différence du nombre d'annotations : encapsulation, ellipse, transposition, choix lexicaux différents, ou encore reformulation. Nous avons noté que les deux premiers phénomènes étaient assez représentés, notamment pour les collocations les plus fréquentes. Par exemple, *ask* encapsule souvent la collocation *poser|question*, tandis que les substantifs *role* et *measure* se suffisaient régulièrement à eux-mêmes et voyaient leur verbe élidé. Cependant, peu de changements étaient visibles entre les genres.

Le dernier corpus parallèle de notre corpus trilingue devait encore être annoté. Nous avons employé la même méthodologie que pour la précédente, à quelques détails près. D'une part, le code source de ZAP devait être un peu adapté pour pouvoir réaliser des alignements entre l'anglais et l'arabe. D'autre part, nous avons dû cette fois créer une table de traduction bilingue anglais-arabe avec GIZA++. Une fois les alignements générés et enrichis, nous pouvions projeter les annotations. À des fins expérimentales, nous avons créé une autre table de traduction bilingue français-arabe grâce à l'anglais en tant que langue pivot, afin d'effectuer une « double » projection (c'est-à-dire qu'un token arabe aligné soit avec un token français, soit avec un token anglais, dont la projection n'a pas déjà été effectuée, recevrait une annotation), et de la comparer à la « simple » projection. Pour évaluer la qualité de ces projections, nous avons à nouveau sélectionné 500 phrases au hasard, effectué la projection manuellement, et confronté cette dernière aux deux sorties. Il s'avère que les résultats entre les deux types de projection sont proches, la simple projection demeurant meilleure que la double (précision inférieure, rappel légèrement supérieur). Toutes les deux obtiennent des scores faibles au niveau

de la collocation (précision aux alentours de 20,50, rappel aux alentours de 40,00, F-mesure légèrement supérieure à 27,00) et moyens au niveau du token (précision aux alentours de 45,75, rappel aux alentours de 63,75, F-mesure légèrement supérieure à 53,25). Après avoir projeté les annotations sur le corpus arabe complet, nous avons une dernière fois entrepris de corriger les résultats, en suivant la même méthodologie que pour l'anglais, si ce n'est que nous n'avons eu accès à aucun outil pour vérifier les scores d'association. Nous nous en sommes remis à l'avis d'un expert, et le corpus final contient 5342 collocations annotées. Si les résultats de la projection sont largement inférieurs pour l'arabe (20 points en moins sur la F-mesure au niveau du token et 35 en moins au niveau de la collocation), c'est principalement à cause du caractère nominal de l'arabe. En effet, la plupart des verbes à forme non finie sont traduits par des *maṣdar*, résultant en une annexion de deux substantifs (c'est-à-dire en un syntagme nominal avec un complément du nom) plutôt qu'une collocation verbo-nominale. Qui plus est, la voix passive est bien souvent rendue par une tournure périphrastique utilisant le verbe تَمَّ *(tamma)* ayant un *maṣdar* pour sujet plutôt que la forme canonique du passif. En outre, nous avons remarqué les mêmes phénomènes traductionnels que précédemment, avec des cas d'encapsulation, d'ellipse, de transposition, ou encore des choix lexicaux différents et autres reformulations.

Une fois les trois corpus entièrement annotés, nous disposions de toutes les données nécessaires pour mener à bien la quatrième et dernière phase de ce projet, c'est-à-dire l'étude linguistique contrastive. Nous avons séparé cette dernière en deux analyses. La première de ces analyses était quantitative. Nous avons calculé les distances observées entre les composants de toutes les collocations annotées entre les langues et entre les genres. Il en est ressorti qu'en moyenne, et ce dans tous les genres, on observait que la distance était la plus élevée en français avec 3,08 tokens entre les composants d'une collocation. Le français est en effet une langue bien moins synthétique que les deux autres (sans compter les différences liées à la tokenisation, notamment avec l'arabe). Au niveau des genres, c'est sans grande surprise que la distance moyenne la plus élevée se trouve dans le corpus des Nations Unies, avec en moyenne 3,19 tokens entre les composants d'une collocation sur le corpus trilingue. C'est notamment dans ce corpus, l'anglais, que nous avons trouvé la collocation dont les composants sont les plus éloignés, avec pas moins de 42 tokens de distance. Nous avons poursuivi en calculant la proportion d'utilisation des collocations continues et discontinues. Tous genres confondus et en moyenne sur les trois langues, environ 1/5^{e} des collocations sont continues. Cependant, la différence entre les langues est très importante : le français en utilise seulement 10,22%, quand l'anglais en utilise 21,90% et l'arabe 34,35%. Cela s'explique notamment par le fait que l'article défini en arabe n'est pas séparé du substantif qu'il définit dans la tokenisation, que l'article indéfini n'existe pas (l'article zéro est un marqueur d'indéfinitude) et que les adjectifs sont rejetés après le substantif qu'ils qualifient. Au niveau des genres cependant, les résultats ne sont pas spécialement significatifs.

Le deuxième de ces analyses était plus qualitative. Nous avons détaillé à grand renfort d'exemples les divers phénomènes traductionnels que nous avons pu observer et lister dans les différentes typologies dressées après projection. Globalement, nous avons discuté *in extenso* du caractère nominal de l'arabe, avant de passer aux phénomènes liés à une traduction de qualité, tels que l'encapsulation, l'ellipse, les transpositions, etc., pour terminer sur ceux liés à une

traduction d'une qualité moindre, tels que l'appauvrissement lexical, les calques et autres omissions. Compte tenu de toutes observations, nous pouvons décidément conclure que traduire les collocations est loin d'être une mince affaire tant les changements observés sont nombreux.

Après avoir résumé notre travail et avant d'y mettre un point final, nous dressons un bilan des apports réalisés, des limites rencontrées et des perspectives envisagées.

16. APPORTS, LIMITES ET PERSPECTIVES

Arrivé au bout de ce projet, nous pouvons dresser un bilan de ce qui a été fait et surtout en tirer des conclusions sur ce qu'il a apporté, quelles ont été les limites auxquelles nous avons été confronté et quelles perspectives pourraient être envisagées.

16.1. Apports

L'apport le plus évident est l'objet même de ce projet, à savoir la création d'un corpus parallèle trilingue de plus de 100 000 phrases entièrement annoté en collocations verbo-nominales[34]. Ce corpus, bien qu'il présente des limites que nous abordons plus bas, pourrait tout à fait être utilisable dans les domaines intéressés par les corpus parallèles. On peut envisager, entre autres, que le nôtre puisse être utiliser pour en extraire des patrons de collocations pour la création de dictionnaires, dans un contexte pédagogique ou d'apprentissage de langues étrangères, ou encore pour une étude linguistique contrastive comme celle que nous avons menée. L'autre utilité d'un tel corpus parallèle serait de réutiliser les données annotées pour développer et évaluer des systèmes d'annotation automatique pour la tâche d'identification de collocations verbo-nominales. C'est exactement ce que nous avons voulu faire.

Pour « boucler la boucle » et arriver à une conclusion définitive des performances de VarIDE quant à l'annotation automatique des collocations verbo-nominales dans les trois langues du projet qui a été le nôtre, nous avons procédé à une évaluation (sur laquelle nous ne dirons que quelques mots) de l'outil avec nos données annotées. Pour chaque langue, nous avons séparé le corpus entier en un jeu d'entraînement et un jeu de test (répartition 80/20). Pour le jeu de test, le contenu est équitablement réparti entre les quatre sous-corpus sur lesquels nous avons travaillé, pour un total de 21 146 phrases (contre 84 584 pour le jeu d'entraînement). Voici les statistiques des jeux de données pour chacune des trois langues :

Corpus	FR	EN	AR
Train (nombre d'annotations)	5694	4643	4169
Test (nombre d'annotations)	1517	1333	1173
Proportion train/test	78,86% / 21,04%	77,69% / 22,31%	78,04% / 21,96%

Tableau 23 : Statistiques des jeux d'entraînement et de test pour évaluation finale

On constate que malgré le nombre moins important d'annotations pour l'arabe, la proportion de la répartition des annotations entre le jeu d'entraînement et le jeu de test est très équilibrée pour les trois langues. Une fois les jeux de données prêts, nous pouvions lancer le processus d'annotation automatique de VarIDE, puis le script d'évaluation. L'annotation s'est faite sur les données de test dont les annotations ont été retirées. Le résultat a été confronté au jeu de test entièrement annoté. Les résultats sont transcrits dans le tableau suivant :

Langue	Précision	Rappel	F-mesure
FR	1423 / 1733 = **82,11**	1423 / 1517 = **93,80**	**87,57**
EN	970 / 1221 = **79,44**	970 / 1333 = **72,77**	**75,96**
AR	77 / 133 = **57,89**	77 / 1173 = **6,44**	**11,79**

Tableau 24 : Résultats de l'évaluation trilingue de VarIDE

[34] Les corpus annotés sont disponibles sur le dépôt GitHub du projet : [lien]

Les résultats sont drastiquement différents d'une langue à l'autre. Si le français et l'anglais obtiennent respectivement des scores très bons et plus que corrects, l'arabe en revanche souffre d'un rappel si bas (6,44) que sa F-mesure arrive difficilement à 11,79. Ces chiffres extrêmement disparates peuvent sembler étonnants de prime abord, mais ils sont congruents avec ceux obtenus[35] au cours de la *Shared Task 1.1* de PARSEME pendant laquelle VarIDE a été évalué. En effet, bien que le système ait obtenu une F-mesure supérieure à 50 pour 6 langues, il a obtenu une F-mesure inférieure à 20 dans 7 autres, dont une à 7,87 pour le turc et une autre à 1,96 pour le lituanien. On en conclut que le système est donc fortement dépendant de la langue, et même si les résultats de l'évaluation arabe sont décevants à titre personnel, ils ne sont pas inexplicables ou saugrenus pour autant.

L'autre apport méritant d'être mentionné est la méthodologie originale que nous avons proposée pour la projection des annotations d'un corpus à un autre. Même si elle demeure imparfaite, elle nous semble, au vu des résultats obtenus, tout à fait valable et surtout perfectible. En outre, on constate qu'il semble que la projection d'une langue bien dotée vers une langue plus rare puisse être une bonne alternative à l'utilisation d'outils d'annotation automatique dépendants de corpus d'entraînement annotés, qui peuvent s'avérer rares.

16.2. Limites

À ces apports, des limites assez importantes viennent faire contrepoids. D'une part, ce projet étant le travail d'un mémoire de Master, il a été mené en majeure partie en autonomie. Cependant, un travail d'annotation, pour qu'il puisse être entièrement valable, devrait, nous le pensons, être le fruit d'efforts menés collectivement. En effet, même si nous nous sommes appuyé sur une méthodologie assez stricte avec le guide d'annotation que nous avons rédigé, un travail d'annotation sans possibilité de le confronter à celui d'autres annotateurs et ainsi mesurer l'accord entre eux, reste difficilement acceptable. Sans nul doute, un nombre important d'annotations n'auraient pas demeuré si plusieurs personnes avaient travaillé dessus. L'avis d'experts, cependant, a été un appui de poids.

En outre, bien que nous ayons tenté une « double projection » pour l'arabe avec les deux premiers corpus annotés, nous nous sommes aperçu que croiser les alignements lexicaux avait peu d'impact et avait même un impact négatif sur les résultats. En revanche, nous l'avons dit, la méthodologie proposée est perfectible. Nous pensons notamment à deux choses assez évidentes. D'une part, pour éviter les erreurs de projection sur des tokens dont les parties du discours ne sont pas en adéquation avec le but recherché, des conditions supplémentaires pourraient entrer en jeu dans le script. Dans le cadre de ce travail, si un token-cible est étiqueté autre chose que `NOUN` ou `VERB`, le token pourrait ne pas être projeté du tout. Si, comme c'était le cas souvent pour l'arabe, les deux tokens-cibles d'une collocation donnée ont la même étiquette, on pourrait s'abstenir tout bonnement de faire la projection. Pour ne pas perdre l'information, un rapport de projection pourrait être généré, mentionnant les raisons de l'absence de projection d'une collocation donnée. Les gains en précision pourraient être significatifs. D'autre part, s'ils existent, ajouter aux tables de traduction des entrées lexicales issues de dictionnaires électroniques fiables pourrait être envisagé, ou encore enrichir les

[35] Résultats de la Shared Task 1.1 de PARSEME : [lien]

alignements lexicaux avec des patrons de collocations avérés dans la langue-cible, s'ils sont connus.

Ces solutions aideraient grandement ce qui, pour nous, apparaît comme la limite la plus importante de ce projet : le travail humain trop conséquent pour le nettoyage de corpus après projection. En effet, cette étape a été extrêmement coûteuse en temps, et le corpus ne faisait « que » 300 000 lignes en tout. Les mêmes efforts sur des corpus plus grands sont inimaginables. Ainsi, il reste de grandes améliorations possibles quant à l'automatisation du post-traitement de la projection. On peut imaginer, par exemple, que le travail manuel que nous avons mené avec les expressions régulières pour retrouver les annotations erronées, manquantes ou incomplètes soit automatisé ou semi-automatisé. Malgré les nombreux scripts que nous avons créés pour nous faciliter la tâche et accélérer le processus, ils demeurent insuffisants.

16.3. Perspectives

Nous venons de le mentionner, mais si ce travail devait être perpétué, la première chose que nous ferions serait de créer un programme pour faciliter l'étape de nettoyage du corpus après projection. Nous ne le répèterons pas davantage, mais cette étape nécessite de trop grands efforts pour demeurer telle quelle, surtout quand il s'agit d'un travail dont la portée est limitée dans le temps.

Par ailleurs, bien que nous ayons proposé une méthodologie qui exploite les alignements lexicaux de ZAP, cet outil mériterait d'être amélioré. Premièrement, pour les projections de l'anglais vers n'importe quelle autre langue, nous l'avons fait, il est assez aisé de créer une table de traduction et de développer la palette des langues-cibles possibles. Deuxièmement, améliorer le programme de sorte à pouvoir projeter des annotations depuis une autre langue que l'anglais pourrait être utile. Cela éviterait notamment de devoir passer par l'anglais en tant que langue pivot pour effectuer une projection. Troisièmement, ce programme serait encore plus intéressant s'il était possible d'intégrer rapidement une catégorie supplémentaire pour des annotations personnelles, comme c'était le cas pour nous. Notre méthodologie évite justement de devoir faire cette opération, mais c'est une amélioration qui pourrait être intéressante.

Enfin, depuis le début de ce travail et surtout depuis la *Shared Task 1.1* de PARSEME, des progrès ont été faits sur la tâche d'identification des expressions polylexicales, notamment avec les systèmes soumis à la *Shared Task 1.2* (Ramisch et al., 2020). En effet, quand le système en 1ère position du classement général de la *Shared Task 1.1* (TRAVERSAL) obtenait une F-mesure de 54 sur 19 langues, le meilleur système de la Shared Task suivante en obtenait une supérieure de plus de 16 points (MTLB-Struct avec 70,14 sur 14 langues). En outre, les résultats concernant l'identification de collocations qui n'ont pas été apprises au cours de l'entraînement sont en hausse constante. Il serait intéressant de réitérer l'évaluation finale menée avec VarIDE avec un des systèmes de cette *Shared Task* pour mesurer les améliorations apportées. L'annotation des autres types d'EP, occultés au cours de ce projet, serait également intéressante à explorer.

V. ANNEXES

Annexe A. GUIDE D'ANNOTATION

Le guide d'annotation peut être consulté sur le dépôt GitHub de ce projet.

Annexe B. TRANSLITTERATION ARABE

Les tableaux suivants font état de la translittération arabe utilisée tout au long du présent travail. La langue arabe est une langue sémitique dont la majorité des lettres sont des consonnes qui en forment le squelette consonantique *(rasm)* :

Consonnes				
Lettre	**Translittération**	**Nom**	**API**	**Exemples**
ب	b	*bāʾ*	[b]	*bateau (fr)*
ت	t	*tāʾ*	[t]	*tableau (fr)*
ث	ṯ	*ṯāʾ*	[θ]	*thick (en)*
ج	ǧ	*ǧim*	[ʒ]	*jouet (fr)*
ح	ḥ	*ḥāʾ*	[ħ]	*wieħed (mt)*
خ	ḫ	*ḫāʾ*	[x]	*naranja (es), Buch (de)*
د	d	*dāl*	[d]	*double (fr, en)*
ذ	ḏ	*ḏāl*	[ð]	*weather (en)*
ر	r	*rāʾ*	[r]	*ragazza (it)*
ز	z	*zāy*	[z]	*zoo (fr, en)*
س	s	*sīn*	[s]	*sabre (fr)*
ش	š	*šīn*	[ʃ]	*chambre (fr), shin (en)*
ص	ṣ	*ṣāḍ*	[sˤ]	/
ض	ḍ	*ḍāḍ*	[dˤ]	/
ط	ṭ	*ṭāʾ*	[tˤ]	/
ظ	ẓ	*ẓāʾ*	[ðˤ]	/
ع	ʿ	*ʿayn*	[ʕ]	*ravn (dk)*
غ	ġ	*ġayn*	[ʁ]	*robe (fr)*
ف	f	*fāʾ*	[f]	*four (fr)*
ق	q	*qāf*	[q]	/
ك	k	*kāf*	[k]	*quille (fr), killer (en)*
ل	l	*lām*	[l]	*lune (fr)*
م	m	*mīm*	[m]	*masse (fr)*
ن	n	*nūn*	[n]	*nacelle (fr)*
ه	h	*hāʾ*	[h]	*heavy (en)*
ء (ئ ؤ إ أ)	ʾ	*hamza*	[ʔ]	*alerte (fr), oubli (fr), ibérique (fr)*

L'alphabet arabe comporte également 3 lettres dites « faibles », dont deux d'entre elles peuvent être alternativement des consonnes ou des voyelles. Quand elles sont voyelles, leur prononciation est longue.

Voyelles longues et semi-consonnes							
Lettre	**Translittération**		**Nom**	**API**		**Exemples**	
ا ou ى	ā		*alif (maqsūra)*	[aː]		*blâme (fr), father (en)*	
و	w	ū	*wāw*	[w]	[uː]	*ouate (fr), willow (en)*	*noon (en)*
ي	y	ī	*yāʾ*	[j]	[iː]	*yaourt (fr), yellow (en)*	*dîme (fr), feeling (en)*

L'arabe ne connaît pas de voyelle réelle. Les voyelles que l'on entend à l'oral ne sont généralement pas marquées, mais lorsqu'elles le sont, c'est sous la forme de diacritiques. On compte 3 voyelles brèves et un signe diacritique symbolisant l'absence de vocalisation. Dans un souci d'exhaustivité, nous avons ajouté le signe diacritique de la gémination (redoublement d'une consonne).

Voyelles brèves et diacritiques				
Diacritique arabe	**Translittération**	**Nom**	**API**	**Exemples**
ـَ	a	*fatḥa*	[a]	*balle (fr)*
ـُ	u	*ḍamma*	[u]	*poule (fr)*
ـِ	i	*kasra*	[i]	*mille (fr)*
ـْ	ø	*sukun*	/	/
ـّ	/	*šadda*	[.]	*collocation (fr)*

Annexe C. EXEMPLE D'UNE PHRASE AU FORMAT CUPT

Les fichiers au format `cupt` sont des fichiers tabulés. Chaque fichier commence par la ligne démarrant par `# global.columns`. Cette ligne indique les en-têtes de chaque colonne. Nous les commentons ici :

- `ID` : identifiant numérique unique attribué au token. Il commence à 1 pour chaque nouvelle phrase et peut être noté avec un intervalle (p. ex. 2-3) pour les tokens « composés » comme *des* en français *(de + les)*.
- `FORM` : forme du token dans la phrase originale ou signe de ponctuation,
- `LEMMA` : lemme du token,
- `UPOS` : étiquette grammaticale *Universal Dependencies*,
- `XPOS` : étiquette grammaticale spécifique à la langue ou blanc souligné si cette information n'est pas fournie,
- `FEATS` : caractéristiques morphologiques (genre, nombre, etc.) ou blanc souligné si cette information n'est pas fournie,
- `HEAD` : identifiant numérique du token tête. Si le token n'a pas de token tête (il est la racine de la phrase), ce champ est marqué par un 0,
- `DEPREL` : relation de dépendance syntaxique entre le token et sa tête. Si le token n'a pas de token tête (il est la racine de la phrase), ce champ est marqué par un 0,
- `DEPS` : graphe amélioré de dépendances syntaxiques ou blanc souligné si cette information n'est pas fournie,
- `MISC` : toute autre annotation,
- `PARSEME:MWE` : colonne pour les annotations utilisées pendant notre projet, c'est-à-dire `X:COLL,X` pour les collocations et `*` pour le reste (voir Annexe A).

Chaque phrase est ensuite introduite par deux lignes de métadonnées. La première indique la source dans le corpus duquel est extraite la phrase (`# source_sent_id`) et lui attribue un identifiant unique, tandis que la seconde indique le texte original (`# text`). Une ligne vide marque la séparation entre les différentes phrases. La page suivante présente un extrait d'un fichier `cupt` pour une phrase en français.

```
# global.columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC
PARSEME:MWE
# source_sent_id = WM/WM.tri.fr::1
# text = La découverte de ces similitudes offre l'espoir d'avancées
thérapeutiques qui pourraient améliorer simultanément de nombreuses
maladies,.
1    La   le   DET  _    Definite=Def|Gender=Fem|Number=Sing|PronType=Art
     2    det  _    _    *
2    découverte    découverte    NOUN _    Gender=Fem|Number=Sing  6
     nsubj     _    _    *
3    de   de   ADP  _    _    5    case _    _    *
4    ces  ce   DET  _    Number=Plur|PronType=Dem    5    det  _    _    *
5    similitudes   similitude    NOUN _    Gender=Fem|Number=Plur  2    nmod
     _    _    *
6    offre    offrir   VERB _
     Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin    0    root _
     _    *
7    l'   le   DET  _    Definite=Def|Number=Sing|PronType=Art 8    det  _
     _    *
8    espoir   espoir   NOUN _    Gender=Masc|Number=Sing 6    obj  _    _
     *
9    d'   de   ADP  _    _    10   case _    _    *
10   avancées  avancée   NOUN _    Gender=Fem|Number=Plur  8    nmod _    _
     *
11   thérapeutiques thérapeutique ADJ  _    Gender=Fem|Number=Plur  10   amod
     _    _    *
12   qui  qui  PRON _    PronType=Rel  13   nsubj    _    _    *
13   pourraient    pouvoir  VERB _
     Mood=Cnd|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin    10   acl:relcl
     _    _    *
14   améliorer améliorer VERB _    VerbForm=Inf  13   xcomp    _    _    *
15   simultanément simultanément ADV  _    _    14   advmod   _    _    *
16   de   un   DET  _    Definite=Ind|Number=Plur|PronType=Art 18   det  _
     _    *
17   nombreuses    nombreux ADJ  _    Gender=Fem|Number=Plur  18   amod _
     _    *
18   maladies maladie   NOUN _    Gender=Fem|Number=Plur  14   obj  _    _
     *
19   ,    ,    PUNCT    _    _    6    punct    _    _    *
20   .    .    PUNCT    _    _    6    punct    _    _    *
```

BIBLIOGRAPHIE

Akbik, A., & Vollgraf, R. (2018). ZAP : An Open-Source Multilingual Annotation Projection Framework. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Altenberg, B., & Aijmer, K. (2000). The English-Swedish Parallel Corpus : A resource for contrastive research and translation studies. *Corpus linguistics and linguistic theory, ed. by Christian Mair and Marianne Hundt*, 15-33.

Altenberg, B., & Granger, S. (2002). Recent trends in cross-linguistic lexical studies. *Lexis in Contrast: Corpus Based Approaches, Amsterdam and Philadelphia: John Benjamins*, 3-48.

Artstein, R. (2017). Inter-annotator agreement. In *Handbook of linguistic annotation* (p. 297-313). Springer.

Attia, M. (2006). *Accommodating Multiword Expressions in an Arabic LFG Grammar*. *4139*, 87-98. https://doi.org/10.1007/11816508_11

Baldwin, T., & Kim, S. N. (2010). Multiword Expressions. In N. Indurkhya & F. J. Damerau (Éds.), *Handbook of Natural Language Processing* (Second Edition, p. 267-292). CRC Press.

Bartsch, S. (2004). *Structural and functional properties of collocations in English : A corpus study of lexical and pragmatic constraints on lexical co-occurrence*. Gunter Narr Verlag.

Benson, M. (1990). Collocations and general-purpose dictionaries. *International Journal of Lexicography*, *3*(1), 23-34.

Bird, S., & Liberman, M. (2001). A formal framework for linguistic annotation. *Speech communication*, *33*(1-2), 23-60.

Blau, J. (1981). *The Renaissance of Modern Hebrew and Modern Standard Arabic : Parallels and differences in the revival of two Semitic languages* (Vol. 18). Univ of California Press.

Boulaknadel, S., Daille, B., & Aboutajdine, D. (2008). *A multi-word term extraction program for Arabic language*. 4.

Brants, T. (2000). Inter-annotator Agreement for a German Newspaper Corpus. *LREC*.

Brashi, A. S. (2005). *Arabic collocations : Implications for translations* [PhD Thesis]. University of Western Sydney.

Chiao, Y.-C., Kraif, O., Laurent, D., Nguyen, T. M. H., Semmar, N., Stuck, F., Véronis, J., & Zaghouani, W. (2006). Evaluation of multilingual text alignment systems : The ARCADE II project. *5th international Conference on Language Resources and Evaluation-LREC'06*.

Christodouloupoulos, C., & Steedman, M. (2015). A massively parallel corpus : The Bible in 100 languages. *Language Resources and Evaluation*, *49*(2), 375-395. https://doi.org/10.1007/s10579-014-9287-y

Church, K., Gale, W., Hanks, P., & Hindle, D. (1991). Using statistics in lexical analysis. *Lexical acquisition: exploiting on-line resources to build a lexicon*, *115*, 164.

Church, K., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, *16*(1), 22-29.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., & Todirascu, A. (2017). Multiword expression processing : A survey. *Computational Linguistics*, *43*(4), 837-892. https://doi.org/10.1162/COLI_a_00302

Cruse, D. A. (1986). *Lexical semantics*. Cambridge university press.

Dagan, I., Church, K., & Gale, W. (1999). Robust bilingual word alignment for machine aided translation. In *Natural Language Processing Using Very Large Corpora* (p. 209-224). Springer.

Dichy, J. (1994). La pluriglossie de l'arabe. *Bulletin d'études orientales*, 19-42.

Dunning, T. E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, *19*(1), 61-74.

Emery, P. G. (1991). Collocation in Modern Standard Arabic. *Zeitschrift Für Arabische Linguistik*, *23*, 56-65.

Erjavec, T. (2004). MULTEXT-East Version 3 : Multilingual Morphosyntactic Specifications, Lexicons and Corpora. *LREC*.

Erjavec, T. (2012). MULTEXT-East : Morphosyntactic resources for Central and Eastern European languages. *Language resources and evaluation*, *46*(1), 131-142.

Evert, S. (2005). *The statistics of word cooccurrences : Word pairs and collocations*.

Evert, S. (2008). Corpora and collocations. *Corpus linguistics. An international handbook*, *2*, 1212-1248.

Evert, S., & Hardie, A. (2011). *Twenty-first century Corpus Workbench : Updating a query architecture for the new millennium*.

Fillmore, C. J., Kay, P., & O'connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions : The case of let alone. *Language*, 501-538.

Gale, W. A., & Church, K. (1993). A program for aligning sentences in bilingual corpora. *Computational linguistics*, *19*(1), 75-102.

Gale, W. A., & Church, K. W. (1991). Identifying Word Correspondences in Parallel Texts. *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California*.

Garcia, M., García-Salido, M., & Alonso-Ramos, M. (2017). Using bilingual word-embeddings for multilingual collocation extraction. *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, 21-30. https://doi.org/10.18653/v1/W17-1703

Green, S., de Marneffe, M.-C., & Manning, C. D. (2012). Parsing Models for Identifying Multiword Expressions. *Computational Linguistics*, *39*(1), 195-227. https://doi.org/10.1162/COLI_a_00139

Green, S., & Manning, C. D. (2010). Better Arabic parsing : Baselines, evaluations, and analysis. *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 394-402.

Gries, S. T., & Berez, A. L. (2017). Linguistic annotation in/for corpus linguistics. *Handbook of linguistic annotation*, 379-409.

Grimm, P. S. (2009). Collocation in Modern Standard Arabic revisited. *Zeitschrift Für Arabische Linguistik*, *51*, 22-41.

Habash, N. Y. (2010). Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, *3*(1), 1-187.

Hausmann, F. J. (1989). *Wörterbücher : Ein internationales Handbuch zur Lexikographie*. W. de Gruyter.

Hausmann, F. J., & Blumenthal, P. (2006). Présentation : Collocations, corpus, dictionnaires. *Langue française*, *2*, 3-13.

Ide, N., & Pustejovsky, J. (2017). *Handbook of linguistic annotation*. Springer.

Ide, N., & Véronis, J. (1994). MULTEXT : Multilingual text tools and corpora. *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*.

Imbert, F., & Pinon, C. (2008). *L'arabe dans tous ses états! : La grammaire arabe en tableaux*. Ellipses.

Izwaini, S. (2015). Patterns of Lexical Collocations in Arabic*. *Zeitschrift für Arabische Linguistik*, 72-99.

Kenning, M.-M. (2010). What are parallel and comparable corpora and how can we use them. *The Routledge handbook of corpus linguistics*, 487-500.

Kraif, O. (2001). Exploitation des cognats pour l'alignement : Architecture et évaluation. *Traitement automatique des langues*, *42*(3), 833-867.

Kraif, O. (2014). *Corpus parallèles, corpus comparables : Quels contrastes?* [PhD Thesis]. Université de Poitiers.

Kraif, O. (2015). Multialignement vs bialignement : À plusieurs, c'est mieux! *TALN 2015, 22e conférence sur le Traitement automatique des langues naturelles*.

Kübler, N., & Aston, G. (2010). Using corpora in translation. *The Routledge handbook of corpus linguistics*, 505-515.

Lamraoui, F., & Langlais, P. (2013). Yet another fast, robust and open source sentence aligner. Time to reconsider sentence alignment. *XIV machine translation summit*.

Langacker, R. (2008). *Cognitive Grammar : A Basic Introduction New York Oxford University Press*.

Lee, D. Y. (2010). What corpora are available. *The Routledge handbook of corpus linguistics*, 107-121.

Liang, P., Taskar, B., & Klein, D. (2006). Alignment by agreement. *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics -*, 104-111. https://doi.org/10.3115/1220835.1220849

Lü, Y., & Zhou, M. (2004). Collocation translation acquisition using monolingual corpora. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, 167-174.

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Bilingual word representations with monolingual quality in mind. *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, 151-159.

Ma, X. (2006). Champollion : A Robust Parallel Text Sentence Aligner. *LREC*, 489-492.

Mathet, Y., Widlöcher, A., & Métivier, J.-P. (2015). The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment. *Computational Linguistics*, *41*(3), 437-479. https://doi.org/10.1162/COLI_a_00227

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies : An advanced resource book*. Taylor & Francis.

McEnery, T., & Xiao, Z. (2007). Parallel and comparable corpora : The state of play. *Corpus-based perspectives in linguistics*, *6*.

McKeown, K., Smadja, F., & Hatzivassiloglou, V. (1996). *Translating collocations for bilingual lexicons : A statistical approach*.

Mel'čuk, I. (1995). Phrasemes in language and phraseology in linguistics. *Idioms: Structural and psychological perspectives*, 167-232.

Mel'čuk, I. (2013). Tout ce que nous voulions savoir sur les phrasèmes, mais. *Cahiers de lexicologie*, *102*(1), 129-149.

Mel'čuk, I., & Zholkovsky, A. (1988). The Explanatory Combinatorial Dictionary. *Relational Models of the Lexicon*, 41-74.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*.

Mitkov, R. (2004). *The Oxford handbook of computational linguistics*. Oxford University Press.

Nofal, K. H. (2012). Collocations in English and Arabic : A comparative study. *English Language and Literature Studies*, *2*(3), 75-93. https://doi.org/10.5539/ells.v2n3p75

Obeid, O., Zalmout, N., Khalifa, S., Taji, D., Oudah, M., Alhafni, B., Inoue, G., Eryani, F., Erdmann, A., & Habash, N. (2020). CAMeL tools : An open source python toolkit for Arabic natural language processing. *Proceedings of the 12th language resources and evaluation conference*, 7022-7032.

Och, F. J., & Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, *29*(1), 19-51. https://doi.org/10.1162/089120103321337421

O'Keeffe, A., & Farr, F. (2003). Using language corpora in initial teacher education : Pedagogic issues and practical applications. *Tesol Quarterly*, *37*(3), 389-418.

O'Keeffe, A., & McCarthy, M. (2010). *The Routledge handbook of corpus linguistics*. Routledge.

Pasquer, C. (2017, juin). Expressions polylexicales verbales : Étude de la variabilité en corpus. *TALN-RECITAL 2017*. https://hal.archives-ouvertes.fr/hal-01637355

Pasquer, C., Ramisch, C., Savary, A., & Antoine, J.-Y. (2018). VarIDE at PARSEME Shared Task 2018 : Are Variants Really as Alike as Two Peas in a Pod? *COLING Workshop on Linguistic Annotation, Multiword Expressions and Constructions*.

Pearson, J. (1998). *Terms in context* (Vol. 1). John Benjamins Publishing.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza : A Python Natural Language Processing Toolkit for Many Human Languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 101-108. https://doi.org/10.18653/v1/2020.acl-demos.14

Ramisch, C. (2012). *A generic and open framework for multiword expressions treatment : From acquisition to applications* [PhD Thesis]. Université de Grenoble.

Ramisch, C., Cordeiro, S., Savary, A., Vincze, V., Mititelu, V., Bhatia, A., Buljan, M., Candito, M., Gantar, P., & Giouli, V. (2018). Edition 1.1 of the parseme shared task on automatic identification of verbal multiword expressions. *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, 222-240.

Ramisch, C., Savary, A., Guillaume, B., Waszczuk, J., Candito, M., Vaidya, A., Mititelu, V. B., Bhatia, A., Iñurrieta, U., & Giouli, V. (2020). Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, 107-118.

Ramisch, C., Villavicencio, A., & Boitet, C. (2010). Mwetoolkit : A framework for multiword expression identification. *LREC*, *10*(5), 662-669.

Ramshaw, L. A., & Marcus, M. P. (1995). Text Chunking using Transformation-Based Learning. *ArXiv:Cmp-Lg/9505040*. http://arxiv.org/abs/cmp-lg/9505040

Reimers, N., & Gurevych, I. (2020, octobre 5). Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. http://arxiv.org/abs/2004.09813

Rychlỳ, P. (2008). A Lexicographer-Friendly Association Score. *RASLAN*, 6-9.

Ryding, K. C. (2005). *A reference grammar of modern standard Arabic*. Cambridge university press.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword Expressions : A Pain in the Neck for NLP. In A. Gelbukh (Éd.), *Computational Linguistics and Intelligent Text Processing* (Vol. 2276, p. 1-15). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45715-1_1

Saif, A. M., & Aziz, M. J. (2011). An automatic collocation extraction from Arabic corpus. *Journal of Computer Science, 7*(1), 6-11.

Santos, A. (2011). A survey on parallel corpora alignment. *Proceedings of MI-Star*, 117-128.

Savary, A., Candito, M., Mititelu, V. B., Bejček, E., Cap, F., Čéplö, S., Cordeiro, S. R., Gülşen Eryiğit, Giouli, V., Gompel, M. V., HaCohen-Kerner, Y., Kovalevskaitė, J., Krek, S., Liebeskind, C., Monti, J., Escartín, C. P., Plas, L. V. D., Behrang QasemiZadeh, Ramisch, C., … Vincze, V. (2018). PARSEME multilingual corpus of verbal multiword expressions. In *Multiword expressions at length and in depth : Extended papers from the MWE 2017 workshop*. Language Science Press. https://doi.org/10.5281/ZENODO.1471591

Savary, A., Ramisch, C., Cordeiro, S. R., Sangati, F., Vincze, V., Qasemi Zadeh, B., Candito, M., Cap, F., Giouli, V., & Stoyanova, I. (2017). The PARSEME shared task on automatic identification of verbal multiword expressions. *Proceedings of the 13th Workshop on Multiword Expression (MWE 2017)*, 31-47.

Savary, A., Sailer, M., Parmentier, Y., Rosner, M., Rosén, V., Przepiórkowski, A., Krstev, C., Vincze, V., Wójtowicz, B., & Losnegaard, G. S. (2015). PARSEME–PARSing and Multiword Expressions within a European multilingual network. *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., & Guzmán, F. (2019, juillet 15). WikiMatrix : Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. *arXiv:1907.05791 [cs]*. http://arxiv.org/abs/1907.05791

Seretan, V. (2008). *Collocation extraction based on syntactic parsing* [PhD Thesis, University of Geneva]. http://archive-ouverte.unige.ch/unige:78

Seretan, V., Nerima, L., & Wehrli, E. (2004). A tool for multi-word collocation extraction and visualization in multilingual corpora. *Proceedings of the 11th EURALEX International Congress*, 755-766.

Side, R. (1990). *Phrasal verbs : Sorting them out*.

Simões, A., & Almeida, J. J. (2003). NATools-a statistical word aligner workbench. *Procesamiento del lenguaje natural, 31*.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Smadja, F. (1993). Retrieving collocations from text : Xtract. *Computational linguistics, 19*(1), 143-178.

Somers, H. (2001, avril). Bilingual parallel corpora and language engineering. *Anglo-Indian Workshop « Language Engineering for South-Asian Languages » (LESAL)*.

Tiedemann, J. (2003). Combining clues for word alignment. *10th Conference of the European Chapter of the Association for Computational Linguistics*.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In N. Calzolari, K. Choukri, T. Declerck, M. Ugur Dogan, B. Maegaard, J. Mariani, J. Odijk, & S. Piperidis (Éds.), *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)* (p. 2214-2218). European Language Resources Association (ELRA).

Todirascu, A., & Cargill, M. (2019). SimpleApprenant : A platform to improve French L2 learners' knowledge of multiword expressions. *CALL and complexity*, 356.

Todirascu, A., Cargill, M., & François, T. (2019). PolylexFLE : Une base de données d'expressions polylexicales pour le FLE (PolylexFLE: a database of multiword expressions for French L2 language learning). *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume I: Articles longs*, 143-156.

Todiraşcu, A., Heid, U., Ştefănescu, D., Tufiş, D., Gledhill, C., Weller, M., & Rousselot, F. (2008). Vers un dictionnaire de collocations multilingue. *Cahiers de linguistique*, *33*(1), 161-186.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work* (Vol. 6). John Benjamins Publishing.

Tutin, A. (2004). Pour une modélisation dynamique des collocations dans les textes. *Actes d'EURALEX*.

Tutin, A., Esperança-Rodier, E., Iborra, M., & Reverdy, J. (2015). Annotation of multiword expressions in French. *European Society of Phraseology Conference (EUROPHRAS 2015)*, 60-67.

Tutin, A., & Grossmann, F. (2002). Collocations régulières et irrégulières : Esquisse de typologie du phénomène collocatif. *Revue francaise de linguistique appliquée*, *Vol. VII*(1), 7-25.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., & Trón, V. (2007). Parallel corpora for medium density languages. *Amsterdam Studies In The Theory And History Of Linguistic Science Series 4*, *292*, 247.

Villavicencio, A., Baldwin, T., & Waldron, B. (2004). A Multilingual Database of Idioms. *LREC*.

Wehrli, E. (2007). Fips, a « deep » linguistic multilingual parser. *Proceedings of the Workshop on Deep Linguistic Processing - DeepLP '07*, 120. https://doi.org/10.3115/1608912.1608931

Williams, G. (2001). Sur les caractéristiques de la collocation. *Actes de la 8ème conférence sur le Traitement Automatique des Langues Naturelles. Tutoriels*, 9-16.

Wu, D. (2000). Alignment. In *Handbook of natural language processing*.

Zaidi, S., Laskri, M. T., & Abdelali, A. (2010). Arabic collocations extraction using Gate. *International Conference on Machine and Web Intelligence*, 473-475. https://doi.org/10.1109/ICMWI.2010.5648038

Zampieri, N., Scholivet, M., Ramisch, C., & Favre, B. (2018). Veyn at parseme shared task 2018 : Recurrent neural networks for vmwe identification. *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, 290-296.

Zhang, W., Yoshida, T., Tang, X., & Ho, T.-B. (2009). Improving effectiveness of mutual information for substantival multiword expression extraction. *Expert Systems with Applications*, *36*(8), 10919-10930. https://doi.org/10.1016/j.eswa.2009.02.026

Ziemski, M., Junczys-Dowmunt, M., & Pouliquen, B. (2016). The United Nations Parallel Corpus v1.0. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 3530-3534. https://www.aclweb.org/anthology/L16-1561

RESUME / ABSTRACT

Résumé : Les expressions polylexicales (EP) sont omniprésentes en langue. Leur nature hétérogène les rend particulièrement difficiles à identifier avec les outils et techniques de Traitement Automatique des Langues (TAL). Bien que de nombreuses méthodologies aient été envisagées au fil des années, l'identification automatique des EP demeure un défi de nos jours. Parmi les différentes catégories d'EP (expressions idiomatiques, mots composés, certaines entités nommées…), les collocations sont définies comme étant des combinaisons récurrentes et arbitraires apparaissant ensemble plus souvent que par le simple fait du hasard. Malheureusement, le nombre de ressources annotées en collocations, d'autant plus lorsqu'il s'agit de corpus parallèles multilingues, est incroyablement limité. Ce mémoire présente une méthodologie d'annotation semi-automatique des collocations verbo-nominales dans un corpus multi-genre en français, ainsi qu'une méthodologie de projection de ces mêmes annotations vers deux corpus parallèles en anglais et en arabe.

Mots-clés : traitement automatique des langues, collocations, expressions polylexicales, annotation, corpus parallèles

Abstract: Multiword expressions (MWEs) are pervasive in language. Their heterogeneous nature makes them especially difficult to identify with Natural Language Processing (NLP) tools and techniques. Even though numerous methodologies have been considered over the years, automatic MWE identification is still challenging nowadays. Among the several MWE categories (idioms, compounds, named entities…), collocations are defined as recurrent arbitrary word combinations appearing more often than chance. Unfortunately, the number of resources annotated for collocations, even more so in multilingual parallel corpora, is incredibly scarce. This thesis presents a methodology for the semi-automatic annotation of verb-noun collocations in a multi-genre corpus in French, as well as a methodology for projecting these same annotations onto two parallel corpora in English and Arabic.

Keywords: natural language processing, collocations, multiword expressions, annotation, parallel corpora